Profiling of DNA methylation and single nucleotide polymorphism for diagnosis, prognosis and targeting DNA methyltransferases for therapeutic intervention of breast cancer

Arunima Shilpi

Department of Life Science

National Institute of Technology Rourkela

Profiling of DNA methylation and single nucleotide polymorphism for diagnosis, prognosis and targeting DNA methyltransferases for therapeutic intervention of breast cancer

Dissertation submitted in partial fulfilment of

the requirements of the degree of

Doctor of Philosophy

in

Life Science

by

Arunima Shilpi

(Roll Number: 511LS102)

based on research carried out

under the supervision of

Prof. Samir Kumar Patra

and

Prof. Bibekanand Mallick

November, 2016

Department of Life Science

National Institute of Technology Rourkela

Department of Life Science
National Institute of Technology Rourkela

November 2, 2016

Certificate of Examination

Roll Number: *511LS102*

Name: *Arunima Shilpi*

Title of Dissertation: *Profiling of DNA methylation and single nucleotide polymorphism for diagnosis, prognosis and targeting DNA methyltransferases for therapeutic intervention of breast cancer*

We below signed, after checking the dissertation mentioned above and the official record books (s) of the student, hereby state our approval of the dissertation submitted in partial fulfilment of the requirements of the degree of Doctor of philosophy in Department of Life Science National Institute of Technology Rourkela. We are satisfied with the volume, quantity, correctness, and originality of the work.

Bibekanand Mallick Co-Supervisor	Samir Kumar Patra Principal Supervisor
Rohan Dhiman Member, DSC	Rupam Dinda Member, DSC
Sujit Kumar Bhutia Member, DSC	Sreenivasulu Kurukuti External Examiner
Surajit Das Chairperson, DSC	Sujit Kumar Bhutia Head of the Department

Department of Life Science
National Institute of Technology Rourkela

Prof. Samir Kumar Patra
Associate Professor

Prof. Bibekanand Mallick
Assistant Professor

November 2, 2016

Supervisors' Certificate

This is to certify that the work presented in this dissertation entitled *"Profiling of DNA methylation and single nucleotide polymorphism for diagnosis, prognosis and targeting DNA methyltransferases for therapeutic intervention of breast cancer*", by "*Arunima Shilpi*", Roll Number 511LS102, is a record of original research carried out by her under our supervision and guidance in partial fulfilment of the requirements of the degree of *Doctor of Philosophy* in *Department of Life Science*. Neither this dissertation nor any part of it has been submitted for any degree or diploma to any institute or university in India or abroad.

Bibekanand Mallick
Assistant Professor

Samir Kumar Patra
Associate Professor

Dedicated to

My late grandparents, parents and brother

Arunima Shilpi

Declaration of Originality

I, *Arunima Shilpi, Roll Number 511LS102* hereby declare that this dissertation entitled *"Profiling of DNA methylation and single nucleotide polymorphism for diagnosis, prognosis and targeting DNA methyltransferases for therapeutic intervention of breast cancer"* represents my original work carried out as a doctoral student of NIT Rourkela and, to the best of my knowledge, contains no material previously published or written by another person, nor any material presented by me for the award of any other degree or diploma of NIT Rourkela or any other institution. Any contribution made to this research by others, with whom I have worked at NIT Rourkela or elsewhere, is explicitly acknowledged in the dissertation. Works of other authors cited in this dissertation have been duly acknowledged under the sections “Reference” or "Bibliography". I have also submitted my original research records to the scrutiny committee for evaluation of my dissertation.

I am fully aware that in case of any non-compliance detected in future, the Senate of NIT Rourkela may withdraw the degree awarded to me on the basis of the present dissertation.

November 2, 2016 *Arunima Shilpi*

NIT Rourkela

Acknowledgement

I would like to express my sincere gratitude for all support and encouragement from my mentor, colleagues, friends and family throughout the period of my doctoral study.

First and foremost, I would like to thank my thesis supervisor Dr. Samir Kumar Patra for his guidance, inspiration and indispensable advice for my graduate study. I am indebted to his teaching and his profound knowledge and supervision. I am especially grateful for his perspective in experimental as well as in theoretical science, which no doubt shaped my philosophy where I am today. He always encouraged me to be an independent thinker that took crucial part of the research training in Epigenetic and cancer research laboratory.

I am very much thankful to my co-supervisor Dr. Bibekanand Mallick for his support. I am deeply grateful to Dr. Sujit Bhutia, Dr. Surajit Das, Dr. Bismita Nayak, Dr. Rasu Jayabalan and all faculty members for their help and suggestion. I am also thankful to my chairperson and all my DSC members (Dr. Rohan Dhiman and Dr. Rupam Dinda) for their suggestion and support in thesis work.

Further I would like express my gratitude to Dr. Ramana Davuluri, Northwestern University, Chicago, USA, for giving me the opportunity as visiting research scholar for the duration of six months. I am especially thankful to him for his time to time guidance in the field of next-generation sequencing leading to complete genome analysis. I am also thankful to Dr. Yingtao Bi for lending me his expertise and knowledge in the field of computational science and R-programming. I would like convey gratitude to Dr. Segun Jung, for sharing his valuable suggestions to improve my writing skills and his perspective in the field of research. I am also thankful to Dr. Manoj Kandpal for his guidance to access quest and being friendly throughout my stay at Chicago. I would also like to convey my gratitude to Joshua Lamb and Abby Cosentino-Boehm for co-ordaining my visit to Northwestern University.

I convey my sincere regards to Prof. Sunil Kumar Sarangi, Director, NIT, Rourkela, who had been a constant source of inspiration and development of resources to carry out innovative research. I am also thankful to, Prof. Banshidhar Majhi, Dean

(Academic) and all the other members of academic section for their help and suggestions. I am also indebted to TEQIP coordinator and chairperson for providing me the financial assistance to my visit to Northwestern University. I would also like to acknowledge the fellowship offered by NIT-Rourkela, which motivated to actively pursue my doctoral study. I am also thankful to Mr. Kailash Kumar Swain and the high-performance computing (HPC) team for their help in the installation of software required multi-processor computing. I am also thankful to Dr. Vinod Devraji of Bangalore for his assistance in learning Schrodinger software.

I would also like to extend my gratitude to all of my friends and colleagues; Chahat Kausar, Dr. Madhumita Rakshit, Dr. Moonmoon Deb, Dr. Laxmidhar Das, Dipta Sengupta, Swayamsidha Kar, Sandip Kumar Rath, Nibedita Pradhan, Sukanya Pati, Priyanka Saha and Priyanka Chakraborty with whom I shared this incredible journey throughout. My special thanks to Sabnam Parbin for her support in the experimental analysis. She was always available to share the scientific ideas during my stay at NIT, Rourkela. Besides the active participation in research, I also enjoyed my interest in sports which included swimming and other athletic activity. I am thankful to Mr. Santosh Naik for being an excellent coach for swimming.

Last but certainly not least, I am eternally thankful to my parents- mother: Mrs. Rita Verma, father; Mr. Avinesh Kumar Verma and brother; Mr. Abhishek Kumar for their inspiration, love, and sound encouragement throughout my doctoral study. Above all, I would thank almighty for the blessings showered on me.

Arunima Shilpi

Abstract

Breast cancer being multifaceted disease constitutes a wide spectrum of histological and molecular variability in tumors. Now, in the wake-up of the Human Genome Project (HGP) several evidences recommend a marked plasticity adopted by tumor cells in modulating the tissue invasion and progression during multiple stages of metastasis. However, the task for the identification of these casualties in a cancer genome is complicated by the interplay of inherited genetic and epigenetic aberrations. These aberrations are like two sides of the same coin. Therefore, in this thesis we provide an extrapolate outlook to the sinister partnership between genetic and epigenetic aberrations in relevance to breast cancer.

DNA methylation is a prototypical epigenetic parameter that lay ground in understanding the gene regulation and their intricate interactions in the normal and diseased state. However, when it is comprehended by the extensive study of the genomic and transcriptomic parameter, it leads to better understanding of complex trait architecture of disease aetiology. The key to our analysis holds in identification of effective model that enables in predicting the phenotypic traits and outcomes, elucidating the presence of diagnostic and prognostic biomarkers and generate an insight into genetic underpinnings of heritable complex traits. In view of this, we explored the emerging approaches based upon data integration and meta-dimensional analysis to deepen our understanding to the relationship between the genomic variations and human phenotypes. This integrated study comprised of Illumina 450 DNA methylation, Affymetrix SNP array and RNAseq dataset retrieved from the Cancer genome atlas (TCGA) portal which elaborated the biological and complex outlay in the diagnosis, prognosis and therapeutic implications of breast cancer.

Owing to the identification of diagnostic marker, the genetic determinants of DNA methylation pattern was extensively interrogated in tumor and matched normal samples. In lieu of this, an overall enrichment in significant CpG-SNP pairs were identified at 50 base pairs upstream and downstream of CpG site. The correlation between the genetic variant and the differential DNA methylation at specific loci was labelled as methylation quantitative trait loci (meQTLs). In a multistep approach to the identification of key drivers of the complex trait, the differentially methylated CpG sites were analysed for the association with the gene expression in unrevealing the differential expression of the tumor suppressor genes in tumor and matched normal sample. The integrated study of genetic variation characterised single nucleotide polymorphism, DNA methylation and gene expression led to the foundation for identification of novel biomarkers for diagnosis of breast cancer. This integrative analysis was further substantiated with the clinicopathological features to stratify the risk associated with the survival of the breast cancer patients. An intensive Cox proportional regression analysis established a significant association between differential methylation and the stratification of breast cancer patients into high and low risk, respectively. The innovative study interrogating the impact of differentially methylated CpGs and SNPs on the survival unwrapped a new horizon in the prognosis of breast cancer.

In view of established study specifying DNA methylation works in concert with genetic variants, several modulators have been identified against the DNA methyltransferase (DNMTs) enzyme to revert malignancy. However, the inherited toxicity and the lack of specificity offer limitations. In the present study, we have identified a novel inhibitor that owes property to rejuvenate the expression of tumor suppressor genes and holds enhanced selectivity towards triple-negative breast cancer cells to normal cells. Thus, the recognition of DNA methylation as a significant contributor to normal and disease state has opened a new avenue for drug discovery and therapeutics in breast cancer.

Keywords: DNA methylation; single nucleotide polymorphism; methylation quantitative trait loci; DNA methyltransferases; inhibitor; breast cancer

Contents

List of figures

List of tables

Abbreviations

HGP	:	Human Genome Project
TCGA	:	The Cancer Genome Atlas
BRCA	:	Breast invasive carcinoma
SNP	:	Single nucleotide polymorphism
meQTL	:	Methylation quantitative trait loci
eQTL	:	Expression quantitative trait loci
ST5	:	Suppression of Tumorigenicity 5
CMAH	:	Cytidine monophosphate-N-acetylneuraminic acid-hydroxylase
FYN	:	Tyrosine kinase
ADAM8	:	A disintegrin and metalloproteinase domain 8
CREB5	:	cAMP responsive element binding protein 5
EXPH5	:	Exophilin 5
DNMT	:	DNA methyltransferases
SAM	:	*s*-adenosyl-L-methionine
SAH	:	*s*-adenosyl-L-homocysteine
NLS	:	Nuclear localization sequence
RFC	:	Replication foci targeting domain
BAH	:	Bromo homology domain
ER	:	Estrogen receptor
PR	:	Progesterone receptor
HER2	:	Human epidermal growth factor receptor
TN	:	Triple negative
GWAS	:	Genome-wide association studies
EWAS	:	Epigenome-wide association studies
NGS	:	Next generation sequencing
DMRs	:	Differentially methylated regions

RNAseq	:	RNA sequencing
ANOVA	:	Analysis of variance
HWE	:	Hardy-Weinberg Equilibrium
KM plot	:	Kaplan-Meir plot
HR	:	Hazard ratio
CI	:	Confidence interval
EGCG	:	Epigallocathechin-3-gallate
ChEBI	:	Chemical Entities of Biological Interest
PDB	:	Protein data bank
UniProt	:	Universal protein knowledgebase
CHARMm	:	Chemistry at Harvard molecular mechanics
PLP	:	Piecewise linear potential
PMF	:	Potentials of mean force
LC_{50}	:	Lethal constant
IC_{50}	:	Inhibition constant
MD	:	Molecular dynamics
RMSD	:	Root mean square deviation
RMSF	:	Root mean square fluctuation.

Notations

r	:	correlation coefficient
kcal/mol	:	kilocalorie per mole
kJ/mol	:	kilojoule per mole
h	:	Hour
°C	:	Degree celsius
%	:	Percentage
µM	:	Micromolar
nm	:	nanometer
mg	:	milligram
µg	:	microgram
bp	:	base pairs
π	:	Pi
Å	:	Angstrom
ps	:	picosecond

Chapter 1

Introduction

1.1 DNA methylation landscape in human genome

DNA methylation is relishing a meteoric rise in the field of epigenetics from the euphoria surrounding the human genome project. This field of epigenetics holds a master key to unfold and unlock the mechanism concomitant with the profound alteration in gene expression in response to the environmental cues [1]. It provides a clue in understanding the tenacity and the genome plasticity associated with chromatin modifications and remodeling engines. Most of these epigenetic modulations known till date is characterized by the covalent and non-covalent modulation of DNA and histone proteins [2]. Of all the modulations, DNA methylation is a core molecular actor that play significant role upon the epigenetic stage influencing the epigenetic stability and heritability and subsequently retaining the integrity of the DNA [3].

In the mammalian genome the primary target for methylation is the cytosine residue; the enzymatic attachment of the methyl to the 5' carbon of the pyrimidine ring creates 5-methylcytosine (5-MeC) [4-6]. This forgotten 5^{th} base being a cognate to cytosine undergoes complementary base pairing with guanine. Usually in mammalian genome, the targeted cytosine residue of methylation machinery resides within the palindromic sequence of the 5'-C-p-G-3' dinucleotide. Nearly 70% of all CpG dinucleotide are methylated; however, the spatial distribution is non-random across the genome. Besides the irrational distribution, there is a small genomic region bearing the higher frequency of CpG dinucleotide at the closer proximity with an average of 1-2 kilobases forming CpG islands [6]. There are about 45,000 CpG islands. Most of the chromosome harbours 5-15 islands per MB being predominant at the promoter region of the genes or lie within the first exon of the genes [7]. These sporadic sequences associated with the epigenetic pattern have the maximal impression on growth and development.

1.2 Significance of DNA methylation

The functionality DNA methylation is integrated to regulate the gene expression in terms of positive correlation between the extent of methylation, transcriptional and recombinational quiescence. This correlation is most conspicuous in transposable elements prevalent across the mammalian genome. It maintains the host defense system

through the transcriptional silencing of these parasitic elements which is a threat to the structural integrity of the genome [8, 9].

The hypermethylation of bulk DNA holds the functional standpoint in the assembly of repetitive DNA into a heterochromatin which maintains the functional compartmentalization of genome into its active and inactive state [10]. While the primordial germ cells and the embryonic stem cells progress with the mitotic division without detectable DNA methylation, the cellular differentiation initiates with DNA methylation [11, 12]. Much of these cellular differentiations are established during the gastrulation stage of embryonic development.

DNA methylation has significant application in the somatic lineage of genes in genomic imprinting in-lieu of embryonic development and physical requirements [13]. The genomic imprinting is characterized by monoallelic or the uniparental expression of genes in the somatic cells [14]. These imprints are transmitted as unique methylation pattern of imprinted genes to the gonads during gametogenesis and after fertilization persists in the somatic cell. The acquisition and propagation of imprinted genes carrying differential methylation pattern play an intrinsic role in mammalian development [15]. Besides, the differential methylation also guides to the transcriptional silencing of the majority of genes on one of the two X-chromosome in each somatic cell of the female. During the early embryonic development, one of the two X-chromosome is randomly selected for inactivation; also an example of parental imprinting [16, 17].

1.3 Catalytic mechanism of DNA Methylation

The chemistry associated with the cytosine methylation hovers around the activity of the enzyme DNA methyltransferases (DNMTs) and the cofactor S-adenosyl-L-methionine (AdoMet), the source for a methyl group [18-20]. This enzymatic reaction brought about by DNMTs implicates via covalent mechanism coupled with acid/base catalysis. In the presence of the nucleophilic addition to the enzyme, the methyl-sulphur bond of AdoMet is destabilized which in turn renders the methyl group to the C5 position of cytosine molecule via the S_N2 mechanism [21]. The stepwise mechanism initiates with transient covalent bond formation between C6 of the target cytosine and thiol group of Cysteine residue (Cys) forming a 6-Cys-S-cytosine adduct [22]. This nucleophilic addition at C6 carbon is expedited by transient protonation of glutamic acid residue at N3 of cytosine establishing 4-5 enamine structure [5]. Thus, the stable covalent bond elevates the electron density at C5-postion promoting the displacement of methyl moiety of AdoMet molecule to provide 5-CH_3-6-Cys-S-5 forming 6-dihydrocytosine complex [23, 24]. Finally, the deprotonation at C-5 position departs the cysteine residue subsequently resolving the covalent intermediate into methylated cytosine and *s*-adenosyl-L-

homocysteine (AdoHcy) as a by-product [25]. The detailed mechanism is elaborated in Figure 1.1.

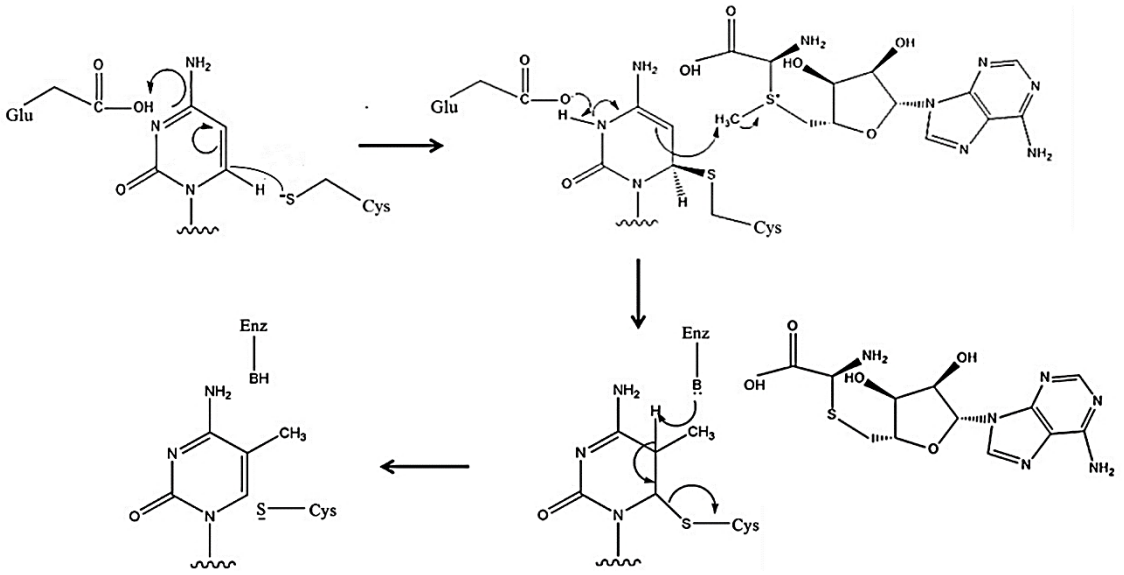

Figure 1.1 Mechanism of DNA methylation: Motif IV of the enzyme active site constitute Cys residue such that the thiol attacks at C-6 of cytosine molecular which results in electron cloud at 5-C. Simultaneously, proton donation (H^+) by Glu-COOH apparently stabilizes the transition state. In step 2, 5C carbanion of cytosine molecule attaches at $–CH_3$ of SAM forms an intermediate complex. In step 3, the abstraction of the proton from the enzyme followed by the β-elimination results in the formation of 5C=6C double bond. In step 4, the methylation group is attached to 5'-C forming stable complex and the enzyme in released by proton addition.

1.4 DNA methylation machinery

The cellular DNA methylation is established and maintained by the complex interplay of family of dedicated enzymes, called DNA methyltransferases (DNMTs) [26, 27]. These DNMTs constitute four members being grouped into two families having discrete structure and function. DNMT1, being the maintenance methyltransferase duplicates the existing methylation mark on the daughter strand of hemimethylated DNA successfully propagating across the successive generation [28] while, DNMT3 family actively participates in *de-novo* methylation during embryonic development [29-31]. This DNMT3 family constitute two active members as DNMT3A and DNMT3B and a regulatory component as DNMT3L [29, 32]. The structural machinery of the active members is integrated into the regulatory domain (N-terminal) and the catalytic domain (C-terminal) exclusively dependent on each other. The catalytic domain establishes nine out of ten conserved motifs being crucial for its function. Topologically the catalytic domain is grouped into two sub-domain [33]. The first half of the domain constitute structurally conserved motifs I-III which enables in co-factor (AdoMet) binding while,

the conserved motifs IV-VIII along with the partner domain is predominantly responsible for the catalytic mechanism [34]. The target cytosine binding site is enclosed within the conserved motif IV (ProCysGln), VI (GluAsnVal) and VIII (GlnXArgXArg) [26, 35, 36].

The large N-terminal domain bearing two glycine-rich loops is implicated in sequence-specific DNA recognition by DNMTs and flipping of target cytosine [37]. This terminal is accreted with the multi-functional domains; the nuclear localization sequence (NLS) domain that escorts in translocation of DNMT1 into the nucleus, replication foci targeting (RFT) domain enriched in glycine residue that recruits DNMTs to replication foci of DNA, the cysteine-rich (CXXC) domain also referred as zinc binding domain that forms interface for binding of unmethylated DNA and the two bromo-homology domain (BAH1 and BAH2) actively involved in protein-protein interaction thus, regulating the chromatin structure [27, 38, 39]. While the catalytic domain is conserved across the DNMTs, the N-terminal domain of DNMT3A/B contains a PWWP (pro-trp-trp-pro) that is functionally significant in non-specific binding with DNA [Figure 1.2] [37, 40, 41].

The subsidiary DNMT3L shares homology with DNMT3A and DNMT3B in both N and C-terminal domains while, it is deficient in conserved amino acid sequence prerequisite to catalytic activity. It is specifically expressed in germ cell and is essential for the establishment of a subset of methylation pattern in both male and female germ cells [42]. DNMT2 exemplified by divergent evolution shares structural homology with known DNA Mtase and its functionality corresponds to cytosine methylation of the anticodon loop of tRNA [43]. Structure elucidation of DNMTs is of considerable interest as its inhibition results in subsequent restoration of aberrantly silenced tumor suppressor genes in cancer.

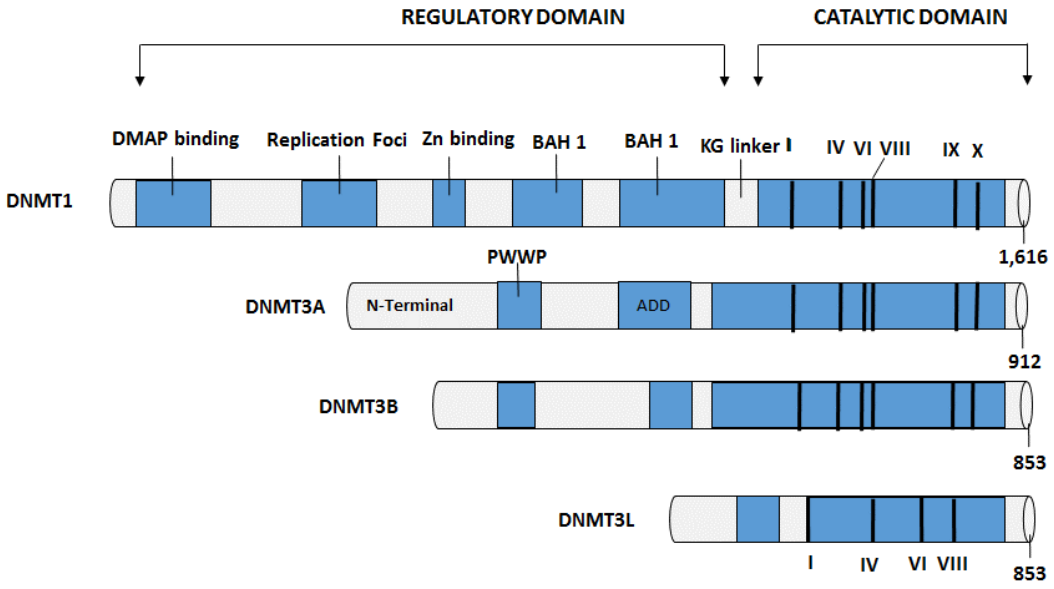

Figure 1.2 Architecture of DNMT1, DNMT2, DNMT3A, DNMT3B and DNMT3L regulatory and the active site domain. Abbreviations: DMAP: DNMT1-associated protein, BAH1: Bromo adjacent homology domain, PWWP: Pro-Trp-Trp-Pro, ADD: ATRX-DNMT3-DNMT3L (related to the plant homology (PHD)-like domain of regulator ATRX); KG linker: Consists of Lys and Gly residues.

1.5 DNA methylation profiling in cancer

The mechanism of gene silencing induced by DNA methylation includes direct inhibition of transcriptional activity by blocking the binding of transcription factors to the methylated sites. In another method, methyl-CpG binding proteins (MBDs) recognizes the methylated DNA and recruits corepressors (HDAC) resulting into compact chromatin structure leading to gene silencing [44, 45]. Gene silencing is characterized by the unique profile of aberrant DNA methylation in different types of cancer. Hence, a myriad of biomarkers based on DNA methylation need to be identified for variable classes of neoplasia [46]. The variable methylation pattern in association with biomarkers is identified both in localized regions and across genome-wide offers platform in the diagnosis, prognosis, therapeutic implications and post-therapeutic monitoring. Aberration in DNA methylation is visualized in the early events of carcinogenesis, some being localized in precancerous lesions [47]. DNA methylation being active readout can be easily identified in tumors with low purity. Moreover, only small fraction of promoter regions of aberrantly methylated genes can be directly correlated with cancer initiation and progression [48, 49]. These epigenetically silenced domains carry a majority of methylated genes actively participating in cancerous stem cell progression [50-52]. In general, the aberration in DNA methylation occurs in higher percentage in tumors as compared to genetic variations, resulting in higher sensitivity.

1.6 Techniques for DNA methylation profiling

The variability in methylation pattern among the cell types and during development or diseases and sometimes in response to environmental cues offers considerable theoretical and technological challenges in comprehensive genome-wide mapping [53]. The standard molecular biology techniques such as cloning and polymerase chain reaction offer limitation as it wipes out the DNA methylation information. Moreover, the standard hybridization technique cannot detect the methyl group being located in the major groove of DNA. Thus, the methylcytosine pre-treatment method was developed to reveal its presence or absence both in the localized regions and genome-wide at cytosine residue. There are methods constituting methyl-sensitive restriction enzyme digestion, affinity enrichment and bisulphite conversion [54, 55]. These technologies based approaches are based upon their ability in discriminating the methylated to unmethylated cytosine. Once

the genomic DNA has undergone one of these methylation-dependent steps, molecular biology techniques, including sequencing and probe hybridization can be implemented to reveal methylated-cytosine loci. Finally, several computational methods and software tools applications can be applied for analysis and interpretation of DNA methylation profile [56]. Thus, in the plethora of techniques for determining DNA methylation and profiles is a consequence of the conjoint analysis of pre-treatment and analytical steps [57-59]. The following section details about the methods for DNA methylation profiling.

1.6.1 Methylation-sensitive endonuclease digestion

Methyl sensitive restriction endonuclease treatment is a powerful tool in the discovery of methylation marker associated with targeted candidate genes as well as systemic genome scanning [60]. There are sequence specific restriction enzyme having particular recognition for methylated CpG regions while, some of them are being inhibited from restriction digestion by 5meC. Some of these methylation-sensitive restriction enzymes for DNA methylation studies are *HpaII* and *SmaI* such that each of this constitute isoschizomer and neoschizomer that are not inhibited by CpG methylation [61, 62]. Beside genome-wide studies, the method is also applicable for locus specific analysis having linkage with DNA methylation across multiple kilobases. This methyl-sensitive restriction digestion is followed by PCR, gel electrophoresis and hybridization on Southern Blotting [63, 64]. However, this method has some limitation as many a times the incomplete digestion results in a false-positive result.

1.6.2 Affinity purification of methylated DNA

Recent advancement in high-throughput technology constitutes protein affinity for the identification of methylated fraction of genomic DNA [65-67]. These methylated fragments are purified either through immunoprecipitation (MedIP26) by using an anti-m5C antibody or by DNA binding domain specific to methyl-CpG-binding protein (MAP27) [68]. These methods are specific to high-density DNA methylation constituting enriched methylated CGIs. Recently, affinity purification using CXXC (CAP; X: any residue) have been introduced in specific to unmethylated DNA [69]. However, the uneven distribution of methylated cytosine or CpG sites offers limitation in terms of the composition of an array for hybridization as a consequence of which individual CpG sites cannot be identified [68].

1.6.3 Bisulphite sequencing of methylated DNA

The analysis of DNA methylation on treatment with sodium bisulphite spurred a revolution in the epigenome-wide associations study (EWAS) of the methylation pattern. The treatment with bisulphite differentially selects cytosine to 5-methylcytosine residues

that are deaminated to yield uracil and are amplified as thymine during PCR [70, 71]. This bisulphite treated DNA can be identified by methylation specific PCR [72], restriction digestion [73], or DNA sequencing [74]. In comparison to other methods, sequencing of subcloned bisulphite converted DNA is more reliable for the detailed study of methylation pattern associated with each CpG sites across the genome. Further ahead, it provides an explicit method for determination of methylation pattern for haplotypes in a qualitative and quantitative manner. Besides, the synergistic application of bisulphite conversion with sequencing aids in the genome-wide study of methylation pattern without being restricted by the presence of restriction enzyme or high CpG density. Genome-wide processing of bisulphite treated DNA follows several steps.

The bisulphite treated DNA results in the conversion of the majority of unmethylated Cs to Ts in the sequencing reads. The absolute DNA methylation level is calculated in terms of percentage of recurrence Cs and Ts frequency in the sequencing reads being aligned to the reference genome. Alignment of these reads is brought about by two alternative approaches. The wild card aligners BSMAP21 [75], RMAP25 [76], RRBSMAP26 [77], Methy-Pipe [78]) replaces Cs is the genomic DNA to wild letter Y which in turn matches to both Cs and Ts in the read sequence. In contrary, the three-letter aligners (Bismark28 [79], MethylCoder32 [80]) converts all Cs into Ts in the reads as well in the genomic DNA sequence. Once the alignment is done, the absolute methylation is determined in terms of frequency of alignment of Cs and Ts to each C in the genomic DNA sequence.

Once the data processing and normalization is accomplished, the next step constitutes visual inspection of methylated regions. The big-Bed format prompts in dynamic visualization of DNA methylation which is based on the colour coding of each CpG site [81]; while, big-Wig format represents methylation level of single CpG sites in terms of variable heights of interspersed vertical bars [82]. These binary files are then uploaded to the web-based genome browser mainly UCSC [83], Ensemble [84] or Human Epigenome [85]for visualization. These genome browser prompts in regions-specific visualization while, the global methylation pattern can be visualized through box plots, Hibert plot [86], scatter plots or tree-like diagrams. R/Bioconductor provides an interface for these plot constructions [87]. Mapping of genome-wide methylation pattern between the groups of samples helps in visualization of systemic differences between the tumor patients and healthy control group. Finally, statistical significance between differential methylation groups can be verified and validated through volcano plots, Q–Q plots or Manhattan plots [88].

1.6.4 Array hybridization

Array-based analysis of methylation pattern is coupled with enzymatic methods. The differential methylation sensitive and cutting of behavior of *SmaI* and *XmaI* is followed by methylated CpG island amplification (MCA) [89]. This method is further associated with representational difference analysis (RDA) or array hybridization [90]. However, the process based upon MCA is significant as it provides coverage of lower resolution. In an alternative approach, differential methylation hybridization (DMH) is based upon restriction digestion of pool of genomic DNA by methylation-sensitive restriction enzymes and mock digestion of another pool [91]. Consequently, the parallel pool of DNA is produced which is amplified and labeled with fluorescent dyes of cyan/red array hybridization [92]. The relative signal intensities of fluorescent dye are used to detect locus- specific DNA methylation. This method is referred as the microarray-based assessment of differential methylation pattern [93].

1.6.5 Next Generation Sequencing

Next-generation sequencing (NGS) offers a platform for harnessing massive-parallel short-read DNA sequences to digitally catechise genome-wide DNA methylation. Several NGS platforms developed so far constitute 1) 454 GS20 pyrosequencing (Roche Applied Science), 2) Solexa sequencing (Illumina) and 3) Supported Oligonucleotide Ligation and Detection: SOLiDTM (Applied Genes) [94-97]. These methods are based upon fundamental principle of immobilization of template DNA to solid surface and parallel sequencing of clonally amplified or single DNA template as a consequence of which thousand to billions of sequence reads are generated in single run [98, 99]. This technology has enhanced drastically thus, reducing the sequencing cost per base and enables genome-wide bisulphite sequencing of DNA methylation pattern in high throughput at a single base resolution in a very short span of time [100]. The large data generated are being co-ordinated by national and international consortia (The Cancer Genome Atlas, TCGA) for data analysis [101]. NGS is advantageous over microarray as it provides higher base resolution with relatively small artifacts such as noise in the form of cross hybridization without any limitation in the genome coverage. Moreover, larger dynamic range and high-coverage increases the efficiency of resultant data [102]. Thus, high throughput analysis based on NGS can be successfully implemented in identification of methylation signatures for diagnosis and prognosis of cancer.

The quantitative based analysis based upon above-mentioned approaches supersedes over the non-quantitative method for detection of aberrant methylation pattern in the clinical settings [96, 103]. These methodologies are even compatible with degraded DNA. Many types of cancer display variability among the patients with similar histopathology and disease stage. Technology based on the high-throughput analysis can

be implemented in molecular characterization of variable grades of a tumor. The digital-based approach in NGS will promote in early detection with minimal methylated residues in biomarker discovery. Finally, with the recent advancement in the technology, DNA methylation has undergone a revolution in the diagnosis, prognosis, therapeutic and post-therapeutic implications of cancer.

1.7 DNA methylation as therapeutic target in cancer

There are plethora of genes and pathways being regulated by DNA methylation. It serves as the biomarker in the restoration of aberrantly silenced genes in cancers [104, 105]. These methylation patterns can be monitored by the introduction of several chemotherapeutic agents or epi-drugs. Epi-drugs can be defined as the modulators that can inhibit or activate epigenetic proteins associated with amelioration, cure or prevention of diseases [106, 107]. The expression of these epigenetic proteins is altered in many human diseases primarily in cancer. These alterations in protein expression are visualized in an early stage of cell transformation; thus, they can be considered as drivers rather than passengers in cancer [108, 109].

Multiple inhibitors targeting DNMTs are deemed to be the most putative anticancer agent having the ability to revert the aberrant methylation pattern at the promoter region of tumor suppressor genes. These DNMTs co-ordinates in mRNA expression in normal tissue and are overexpressed in tumors [110]. The elevated expression has been reported in cancers of the liver, colon, prostate, breast cancer and leukemia. Thus, inhibitors against DNMTs promise anticancer agents as they restore the expression of epigenetically silenced tumor suppressor genes in these cancers. Two such FDA approved nucleoside DNMT inhibitors, 5-azacytidine (Vidaza) [111] and 5-aza-2'-deoxycytidine (Dacogen) [112] had been reported to be effective in the treatment of bone marrow disorder in myelodysplastic syndrome. These inhibitors get incorporated into DNA in place of cytosine. 5-aza-2'-deoxycytidine (decitabine) when, co-administered with carboplatin reverses the platinum resistance in ovarian cancer-promoting in prolonged progression-free survival [113]. These inhibitors have been identified in activating the dormant gene expression of the *p16* gene subsequently, decreasing the growth of cancer cells [114]. Besides regulation of gene expression through DNMTs, these nucleoside inhibitors also get incorporated into RNA thus, inducing ribosomal disassembly and preventing the expression of oncogenic proteins at the translation level. However, these nucleoside inhibitors offer some limitations [115, 116]. The ability of these inhibitors to get incorporated into DNA and RNA arrests the cell cycle forming DNA/RNA covalent protein-adduct is toxic [117]. Moreover, in aqueous solution these inhibitors are readily hydrolysed by cytidine deaminase. Thus, toxicity and instability of these inhibitors inevitably presents a challenge to their applications clinically.

1.8 DNA methylation in breast cancer

Most of the epigenetic studies unravels the hypothesis behind the disease predisposition is a consequence of the mismatch between prenatal and postnatal environment [118]. This epigenetic mismatch because of DNA methylation is widely associated with the developmental origin of health and diseases mainly the non-communicable diseases such as diabetes, cardiovascular and neuro-developmental disorders [119]. Of all the diseases known till date, cancer remains elusive, and it is widely accepted that the co-ordinated effect of genetic and epigenetic disorders leads to cancerous state [120, 121].

DNA methylation characterized by genome-wide hypomethylation of sparsely populated CpG sites in intergenic and repetitive sequences and hypermethylation of densely packed CpG islands at promoter regions leads to cancer [122]. Hypomethylation of the repetitive sequences primarily in the transposons causes genomic instability and DNA breakage, and the intergenic region of chromatin undergoes de-condensation [123]. In many cases, hypomethylation also results in loss of imprinting or demethylation of retrotransposon leading to cancer [124-127]. On the contrary, the hypermethylation of tumor suppressor genes at the promoter region leads to somatic aberrations in cancer [128]. The driving force associated with cancer is mainly focused on promoter hypermethylation of CpG islands as it clearly demonstrates the permanent gene silencing both physiologically and pathologically. This anomaly in gene silencing compels in the aberrant clonal expansion of cells subsequently fostering to tumor progression [Figure 1.3] [129].

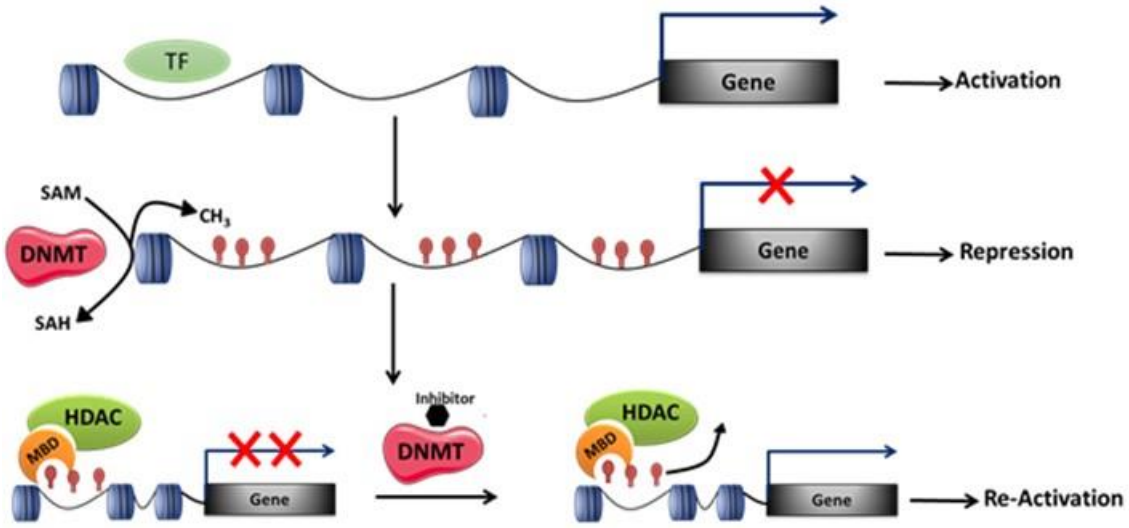

Figure 1.3 DNA methylation-mediated gene silencing in cancer.

Of all the cancers were known till date, breast cancer occupies the top most slot in morbidity and mortality of women in developed countries while, the developing countries are on a rise. In 2014, the invasive nature of breast cancer accounted for 232,670 newly-diagnosed cancer cases and 40,000 cancer death in women in USA[130]. This high mortality rate is explicated by the histological and morphological heterogeneity of the disease. According to World Health Organization (WHO), the standard classification of breast cancer defines 18 different histological types [Table 1.1] [131, 132]. This histological variability contributes to the differences in prognosis and target-specific response in chemotherapy. Many a times these tumors offer resistance to drug treatment, as a consequence of which the disease relapse.

Subsequent studies have classified this heterogeneous group of disease into a spectrum of subtypes having distinct genotype and phenotype. This classification system is based upon presence of estrogen receptor (ER+), progesterone receptor (PR+) and human epidermal growth factor receptor 2 (HER2+); however, their mere absence results in triple negative breast cancers (ER/PR/HER2-) [133, 134]. Based on the presence of these receptors, patients are grouped into four major sub-groups of Luminal A (ER+ and/or PR+, HER2-), Luminal B (ER+ and/or PR+ , HER2+), HER2 (ER-, PR-, and HER2+) and triple-negative (ER-, PR-, HER2-) [135, 136]. Over last decade, several efforts have been sought to improve this stratification of breast cancer; however, there are still in the subject of controversy. Increasing shreds of evidences have substantiated the critical role of epigenetic deregulation in the early event of carcinogenesis and subsequently prompts to assess the epigenetic cause of breast cancer [108, 137]. More significantly, methylation signatures are regularly employed in the stratification of breast cancer patients in diagnosis and prognosis [138]. With recent advancement in technology like microarray and next generation sequencing associated with genome-wide DNA methylation profiling will guide new pavement in better understanding of breast cancer etiology [139].

Table 1.1 Histopathological types of invasive breast carcinoma (Adapted from Weigelt *et.al*, 2005)

Histological types of invasive breast Carcinoma	**Frequency**	**10-year survival Rate**
Invasive ductal carcinoma	50-80%	35-50%
Invasive lobular carcinoma	5–15%	35–50%
Mixed type, lobular and ductal features	4–5%	35–50%
Tubular/invasive cribriform carcinoma	1–6%	90–100%

Mucinous carcinoma	<5%	80–100%
Medullary carcinoma	1–7%	50–90%
Invasive papillary carcinoma	<1–2%	Unknown
Invasive micropapillary carcinoma	<3%	Unknown
Metaplastic carcinoma	<5%	Unknown
Adenoid cystic carcinoma	0.1%	Unknown
Invasive apocrine carcinoma	0.3–4%	Unknown
Neuroendocrine carcinoma	2–5%	Unknown
Secretory carcinoma	0.01–0.15%	Unknown
Lipid-rich carcinoma	<1–6%	Unknown
Acinic cell carcinoma	7 cases	Unknown
Glycogen-rich, clear-cell carcinoma	1–3%	Unknown
Sebaceous carcinoma	4 cases	Unknown

1.9 Work done so far in diagnosis of breast cancer

The statistics of breast cancer is startling and calls for early diagnosis. Multifactorial etiology is characterized by constellations of risk factors. These risk factors are concomitant with genetic and epigenetic predispositions, loss of host immunological defense, viruses as well as other carcinogens. Hormonal imbalance in estrogen is considered to be one of the most significant promoters of carcinogenesis [140]. Despite the ongoing research in finding the cause of breast cancer, this avenue does not hold great promise in the scenario of combating this deadly disease. Besides, finding the cause, the most important aspect is the early diagnosis of the disease such that the prognosis for a cure will guide into appropriate therapeutic interventions.

1.9.1 Methods for early diagnosis of breast cancer

The association between survival and stage of disease diagnosis are two concomitant aspects in disease cure. If a patient is diagnosed in its early stage of tumor proliferation; an appropriate therapy and medication will lead to the long-term survival. As per the instruction of physician, the art of periodic breast examination preferentially in the patients with an increased risk (family history of breast cancer) will be judicious in early diagnosis and very often highly curable [141]. This early diagnosis can be associated with factors such as the common type of breast lesions, recurrence of such lesions, characteristic symptoms and family history. The most common lesions in women are characterized by fibrocyst, fibroadenoma, intraductal papilloma and duct ectasia while, in

men gynecomastia is more predominant [142]. While monthly breast examination is of great importance in early diagnosis, it can only identify palpable lesions. However, techniques based upon X-rays such as mammography or xerography can detect in the preclinical stage before lesions enter into the clinically palpable size. Thus, breast X-rays are implemented for identification of clinical lesions benign or malignant state. However, a great deal of concern is associated with exposure to radiation by X-rays. Moreover, women with high breast density are sensitive to mammography, resulting in only 24-46% of the detection of malignancies [143]. Similarly, Magnetic Resonance Imaging (MRI) has been useful in detecting aberrations associated with benign and malignant lesions but, its poor specificity results in unusual breast biopsies and associated uncertainties [144-146]. Thus, the methods for early detection need to be fortified by the advent of molecular technologies related to cellular changes in genome or proteome. Since last decade, there had been a substantial advancement in biomarkers discoveries, having a decisive role in understanding the cellular and molecular mechanism oftransformation of the normal cell to a malignant state.

1.9.2 Diagnosis based upon biological marker

Biological markers offer a way around to the hurdles in this era of genomic medicine. These markers are characterized by an indicator that can measure normal biological process, pathogenic or pharmacological process in response to therapeutic interventions. It can be instigated at any stage of disease diagnosis, prognosis or predictive outcome. These biomarkers can also be associated with changes in the environment and are referred as exposure biomarkers [147]. Thus, biomarkers antecedent to the disease are influenced by both genetic and epigenetic variations. Further ahead, these markers can be implemented in the stratification of individual based upon associated risk or prognosis and can be a surrogate endpoint in clinical trials [148, 149]. An ideal biomarker must compliment with clinically relevant information ideally across multiple individuals and populations. Typically, a molecular marker in breast cancer are obtained from breast epithelial cells which include primarily ductal lavage, periareolar fine-needle aspiration, fine needle biopsies or core-needle biopsies [150]. Herein we elaborate about the genetic and epigenetic biomarker primarily DNA methyl markers known till date in breast cancer.

1.9.3 Diagnosis based upon genetic markers

The autosomal inheritance of dominant allele exemplifies significant predisposing factor in 10% of women with breast cancer. *BRCA1* and *BRCA2* are identified to be the most susceptible genes linked to germline mutation and hereditary cause in most of the women. Women having mutation in either of these genes are associated with cumulative lifetime risk of 60-80% in development of breast cancer [151]. Understanding the normal

biological function and regulations of these two genes will lead to the study of molecular basis of heredity and will provide new driving force in disease diagnosis and therapeutic strategy. The functional characterization of these genes constitutes the maintenance of genome integrity by compromising unusual loss, duplication or chromosomal rearrangement of DNA. The developments of breast rely upon estrogen and progesterone for growth, differentiation, and homeostasis [152]. The inactivation or mutation of these genes results in estrogen-induced DNA damage. Thus, DNA damage results in error-prone DNA repair leading to global genomic instability and concomitant accrual functionality leading to tumorigenesis. Mutation in *BRCA1* and *BRCA2* causes repression of transcriptional activity of estrogen and progesterone receptors leading to the unusual proliferation of the epithelial cell and altered the hormonal response. Thus, the study of mutations associated with *BRCA1* and *BRCA2* is identified to be beneficial in diagnosis and treatment of breast cancer patients [153, 154]. Being a caretaker of genome integrity, it has been recognized as a prime target for therapeutic interventions. These genes also unfold the risk associated with the genetic context in different populations and historical groups. However, the inconsistency in mutation prevalence and penetrance brings about controversies in understanding the risk associated with each patient [155, 156]. Penetrance is defined as the percentage of individuals carrying particular variant of a gene may be associated with risk for cancer predisposition [157, 158]. Some of these genes having high penetrant include the following.

TP53 being tumor suppressor gene plays significant role in the regulation of cell growth. Germline mutation associated with this gene results in a spectrum of malignancies including sarcoma, adrenocortical, sarcoma and leukemias. Females carrying TP53 have a higher frequency of malignancy and susceptibility to Li Fraumeni syndrome [159]. Besides, Phosphate tensin (*PTEN*) homologs have been identified to be actively participating in phosphatidylinositol-3-kinase (*PI3K*) phosphatase activity [160] [161]. However, the dysfunction associated of this gene leads to cell cycle arrest, apoptosis, and anomalous cell survival. Germline mutation in PTEN results in Cowden syndrome (CS) and is characterized by multiple hamartomas and elevated malignant transformation [162]. In breast cancer, 50% women at an average age of 36-46 years are diagnosed with CS. Frequency of this multifocal and bilateral disease has been identified to be elevated in patient with ductal adenocarcinoma. More than 67% of women bearing CS are also associated with benign breast diseases, such as adenosis, adenosis, fibroadenomas and apocrine metaplasia [163]. Besides genes having high penetrant, there are some genes associated with moderate penetrance and the risk associated varies from 1.5 to 5. Some of these genes and the associated risk are described in the following section.

Checkpoint kinase 2 (*CHEK2*) encoding for serine-threonine kinase are activated in response to damage caused by double-stranded DNA breaks (DSBs) and transmits signal for the repair of the proteins and the processors in the downstream [164]. It also phosphorylates *BRCA1*, expediting its role in DNA repair. However, female carrying *CHEK2* mutations in the homozygous state have six-fold increase in bilateral or recurrent breast cancer [165]. Similarly, BRCA1-associated RING domain 1 (*BARD1*) encodes for a protein having structural and functional homology with BRCA1 [166]. Mutations in these genes are deleterious and results in aberration in DSB repair and arrest of apoptosis. However, the mutation in *PALB2* gene conjointly with *BRCA2* interrupts in DNA repair mechanism, leading to tumor suppression [167]. Some of the moderate penetrate genes having frequent mutation includes, *ATM*, *RAD51C*, *MRN* complex and others [168-170].

In short, the mutation associated with the genes mentioned above results in successful screening and detection of malignancy, however, the complexity associated with its outcome, requires clarification as to whether these mutations act as driver or accelerator. Thus, the selection of the appropriate biomarker for particular settings and cohorts is very essential. Finally we would like to say that the selection of the most promising biomarker for specific settings and cohorts will lead to enhanced diagnosis.

1.9.4 Single nucleotide polymorphism in breast cancer predisposition

Identification of genetic risk associated with the allelic polymorphism, either at a single locus or epistatic effect will promote in screening and stratification of breast cancer. Several studies have revealed the presence of SNPs in association with DNA repair genes (*PALB2*, *BRIP1*, *CHEK2*, *ATM*, and *RAD50*) which can be implemented in screening and stratification of breast cancer. Genome-wide studies reveals the presence of SNPs associated with genes and loci (*LSP1, TOX3, FGFR2, TGFB1, MAP3K1, 2q35, and 8q loci*) in Ashkenazi Jewish ancestry accounting 80-90% of hereditary breast cancer [171]. Similarly, studies conducted by *Johnson et al.* in 437 patients bearing primary breast cancer disclosed the presence of 25 SNPs in association with *BRCA1*, *BRCA2*, *ATM*, *TP53* and *CHEK2* genes. Some of these SNPs associated with high risk include; (rs1799950, rs4986850, rs22279945, rs16942, and rs1799966) of *BRCA1* gene, (rs766173, rs144848, rs4987117, rs1799954, rs11571746, rs11571747, rs4987047, rs11571833 and rs1801426) of *BRCA2* gene, (rs3218707, rs4987945, rs4986761, rs3218695, rs1800056, rs1800057, rs3092856, rs1800058 and rs1801673) of *ATM* gene, rs1787991 of *CHEK2* gene and rs1042522 of *TP53* gene [172]. Similarly, *Cox et al.* genotyped 9 SNPs and found evidence for association of rs1045485 with *CASP8-D302H* in breast cancer [173]. Stacey et al. studied approximately 300,000 SNPs in 1,600 individuals from Iceland and reported the presence of rs13387042 and rs3803662 on chromosome 2q35 and 16q12 respectively [174].

Zhang et al. studied the regulatory regions of *ER-α* gene in Chinese population and identified rs379857 to be predominant in 300 breast cancer patients [175]. Similarly, *Abbasi et al.* studied 150 Iranian patients and identified 3 SNPs associated with codon 10, 235 and 594 in *ER-α* gene and other being associated with codon 392 of *ER- β* gene having additive effect in development of breast cancer [176]. *Hosseini et al.* also reported the elevated expression of *ERα* gene and down-regulation of *ERβ* gene in cancer tissues, having a significant role in breast cancer development. Thus, SNPs regulating the expression of genes in breast cancer can be a potential tool for diagnosis and therapeutic implications. Although the success of GWAS endorses the identification of genetic variants asociated with diseases, however, it requires complementary approaches for answering the unidentified variants.

1.9.5 DNA Methylation: An epigenetic biomarker in diagnosis of breast cancer

Several studies conducted so far elucidate the gradual transformation of methylation pattern in normal, pre-malignant and malignant breast tissue. This gradation in methylation pattern promotes in early detection and classification of breast cancer subtypes. Hypermethylation in the promoters region of tumor suppressor genes primarily in *APC, CCND2, CDKN2A (p16^{ink4a}), HIN-1* (*SCGD3A1*), *NES1* and *RARB* are best-described methylation in breast cancer [177-182]. Several studies disclose that aberrant promoter methylation is largely associated with gene silencing and dysregulation of the cell cycle. Recent studies have also revealed that hypermethylation in developmental and differentiating genes mainly *HOXB13, HOXA1*, and *HOXA9* and *PAX6* are concomitant with breast carcinogenesis [183, 184]**.** Aberrant methylation pattern in polycomb-regulated genes is implemented in identification of basal-like cancers. Significant CpG hypermethylation have been reported to be associated with genes *ESR1, E-cadherin, CCND2, 14-3-3-σ, RASSF1A* and*SFRP1* in both ductal carcinoma *in-situ* (DCIS) and invasive ductal carcinoma (IDC) [185-187]. Cancer-specific methylations markers in conjunction with methylation-specific PCR (MSP) can be successfully implemented in the detection of breast cancer. *Evron et al.* identified three-gene panel of *Cyclin D2, RARβ* and *TWIST* for detection of malignancy by extracting ductal fluid and lavage [180, 188, 189]. However, *Fackler et al.* improvised three-panel to nine-panel of gene *RASSF1A, TWIST, HIN1*, Cyclin *D2, RARβ, APC, BRCA1, BRCA2* and *p16* being associated with the detection of malignant cells in breast cancer [190].

According to Knudson's two-hit model theory, complete inactivation of tumor suppressor gene is characterized by loss-of-function of both the copies of the gene. Genetic mutation and epigenetic gene silencing characterized by hypermethylation, synergistically leads to the deactivation of tumor suppressor genes. Some of these genes

under the conjoint effect are *p16INK4a, APC* and *BRCA1* [191]. Hypermethylation of *p16INK4A* promoter results in loss of function required for human mammary epithelial cell growth successfully bypassing cell senescence leading to malignancy. Similarly, genes such as, *BRCA1* and *MGMT* associated with DNA repair undergoes DNA methylation mediated inactivation fostering the malignant transformation of the mammary gland. More recently, genes *SFRP1* and *WIF1* associated with WNT oncogenic pathway have been also identified to be hypermethylated in primary breast tumors [192].

Thus, from the above study it is apparent that cells in response to carcinogenic stimuli aggregate anomalous methylation pattern across the promoter region of tumor suppressor genes and are prone to be transformed into malignant cells. Once the patient have been diagnosed with breast cancer, the next question arises is the survival probability and recurrence of this deadly diseases. These two factors can be explained in terms of prognosis in breast cancer.

1.10 Work done so far in prognosis of breast cancer

If a woman is diagnosed with breast cancer at a younger stage of fewer than 50 years, the chemotherapy increases their 15-years of survival rate by 10% while, in older women the increase is only 3%. However, the chemotherapy substantially affects the patient leading to a wide range of acute and long-term side effects. Many a times it is not possible to accurately predict the risk associated with metastasis development and progression, as a consequence of which adjuvant therapy leads to relapse in 40% of patients and ultimately that die [193]. Thus, new prognostic markers are required to assess the patients who are at high risk of developing metastasis.

1.10.1 Established and recent prognostic markers

In oncology, prognostic markers are a clinical measure that enables to elicit patient's risk associated with recurrence of disease after primary treatment. These markers play a vital role in distinguishing patients into different risk groups for which specific treatment strategies can be advised during patient counseling. They also have application in defining the strata in clinical trials in order to ensure comparability of treatment groups. In breast cancer, the risk of metastasis development is characterized by the presence of lymph node metastasis and the loss of histopathological differentiation. The vessel invasion in the patients with tumour-negative axillary lymph nodes results in distant recurrence [194]. However, many a time's women bearing breast tumor without holding the spread in lymph node develop metastasis while, those having tumour spread to the lymph nodes do not develop distant metastases in 10 years after local therapy. Thus, markers to predict the metastasis loci are in scarcity. It has also been predicted that ER+ breast tumors metastasize to the bone while, the invasive lobular carcinomas recur with

increased frequency in ovaries and gastrointestinal tract [195, 196]. These traditional prognostic markers can identify the risk associated with approximately 30% of patients, while, the remaining 70% of patients require a new set of prognostic marker to classify patients into high and low risk. In order to identify the potential novel marker, it should be tested retrospectively in large patient cohorts along with long follow-up period. The conjoint multivariate analysis of established and novel marker should be done in order to assess its significance. Some of these metastatic prognostic markers are enlisted in Table 1.2. Among the enlisted markers, *ERBB2* (epidermal growth factor receptor 2) has raised attention as a plausible prognostic marker. *ERBB2*, proto-oncogene codes for a transmembrane receptor with constitutive tyrosine kinase activity. This gene is overexpressed in 15-30% of breast cancer patients [197]. The prognostic potential was evaluated by Ros and colleagues in recently published literature, including 81 studies and 27,161 patients. Most of the studies done so far has reported that the amplification of *ERBB2* gene is associated with prognosis of patients with axillary lymph metastasis [198]. The increasing evidence based upon its response to adjuvant chemotherapy and endocrine therapy substantiate to its finding of the potential prognostic marker. However, its weak prognostic determinant in lymph-node-positive breast cancer, as declared by WHO, offers limitations and requires adequate studies to validate its prognostic significance.

Table 1.2. Metastatic prognostic marker in breast cancer

Established Marker	Clinical Study	Metastatic Determinant	Details
Histological grade	Established	Grade 1 tumor: low metastasis risk; grade 2: intermediate risk; grade 3: high-risk metastasis	Grading depends ontumor size [199]
Tumour size	Established	Tumour size < 2 cm diameter: low risk metastasis; Tumour size 2-5 cm diameter: Intermediate risk; Tumour size > 5 cm diameter: High risk	Independent prognostic marker [200]
Axillary lymph node	Established	Absence of lymph-node metastases: low-risk metastasis; Presence of lymph-node metastases: high-risk metastasis	Depends upon tumor size [201]

Angio-invasion	Established in patients having lymph-node-negative tumor	Tumour gets associated with 3 blood vessels to undergo metastasis	Localized in patients with lymph node-negative tumors[202]
Steroid receptor expression	Established with adjuvant therapy decision	Low steroid level is associated with metastasis	Related to histological grade: short term metastasis [200]
PAI1/uPA protein level	Newly established	High level of uPA and PAI1 protein: high-risk metastasis	Independent marker [203]
ERBB2 genes	Established	*ERBB2* amplification is associated with metastasis	Patients with lymph node + tumor[204]

1.10.2 Gene expression pattern based prognostic markers

The heterogeneity was taken into consideration, and the prediction of metastatic potential requires concurrent analysis of wide range of markers. The introduction of microarray technology and next generation sequencing has enabled in genome-wide analysis of gene expression and the associated mutations. Unsupervised analysis of gene expression pattern leads to the classification of breast tumors into four distinct subgroups as Luminal A, Luminal B, HER2 and Triple negative (TN). The basal-like subgroups (HER2 and TN) bearing estrogen-negative-receptor shows high expression of cytokeratin-5 and cytokeratin-17 [134]. However, estrogen-positive receptor subgroup Luminal A exhibits high-level expression of cytokeratin-8 and cytokeratin-18 while, luminal B has low expression of these genes. These findings reveal that the differential gene expression associated with subtypes holds characteristic clinical significance and are a potential prognostic target for therapeutic implications [205].

In another approach, the supervised classification of gene-expression pattern can predict the clinical behavior of tumors. This classification method was based on the expression profile of 70 genes to predict the likelihood of distant metastasis in young patients (< 55 years of age) having lymph-node-negative tumors [206]. The primary breast tumors were classified as poor prognosis and good prognosis signatures based on the expression profile. Poor prognosis signature comprised of the genes that convoluted cell cycle, invasion, metastasis, angiogenesis and signal transduction. It also included the genes that are exclusively expressed in the stromal cells surrounding the epithelial cells in the tumor. Some of these genes constituting of *MMP1* and *MMP9* promotes in

extracellular matrix (ECM) degradation and tumor invasion. The upregulation of these genes in stromal cells offers significant prognostic signature for breast cancer metastasis. Thus, multivariate analysis of gene-expression signatures holds strong prediction for metastasis-free survival and overall survival. During the analysis of 151 patients having lymph-node-negative tumors, 60% of the patients were in high metastatic risk (poor-prognosis) while, 40% of the patients were in low metastatic risk (good prognosis). However, after 10 years of the follow-up period, 56% of the patients had poor-prognosis, and only 13% were with good prognosis [207]. Recently, studies based upon RT-PCR analysis of 21 genes exhibited the metastatic potential to be associated with an expression ratio of *HIXB13* and *IL17BL* genes [208]. Thus, gene-expression profiling defines the prognostic classification of breast cancer, however, many a times the presence of mutations and polymorphism in the proximity of gene expression offers limitations to its potential prognostic significance.

1.10.3 Analysis of mutations including single nucleotide polymorphisms for identification of prognostic biomarkers

Identification of numerous breast cancer predisposition factors associated with single-locus or epistatic effects can be largely used for breast cancer risk assessment [209]. The conjoint effect of multiple genetic risk loci increases the risk prediction accuracy and eventually upholds in developing population-based risk screening and stratification programs [210]. These genetic loci are associated with germline DNA variations mainly the SNPs and copy number variations. Several studies have demonstrated that a germline mutation in *BRCA1* and *BRCA2* genes results in a translational shift and aberrantly spliced site leading to premature truncation of encoded proteins [211]. However, the germline mutation associated with *BRCA1* and *BRCA2* genes are very rare, and the predisposition of these 2 genes could explain only 15-20% of the genetic risk in overall populations [212, 213]. Similarly, germline mutations associated with *TP53* and *PTEN* genes exhibit moderate penetrance for breast cancer predisposition [214-216].

With the advancement in genotype technologies and completion of Human Genome, Hap Map and 1000 Genomes projects, the paradigm shift of genetic association with limited candidate genes has expanded to the genome-wide investigation of genetic variants. Thus, GWAS investigation have identified >4,500 low-penetrance SNPs associated with >700 different diseases or traits [217]. Studies conducted by Easton *et al.* evaluating breast cancer cases in United Kingdom have identified 4 SNPs associated with the genetic loci of *FGFR2, TNRC9, MAP3K*, and *LSP1* [218]. Furthermore, Ghoussaini *et al.* conducted a large-scale replication study in European women based upon multiple independent breast cancer GWAS and identified 3 novel loci susceptible to breast cancer on chromosome 12p11, 12q24 and 21q21.While, SNPs on 12q24 and 21q21 loci were

strongly linked to ER+ breast cancer, SNPs located on 12p11 chromosome offered risk for both ER+ and ER- breast cancer [219]. More recently, the Collaborative Oncological Gene- Environment Study (COGS) conducted by Michailidou *et al* on the largest GWAS study constituting > 100,000 breast cancer individuals of European ancestry, identified 41 novel loci susceptible to breast cancer susceptible located on chromosomes 1-14, 16, 18, 19 and 22. These variants were associated with high, moderate and low penetrance genes and explained about 50% of the familial risk of breast cancer. Thus, the genetic loci harboring the risk variants included *MDM4, TET2, TERT, KLF4, POU5F1B, RAD51B,* and *BABM1* genes [220]**.** Importantly, these results shared genetic susceptibility for breast, ovarian and prostate cancer, providing evidence that the development and progression of these hormone-related cancers share common genetic etiology. Besides, the identification of prognostic marker based upon genetic variants, the epigenetic modulation does hold large significance in the detection of the risk associated loci across the genome.

1.10.4 Risk associated with DNA methylation in prognosis of breast cancer

Aberrant epigenetic regulations in breast cancer are emphasized on the molecular mechanism of cancer development, prediction of aggressiveness and potential epigenetic therapeutic implications. The investigation carried out so far is propounded on the identification of novel biomarkers to predict the risk associated with survival of breast cancer patients. There are many pieces of evidence that the hypermethylation of tumor suppressor genes in breast cancer plays a decisive role in cell-cycle regulation, tissue invasion, apoptosis, metastasis, and angiogenesis [221-224]. Thus, the aberrant methylation profiles of these genes are highly associated with cancer staging and prognosis.

Esteller *et al.*, identified the significant role of *BRCA1, p16, GSTP1* and *CHD1* in tissue invasion and metastasis. In addition, *ADAM23* gene responsible for cell adhesion process exhibited increased promoter hypermethylation [225]. Similarly, Fang *et al*. analyzed 39 primary breast tumor specimen using Infinium 27K platform identified DNA methylation signatures concomitant with breast cancer metastasis. The methylation signature of three genes, primarily, rho guanine nucleotide exchange factor (*ARHGEF7*), ALX homeobox 4 (*ALX4*) and RAS-protein-specific guanine nucleotide releasing factor 2 (*RASGRF2*), holds strong determinant for metastasis-free survival and overall survival. In particular, these signatures shared common prognostic space in gliomas, colon, and breast cancer [226]. In another study carried by Dedeurwaerder *et al.,* profiling 248 breast cancer samples recognized immune genes holding significant prognostic value. In particular, the promoter hypermethylation of lymphocyte transmembrane adaptor1

(*LAX1*) and *CD3D* genes strongly determined the survival in breast cancer subtypes [227]. In recent study, Conway *et al*. evaluated 935 CpG sites in 517 invasive breast tumors from Carolina Breast Cancer study. Array-based DNA methylation profiling led to the identification of 266 differentially methylated CpG loci associated with hormone receptor (HR+ and HR-), luminal A and p53 wild-type and mutant breast cancer. Hypermethylation of *FABP3*, *FGF2*, *FZD9*, *GAS7*, *HDAC9*, *HOXA11*, *MME*, *PAX6*, *POMC*, *PTGS2*, *RASSF1*, *RBP1*, and *SCGB3A1* genes were associated with the CpG loci of HR+, luminal A and p53 wild-type breast cancer. Similarly, highly methylated loci in HR-, basal-like and p53 mutant tumors comprised of *BCR*, *C4B*, *DAB2IP*, *MEST*, *RARA*, *SEPT5*, *TFF1*, *THY1*, and *SERPINA5* genes. Hypermethylated luminal-tumours were also enriched for homeobox and developmental genes (*ASCL2*, *DLK1*, *EYA4*, *GAS7*, *HOXA5*, *HOXA9*, *HOXB13*, *IHH*, *IPF1*, *ISL1*, *PAX6*, *TBX1*, *SOX1*, and *SOX17*) [228]. These differentially methylated genes had a substantial role in establishing and maintaining tumor phenotypes and clinical outcomes. Methylome sequencing in triple-negative breast cancer carried out by Stirzaker *et al.* identified distinct methylation cluster associated with 17 differentially methylated regions holding a strong association with overall survival. Notably, these DMRs predominantly overlapped with conserved transcription factor binding regions and DNAase 1 hypersensitive regions. Of the genes enlisted, many were associated with *WT1* (Wilson Tumour 1), *WT1-AS* (Antisense *WT1*), *DMRTA1* (DMRT-like family A1) and *HOXB13* (Homeobox gene family) [229].

Besides, the exclusive analysis of genetic and epigenetic aberration, the integrated study will embark on a contextual framework for unraveling the cryptic details of recurrence and overall survival. Once the diagnosis and prognostic markers have been identified the next step follows is to identify suitable inhibitor which can minimise the load of this deadly disease. Herein we enlist the inhibitors that have been implemented against breast cancer.

1.11Molecular targets and inhibitors known till date for treatment of breast cancer

Our therapeutic armamentarium presents several chemotherapeutic agents against breast cancer. However, the vast majority of patients develop resistance and eventually capitulate to the disease. Inhibitors targeting specific molecular target in breast cancer holds promise for improving clinical outcomes. The large success of lapatinib and trastuzumab is treating HER2-overexpression in combination with endocrine therapy against positive hormone receptor exemplify this [230, 231].

1.11.1 Targeting genetic regulators

Basic research concerned with the better understanding of the biology underlying the malignant progression of breast cancer has motivated us to identify promising molecular targets in breast cancer. Advancement in the modern oncology has expanded the spectrum of potential molecular targets. However, the intra-tumour heterogeneity in the microenvironment presents a biased assessment to the complete spectrum of genomic alterations of the corresponding cancers. The complete spectrum of targets can be visualised according to the cellular component targeted, namely, breast cancer cells [Table 1.3], breast cancer stem cells [Table 1.4] and the breast cancer tissue microenvironment [Table 1.5]. Most of these genetically regulated targets and the therapeutic agents are specific to malignant cells and possess higher therapeutic index than the conventional chemotherapeutics. However, toxicity is the major concern.

Table 1.3 Targeted genetic agents against breast cancer cells

Cellular Target	Therapeutic Agents	Application on Patients	Clinical Study
mTORC1/2	INK128	Advanced or metastatic solid tumors	Phase I [232]
	AZD2014	ER+ or advanced MBC	Phase I [233]
Dual *PI3K–mTOR*	XL765	HR+, HER2– recurrent or MBC	Phase I–II [234]
	BEZ235	HR+ MBC, HER2+ locally advanced MBC HER2+ MBC	Phase I [235]
	GDC-0980	ER+ locally advanced or MBC	Phase II [236]
	GSK2126458	Solid tumors or lymphoma	Phase I [235]
Pan-PI3K	XL147	HER2+, MBC, HR+, HER2	Phase I-II [237]
	BKM120	HER2–, HR+, MBC	Phase II [238]
	GDC-0941	ER+ locally advanced or MBC	Phase II [236]
PI3Kα	BYL719	Advanced solid malignancies	Phase I [239]
	GDC-0032	Locally advanced or metastatic solid tumors	Phase II [240]
PI3Kβ	GSK2636771	Advanced solid tumors with PTEN deficiency	Phase I-II a [241]

AKT	MK-2206	ER+, MBC, Advanced BC with a PIK3CA mutation and/or PTEN loss	Phase I [242]
	AZD5363	Advanced ER+ BC	Phase I [243]
IGF-1R	Cixutumumab	Locally recurrent or MBC HER2+ previously treated BC	Phase I-II [244]
	Dalotuzumab	ER+ BC	Phase II [241]
Multitargeted FGFR	Dovitinib	HR+, HER2– BC	Phase I-II [245]
	E-3810 (EOS)	Locally advanced or metastatic solid tumors	Phase I [246]
MET pathway	Onartuzumab	TNBC	Phase II [247]
	Foretinib	HER2+ MBC, TNBC	Phase I-II [248]
	Cabozantinib	HR+, HER2– BC	Phase II [249]
Cyclin-dependentkinase	PD0332991	MBC, HR+ advanced BC	Phase I [248]
	Dinaciclib	Metastatic TNBC	Phase I [250]
	Seliciclib	Advanced solid tumors	Phase I [250]
MAPK pathway	AZD6244	Locally advanced or metastatic solid tumors	Phase I [251]
	GSK1120212	Advanced solid tumors	Phase I [252]
	TAK-733	Advanced solid tumors	Phase I [253]
EGFR–HER3	MEHD7945A	Locally advanced, or metastatic epithelial malignancies	Phase I [254]
Aurora kinases	ENMD-2076	Locally advanced or metastatic TNBC	Phase II [255]
Androgen receptor	Bicalutamide	Androgen receptor-positive, HR– MBC	Phase II [256]
	Abiraterone	ER+ MBC progressing after letrozole or anastrozole	Phase II [257]
Prolactin receptor	LFA102	Metastatic Breast Cancer	Phase I [258]

Table 1.4 Target agents against breast cancer stem cells

Cellular Target	Therapeutic Agents	Application on Patients	Clinical Study
γ- secretase	MK-0752	Metastatic or locally advanced	Phase I [233]

		solid tumor	
	RO4929097	HER2– unresectable or MBC	Phase I [259]
	BMS-906024	Advanced or metastatic solid tumors	Phase I [260]
Delta-like ligand 4	MEDI0639	Advanced solid tumors	Phase I [261]
Smoothened receptor	XL139	Solid tumors, Advanced or metastatic solid tumors	Phase I [262]
	Vismodegib		Phase I [263]
	PF-04449913	HER2– unresectable or MBC	Phase I [264]
	LDE225	Advanced or metastatic solid tumors	Phase I [265]
	TAK-441		Phase I [266]
	LEQ506	Advanced tumours	Phase I [267]
		Advanced solid tumors	
		Advanced solid tumors	
Frizzled receptor	OMP-18R5	Advanced or metastatic solid tumors	Phase I [268]
	OMP-54F28		Phase I [245]
		Advanced or metastatic solid tumors	
β-catenin	PRI-724	Advanced solid tumors	Phase I [269]
Porcupine	LGK974	Melanoma (except uveal), lobular or triple-negative BC, or pancreatic adenocarcinoma	Phase I [270]

Table 1.5 Targeted Agents against breast cancer microenvironment

Cellular Target	**Therapeutic Agents**	**Application on Patients**	**Clinical Study**
PD-1	Nivolumab	Locally advanced or metastatic solid tumors	Phase I [271]
	AMP-224	Advanced cancer	Phase I [258]
PD-L1	BMS-936559	Relapsed breast cancer	Phase I [272]
	MPDL3280A	Advanced solid tumors	Phase I [244]
Lysyl oxidase	Simtuzumab	Advanced solid tumors	Phase I [273]
Chemokine receptor	PLX3397	Advanced solid tumors	Phase I [231]

Integrin	Cilengitide	Unresectable solid tumors, excluding lymphoma	Phase I [274]
	PF-04605412	Advanced or metastatic solid tumors	Phase I [246]
	IMGN388	Advanced solid tumors	Phase I [225]
Hypoxia	EZN-2968	Advanced solid tumors	Phase I [275]
	TH-302	Advanced solid tumors	Phase I [276]

1.11.2 Targeting epigenetic regulators for breast cancer therapy

Epigenetic aberrations characterized by DNA methylation, histone modifications, miRNA downregulation and chromatin remodeling offer new therapeutic targets in breast cancer [277, 278]. HDAC inhibitors have shown the reactivation of *ESR1* and *PGR* gene expression in ER-negative breast cancer cells [279]. Some of these inhibitors primarily vorinostat, entinostat, and panobinostat have already passed through the clinical trials. Triple-negative breast cancer cells, when targeted with HDAC inhibitors in combination with aurora kinase inhibitors, enhances the antitumor activity [280, 281]. HDAC inhibitors also restore the sensitivity to trastuzumab through small molecule acting as EGFR/HER2 inhibitor [282]. Phase II trial assessing the inhibitory effect of entinostat in combination with exemestane in ER+ breast cancer shows reduced risk of progression of breast cancer [283, 284]. DNA methyltransferases inhibitors, specifically 5-azacytidine and decitabine have been successfully implemented in the treatment of hematological malignancies have also shown efficiency in treating metastatic breast cancer [285-287]. Transient low doses have exhibited antitumor efficacy in the *in-vivo* condition in a breast cancer xenograft model and restoration of expression of hypermethylated genes. Moreover, the hypermethylation of tumor suppressor genes including *BRCA-1*, *E-cadherin,* and *MASPIN* are also restored in breast cancer [288]. DNMT inhibitors also sensitize breast cancer lines to chemotherapeutic agent doxorubicin by inducing tumor necrosis factor related apoptosis inducing ligand (TRAIL) [289]. Moreover, the conjoint effect of HDAC and DNMT inhibitors induces enhanced expression of *ESR1* gene [290]. The ongoing research elucidates the novel combination strategies of epigenetic modifiers with tamoxifen, aromatase inhibitors, trastuzumab and other cytotoxic agents in breast cancer treatment. A number of early phase ongoing or completed clinical trials for early phase solid tumor diagnosis are included in table 1.6.

Table 1.6 Epigenetic modifiers in breast cancer

Cellular Target	Agent	Application on patients	Clinical Trials
HDACs	Vorinostat	Advanced breast cancer, median prior chemotherapy cycles	Phase II [291]
	Vorinostat + tamoxifen	Advanced ER-positive breast cancer hormone-resistant	Phase II [292]
	Vorinostat + aromatase	Advanced ER-positive breast cancer	Phase II [283]
	Entinostat + exemestane	Advanced ER-positive breast cancer, progression on prior non-steroidal AI	Phase II [293]
	Entinostat + Anastrozole	Primary operable triple-negative breast cancer	Phase II [294]
	Vorinostat + paclitaxel + bevacizumab	Primary operable triple-negative breast cancer	Phase I-II [283]
	Vorinostat/placebo + nabpaclitaxel + carboplatin	Advanced breast cancer	Phase II [295]
	Vorinostat + ixabepilone	Primary operable breast cancer, triple-negative or high-grade ER-positive	Phase I [296]
	Vorinostat + trastuzumab	Advanced breast cancer	Phase I-II [297]
	Vorinostat + lapatinib	Advanced HER2-positive breast cancer	Phase I-II [298]

	Entinostat + lapatinib	Advanced solid tumors and advanced HER2-positive breast cancer	Phase I-II [299]
DNMTs	AZA single agent	Primary operable breast cancer “window trial”, a triple-negative breast cancer	Phase II [300]
	AZA + entinostat	Advanced breast cancer: triple-negative and hormone-resistant	Phase I-II [266]
	Decitabine + panobinostat +/− tamoxifen	Advanced triple-negative breast cancer	Phase II [301]
	AZA + nab-paclitaxel	Advanced solid tumors and breast cancer	Phase I-II [302]

1.11.3 Other molecular targets

There are numerous other molecular target agents also being under clinical trial. These include the compounds targeting SRC complex, a tyrosine kinase regulating numerous oncogenic targets, primarily cell proliferation, survival, induction of angiogenesis and promoting cell migration or invasive phenotype (dasatinib, bosutinib, and saracatinib) [303-305]. Interestingly preclinical studies substantiate the conjoint effect of anti-SRC agents and trastuzumab. SRC complex are activated in cells despite acquired and *de-novo* trastuzumab resistance, and acts in the downstream of trastuzumab resistance which in turn can be pharmaceutically reversed. Similarly, HER3 has emerged as potential drug target candidate against U3‑1287 and MM‑121 inhibitors have been identified [306, 307]. Targeting HER3 is of greater significance in HER2+ metastasis breast cancer, as the data have disclosed that the formation of HER2-HER3 dimer promotes the malignant progression of HER2+ breast cancer cells [308]. Another group of targeting agents corresponds to the androgen receptor or prolactin receptor. The inhibitors targeting these receptors are bicalutamide, enzalutamide and abiraterone [256]. Molecular characterization of triple negative breast cancer spectacles the presence of luminal androgen receptor subtypes. Targeting these receptors is of greater significance as it simulates the growth of tumor cells in context with the stimulation of WNT and HER2 oncogenic pathway.

Identification of genetic and epigenetic aberrations described so far has led to the development of targeted therapeutics in breast cancer. However, the clinical results obtained so far do not meet the requirement for treatment of intra-tumor heterogeneity in breast cancer. With the advent of next-generation sequencing techniques, the interrogation of large-scale genomic alterations regulating DNA methylation will significantly expand to the new therapeutic arsenal. On targeting the DNA methyltransferase enzyme (DNMTs) will cause the reversal of the differential methylation (hypermethylation) opening a new window to the therapeutic intervention in breast cancer.

1.12 Lacuna in the understanding of the problem

With the advent of whole genome sequencing program, several human cancers have come up with an explicated results that the mutated genes associated with epigenome can remodel the complete cellular programming leading to cancerous state. The presence of these associated mutations was unknown and overlooked, however the analysis of 1,000 of cell lines by whole exome sequencing disclosed the presence of large number of potential mutations regulating epigenetic modifications, preferentially DNA methylation [309]. Genome-wide association studies (GWAS) have identified the presence of these

mutations in the form single nucleotide polymorphism (SNPs) holding an increased risk in several diseases including cancer [310]. Surprisingly, cancer associated SNPs are highly enriched in the defined region of functional enhancers and alter the chromatin landscape [311]. Moreover, several genome-wide expression quantitative trait loci are linked to genetic variations and changes in gene regulations [312]. More recently, these genetic variants have been identified to be strongly associated with transcription factor binding site, there-by leading to differential gene expression. Although many studies have revealed allele-specific DNA methylation and gene expression related to genome imprinting and X-chromosome inactivation, recent studies have shown that these allele-specific phenomena are involved in other cellular activities [313]. Notably, most of the allele-specific DNA methylation are strongly correlated with SNP genotypes affecting the binding of transcription factors and long-range chromosome structure. Conversely, the presence of SNPs in the vicinity of CpG site can create or delete the loci, subsequently influencing the binding of transcription factor and methyl-binding proteins (MBDs) [314]. Further studies need to be associated with epigenetic variation (epigenotype), genetic variation (genotype) and trait or disease (phenotype) to explain the functional causality of diseases [Figure 1.4] [315]. Moreover, there are increasing the number of nucleosides, and non-nucleoside analogs being studied as anti-cancer drugs. Inhibition of DNA methyltransferase (DNMTs) by 5-azacytidine (Vidaza; azacitidine) and 5-Aza-2'-deoxycytidine (Dacogen; decitabine) have been approved by FDA for cancer treatment [316]. However, owing to their genotoxic effect in high dose offers limitation for further implementation in clinical settings.

Ever growing evidence of epigenetic alterations characterized by DNA methylation in cancer, offers a chance to enhance the increased sensitivity and specificity in its diagnosis and therapy. Several genome-wide consortia such as 1000 Genome Project, IHEC, and Roadmap projects are blueprints for methylome mapping [317-319]. These databases constitute massive data for reference in research and clinical trial. Integration and management of these data from several “omics” approach mainly genomics, transcriptomics, and epigenomics will lead to the production of predictive models for identification of novel epigenetic biomarkers and signatures [Figure 1.4]. Moreover, combinations of hypermethylated biomarkers in breast cancer will enhance the sensitivity of detection and prediction of tumor progression. Implementing high-throughput study of next generation sequencing and statistical analysis will increase in the characterization of cancer subtypes. Owing to the lacunae in the analysis based upon exclusive genetic and epigenetic aberrations, our present work elaborates the conjoint study of genetic and epigenetic anomalies that will be effective in diagnosis, prognosis and therapeutic implication in breast cancer.

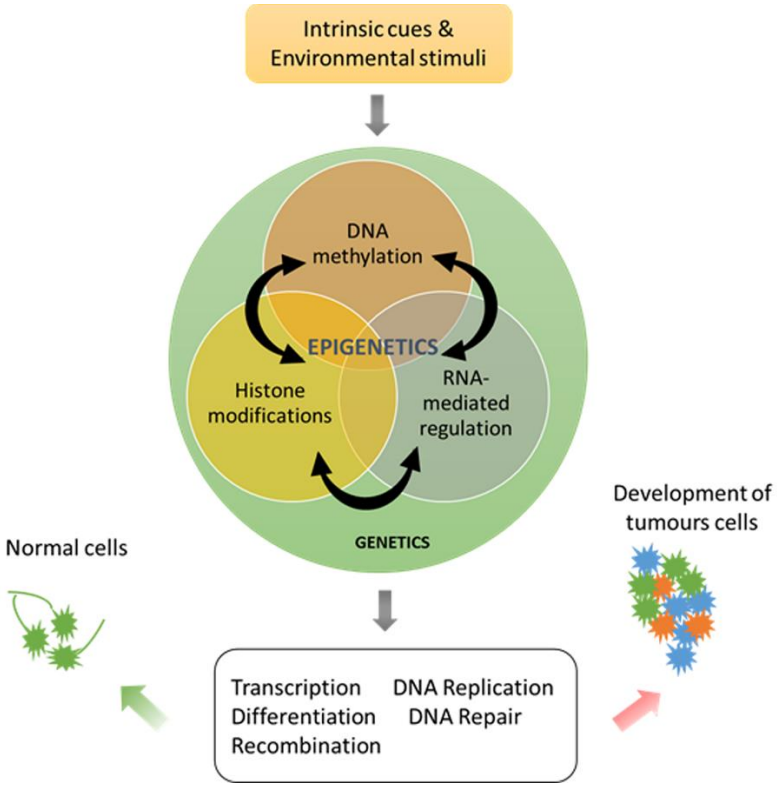

Figure 1.4: Epigenetic and genetic aberrations synergistically regulate the cellular function leading to carcinogenesis.

1.13 Objectives

In many instances genetic and epigenetic aberrations have been thought to be an independent entity relishing an active participation in carcinogenesis. However, with the recent outcome of complete genome sequencing of thousands of human cancer have resulted in an unexpected findings of many mutations that control the epigenome. Of all the epigenetic process, the imbalance in DNA methylation conspires with the mutation to drive the process of cancer development and progression. Thus, it offers a daunting challenge to identify a stochastic change in genome-wide DNA methylation and associated mutations in revealing the anomalies in breast cancer. Moreover, the active participation of DNA methyltransferases in differential methylation needs to be targeted to slow down its hyperactivity leading to therapeutic interventions. Keeping these views in mind, we have designed objectives as follows:

1. To understand how differential allelic distribution regulates CpG methylation in tumor and normal samples leading to the diagnosis of breast cancer.
2. To decipher how single nucleotide polymorphisms affect DNA methylation at nearby CpGs and impact breast cancer prognosis among individuals.

3. To identify novel inhibitor(s) targeting DNA methyltransferases for therapeutic intervention in breast cancer.

The success of genome-wide association studies has been successfully comprehended with the epigenome-wide association studies in identification in capturing the disease association epigenetic variant primarily differential DNA methylation in the diagnosis, prognosis and therapeutic implications of breast cancer.

1.14 Overview of this thesis

Integration of genetic and epigenetic marks holds the key to the understanding the underlying biology of the complex interaction of inherited trait and environmental cues in the catastrophe of the deadly disease like cancer. This interplay between the two layers of information reveals correlation between single nucleotide polymorphism (SNPs) and the DNA methylation at a particular site leading to the identification of methylation quantitative trait loci (meQTL). In chapter 2, we begin with genotype-epitype interactions and the associated phenotypes in identification of diagnostic marker in breast invasive carcinoma (BRCA) samples obtained from TCGA database. Realizing the fact that the large section of cancer-related SNPs resides in the noncoding region and holds incredible functional impact, we look forward to identify novel diagnostic biomarkers with respect to the presence of these meQTLs. Once we have established a platform for diagnosis, we also need to analyze the longevity related to breast cancer patient survival. Chapter 3 details about the epigenome-wide association analysis (EWAS) of the myriad of meQTLs in association with the risk to the survival of breast cancer patients. In our study, we mainly reveal the complex interplay of genetic and epigenetic variants in predisposing the diagnosed BRCA patients to a lethal stage. Comprehensive assessment of these risk variants at different stages will lead to the identification novel biomarkers in breast cancer prognosis. Once we have identified the biomarkers in association with diagnosis and prognosis, the next step follows therapeutic implications. In chapter 4, we describe the pharmacological manipulation of key epigenetic enzyme DNA methyltransferase in breast cancer reprogramming. In light of emerging concept of chemoinformatics, molecular docking, and simulation studies have been employed to accelerate the development of novel DNMT inhibitors having medicinally relevant space. The *in-silico* analysis has been comprehended by an *in-vitro* study to visualize the effect of the novel inhibitor inhibiting DNMT activity and the ability to restore the expression of silenced tumor-suppressor gene devoid of being toxic to normal cells.

The strategy for the effective treatment can be based upon the combinatorial analysis. If we can integrate candidate-altered genetic and epigenetic profile into a predictive model in conjunction with novel therapeutic implications, it will lead to the low-dose and customized high-impact treatment we seek.

Chapter 2

To understand how differential allelic distribution regulates CpG methylation in tumor and normal samples leading to the diagnosis of breast cancer

2.1 Introduction

The susceptibility to inherit breast cancer is estimated to be 25-50% however, only 5-10% of cases are explained by the genetic variant in association with *BRCA1*, *BRCA2,* and *TP53* genes [320]. In some instances, it is characterized by the conjoint effect of multiple genetic variant loci [321]**.** Thus, heredity is not the only cause for genesis in most of the breast cancer. It is a consequence of a gradual accumulation of mutational load, telomere dysfunction, and epigenetic gene silencing with developing age [322]. Exposure to estrogen hormone causes an anomalous change in breast epithelial stem cell subsequently propagating cell to divide. Besides the direct stimulation of epithelial cells development, growth hormone also influences stromal microenvironment for tumor cell development leading to profound tumor progression. During metastasis, this microenvironment is regulated largely by paracrine signaling between epithelial and neighboring stromal fibroblast [323]. This eccentricity in breast development leading to the cancerous ailment is affected by variability in environmental cues associated with epigenetic aberrations. Thus, it can be inferred that genetic variations in conjunction with epigenetic anomalies regulate the aberrant division of epithelial and stromal cells leading to breast cancer.

Elucidating the genetic and non-genetic determinants in the diagnosis of breast cancer is one of the principal challenges in the field of biomedical research. Despite GWAS discloses > 800 SNPs in several diseases, still a substantial portion of the causality remains enigmatic [324]. The epigenomic equivalent of GWAS characterized by epigenome-wide association studies (EWAS) presents novel opportunities in confounding the factors and follow-up influencing the disease etiology. DNA methylation is a significant epigenomic marker that represents a molecular phenotype that links to the genotype in resolving disease complexity. This variation in the genotype is characterized by the presence of SNPs in the vicinity of CpG sites which in turn disrupts methylation status at each CpG sites. These SNPs form major class of methylation quantitative trait loci (meQTLs) [325, 326]. It has also been reported that these SNPs are associated with

CpG sites within consensus sequence of methyl-CpG binding proteins; thus, it can be hypothesised that strategies focusing on the identification of SNPs for genotyping will contribute in elucidating the genetic epidemiology of breast cancer.

The dynamic characterization of DNA methylation facilitates the determination of diagnostic biomarker by considering inter and intra-individual variations. The SNPs associated with each CpG sites influences the methylation pattern leading to differentially methylated regions (DMRs) [327]. These DMRs across the healthy individual and the cancerous tissue helps in estimating the variance across particular CpG site located in the intergenic or intragenic regions. However, the loci constituting unstable methylation pattern are precluded as false positive hits. Nowadays reference data set consortia such as 1000 Genome Project has been created based on the epigenomic profiles of stem cells and developmental somatic tissue profile from healthy individuals [138]. Systemic screening of these reference data set obtained from different individuals enables in identification and exclusion of variable CpG sites and regions facilitating in biomarker selection.

Methylation at any CpG site is quantified in terms of *beta* value that is defined as the ratio of intensities between methylated (*M*) and unmethylated (*U*) allele. Thus, beta value is given by the equation 1:

$$beta = \frac{\max(M,0)}{\max(M,0)+\max(U,0)+100} \qquad \text{(Equation 1)}$$

Where, M and U codes for signal A and B produced by two different beads in Illumina methylation assays. Here the constant 100 has been used to normalize the beta value provided the value with respect to M and U are comparatively small [328]. The beta value of any locus range from 0 (unmethylated) to 1 (completely methylated). While, Illumina platform corroborates with the genome-wide association of beta values at each CpG site, the genotypic information with respect to allele frequency distribution can be excavated from Affymetrix high-throughput SNP array database. For each SNP, the intensities of two alleles denoted as A and B are measured as four sets of perfect match probe from sense and antisense strands, denoted as +/-. The intensities are normalized by excluding background noise as mismatch probes (MM) [329, 330]. TCGA database repository supports the platform for Illumina DNA methylation 27/450 and Affymetrix genome-wide human SNP array 6.0 data covering 33 cancer types [331]. Thus, beta value and genotypic details across the breast cancer and matched normal patients can be assembled in the identification of meQTLs as the diagnostic marker.

The present study unravels the combinatorial effect of meQTLs and gene expression in tumor growth and development. The contribution of genetic variants (SNPs) in regulating DNA methylation and gene expression illuminate their potential function in unfolding the complexity of deadly disease like cancer. Many studies conducted till date already discloses the significance of SNPs in the disease phenotype. Besides, targeting the

CpG sites in protein-coding regions, SNPs were also identified to influence the non-coding regions mainly the promoters, introns, alternates spliced regions and the intergenic regions. The complete analysis is based on recently developed statistical methodology at R-interface, overarching the confounding paradigm of differential methylation, single nucleotide polymorphism and gene expression in the diagnosis of breast cancer. In summary, we have demonstrated the systematic assessment of methylation and expression data being influenced by genetic variations in breast cancer and the analysis were based on enriched publically accessible TCGA cohort. Thus, the combinatorial effect of differential methylation and gene expression will be a gateway towards the understanding of the underlying mechanism behind breast cancer pathogenesis.

2.2 Materials and Methods

2.2.1 Dataset retrieval from TCGA repository

The Cancer Genome Atlas (TCGA) in a national research consortium spearheaded by National Cancer Institute (NCI) in collaboration with National Human Genome Research Institute (NHGRI). The database offers comprehensive profiles of cancer genomes through the application of high-throughput technologies, primarily microarrays and next generation sequencing. It is affluent with more than 6000 patients' tumor and matched normal samples profiles, extending up to 37 types of genotypic and phenotypic data across 33 cancer types. The data generated are categorized based on data type and data level. This categorised data include level 1 (Raw, non-normalized), level 2 (processed data), level 3 (segmented/normalised) and level IV (summarised) data. These data integrate samples details as "TCGA barcode describing the participants and biospecimens (blood, tissue) [332]. We downloaded the Level-III DNA methylation and RNA seq and Level-1 SNP array data of breast invasive carcinoma (BRCA).

2.2.2 Illumina 450 k DNA methylation data

Illumina has established Infinium-based Human Methylation microarray assay for quantitative analysis of methylation across the genome. This high-throughput assay Human Methylation450 (450K) Bead Chip consists of 485,577 probes that cover 482 421 CpG sites, 3091 non-CpG sites and 65 random SNP. The level 3 methylation normalized dataset for BRCA encompasses the detail for 746 tumors and 96 matched normal samples. Of the total, 740 tumor samples were obtained from the primary tumor, while remaining 6 samples pertaining to metastatic class were filtered out. Each of these normalized data sheets incorporated the details for genomic coordinates and beta-values for each CpG sites, while the associated gene information was optional. 65 non-random SNPs were excluded and 485,512 CpG sites were processed for further studies. These

methylation files were processed to interrogate the SNPs associated with each CpG loci. The entire set of SNPs information was based upon the Affymetrix Genome-Wide Human SNP Array 6.0 genotypic platform.

2.2.3 Affymetrix SNP arrays dataset preparation

Affymetrix system offers series of microarray platform feasible for exploring biological mechanisms such as genotyping, copy number variations and the differential expression, on the whole genomic scale. In our present study, we mainly focus on SNP based microarrays for high-throughput genotyping in genome-wide association studies. Level 1, raw SNP array data for 1076 for BRCA tumor, 137 matched normal and 975 blood samples were downloaded from TCGA Data Portal. Data normalization and genotype call for each sample were performed by "Corrected Robust Linear Model with Maximum likelihood distance" algorithm [333]. CRLMM algorithm estimates the genotype based upon two-stage hierarchical model (M) for log ratio of I_A and I_B ($M = \log_2 (I_A/ I_B)$. The model follows the empirical Bayes approach in which the mean conditioned on genotype has multivariate normal distribution while the variance has an inverse gamma distribution. Based upon the information for mean and variance, CRLMM computes the posterior probabilities for each genotype given the observed log ratio M. The algorithm estimates the genotype using linear mixture model and for each SNP-genotype combination, the uncertainty parameter is corrected using HapMap samples. In order to process the large data set, the crlmm-package was substantiated with ff package to reduce memory footprint (http://cran.at.r-project.org/web/packages/ff/index.html). The algorithm was implemented to decode the genotype calls for SNPs as 1 (AA/Reference allele), 2 (AB/Heterozygous allele) and 3 (BB/Alternate allele). The genotype calls at the threshold of 0.05 were filtered while, those having more than 25% low confidence calls was excluded. The complete process of data normalization and data filtering resulted in 905,422 SNPs for further analysis.

2.2.4 RNAseq dataset preparation

Direct sequencing of the transcriptome by RNA-sequencing (RNA-seq) method is now possible with the advent of next generation technology. Sequence data from RNA-seq method can be used to identify (*de-novo* assembly of transcripts) and quantify the expressed transcripts. RNA-seq data also facilitates detection of transcript fusion and alternate splicing of isoforms. TCGA Data Portal offers enormous resource for identification of differentially expressed genes between different tissue types (for example, cancer vs. normal or different cancer types) [334]. Methodology for RNAseq data development involves the alignment of the fragmented transcript (short reads) to the reference genome. RNASeqV2 level 3 released gene expression for RNAseq were

downloaded from TCGA. The dataset constitutes the details for 1056 tumor and 112 matched normal samples. The data processing and quality control was done by Broad Institute TCGA workgroup [http://gdac.broadinstitute.org/]. The reference for gene transcript was based upon HG19 UCSC track (http://hgdownload.cse.ucsc.edu/downloads.html). The Map-Splice was used to do the alignment and the quantification was carried by RSEM [335, 336]. We downloaded the upper quantile normalized RSEM count estimates.

2.2.5 R statistical programming software

The complete statistical analysis detailed in the study was carried out at R-interface [http://www.R-project.org/]. R is acquainted with the substantial collection of the statistical algorithms for easy handling of data and well-designed extension system and excellent visualization platform. It constitutes several integrated modules and packages while, new modules can be submitted to the central repository of the Comprehensive R Archive Network (CRAN) or to, the Bioconductor [337].

2.2.6 Procedure for the identification of regulatory CpG-SNP candidates associated with breast cancer diagnosis

Figure 2.1 is a detailedoutline of the procedure for identification of regulatory Cp-SNP candidates involved in thediagnosis of breast cancer. We describe the details in the following steps.

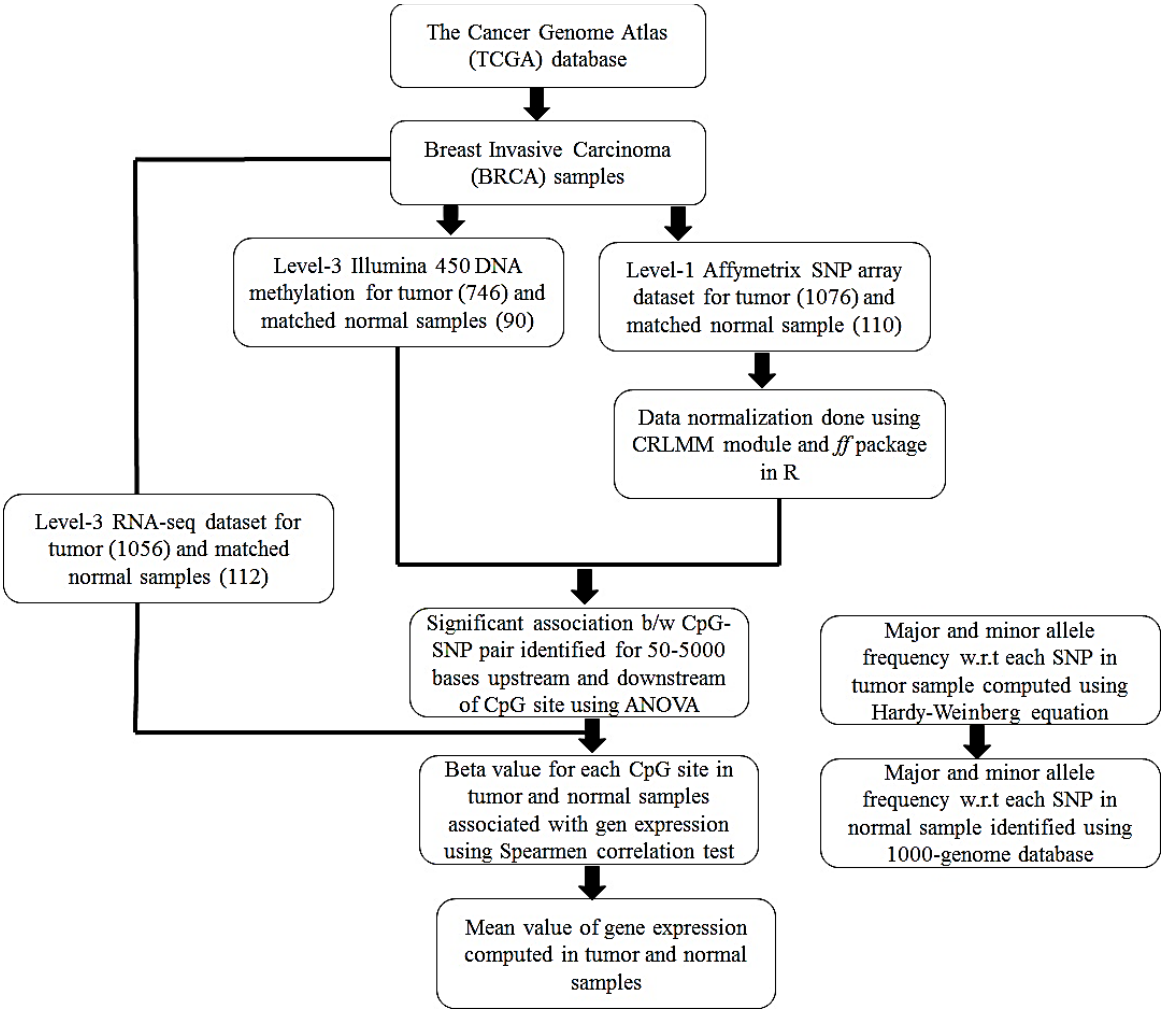

Figure 2.1 Detailed outline for identification of CpG-SNP pair candidates in the diagnosis of breast cancer. Complete study is based upon DNA methylation, SNP-array and RNAseq dataset.

Step 1. The "BED" format file of DNA methylation and SNP array data constituting 485,512 CpG sites and 905,422 SNPs, respectively were prepared as an input to the OverlapSelect program from UCSC Kent source library. This data integration is based on genomic positions that find all the SNPs lying in the vicinity of the CpG sites. The program was used to search the neighboring SNPs of the chosen CpG site using the following command:

overlapSelect -selectFmt=bed -selectRange -inRange -mergeOutput CpG.bed, SNP.bed output.bed.

Note: Only unique CpG-SNP pairs were analyzed for further studies.

Step 2. For each of the CpG-SNP pair obtained in step 2, we extracted the corresponding beta value for each CpG site and the genotype with respect to each SNP genotype from the common patient samples (731 samples in breast cancer). We applied the Analysis of Variance (ANOVA) to assess the statistical significance between the beta-values and the neighboring genotype [338].

Step 3. For each of the significant CpG site obtained from Step 2 was evaluated for differential methylation in tumor and matched normal samples in breast cancer. The beta value associated with CpG site in the tumor and matched normal samples sharing common space (86) were assembled. The significant difference in beta-values associated with respect to CpG site was computed based on paired *t-test*. The mean beta value associated with each significant CpG site across the sample was calculated. All the significant CpG site having mean beta value across the tumor sample greater than normal were retained, and the remaining were filtered out.

Step 4: For each significant SNP obtained in step 2, the frequency associated with major and minor allele in tumor sample were calculated using Hardy-Weinberg equation as follow [339].

$$p^2 + 2pq + q^2 = 1 \qquad \text{Equation 1}$$

$$p + q = 1 \qquad \text{Equation 2}$$

Here in the above equation p and q corresponds to major and minor allele, respectively. The respective allele frequency for each SNP in normal sample was obtained from "1000 genome population" database (http://www.1000genomes.org/). Moreover, the differential allelic distribution in tumor was compared with respect to normal in order to identify the percentage of germline and somatic mutation associated with each SNP.

Step 5. Finally, the differentially methylated sites obtained from step 3 were studied for their effect on gene expression. Spearman correlation test was implemented to study the significant association between DNA methylation and gene expression in tumor and normal sample, respectively. The complete analysis was carried at threshold p-value of 0.05 and correlation coefficient was computed. Moreover, the average value of expression of the gene associated with DNA methylation was calculated in tumor and normal sample.

2.3 Results

2.3.1 Interpretation of genotype, methylation and gene expression dataset in breast cancer

To study the correlation between the genetic and epigenetic codes in breast cancer comprehensively, Affymetrix genome-wide SNP array and Illumina Methylation450 dataset were merged for analysis. Prior to the processing of conjoint methylation and genotypic data, an intermediate step of pre-processing was carried out to filter non-significant CpG site across the genome. One dimensional matrix was constructed to compare the overall methylation pattern associated with 485, 5512 CpG sites across 740 tumors and 90 normal samples, respectively [Figure 2.2 a]. All the statistical parameters were set to compute the variance across these samples at the provided R-interface. Of the total 485, 5512 CpG sites, 448, 2886 CpG sites bearing zero variance across the samples were filtered out, and the remaining 37,2626 methylated CpG loci were processed for subsequent analysis. Illumina file, level 3 data substantiates β-value for given CpG site. Beta-value was then converted to M-value based on equation 2. M-value imparts better Detection and True Positive Rate (TPR) for both methylated and unmethylated CpG probes. Moreover, minimal threshold difference imposition enhances the performance of M-value in comparison to beta-value application.

$$M_i = log_2(\frac{\beta_i}{1-\beta_i}) \qquad \text{(Equation 2)}$$

The heterogeneity in methylation was processed for CpG sites positioned on each chromosome, and it was evident that the variation expanded profoundly in tumor samples to normal samples [Figure 2.3]. This hypervariability in the methylation pattern is characterized by the quantitative difference in aberrant methylation associated with CpG islands in different individual tumors. This increased variability across tumor sample was striking feature as it largely distinguished cancer from the normal cells. Polymorphism in allele distribution surrounding the CpG site may define the cause for the variation in methylation pattern. Thus, an integrative analysis based upon the identification of significant CpG-SNP pair was carried out for 731 tumor samples sharing a common interface for methylation and SNP array data [Figure 2.2b]. Finally, differentially methylated regions (DMRs) were correlated with the differential gene expression for 86 samples constituting both methylation and RNAseq data as obtained from TCGA [Figure 2.2a].

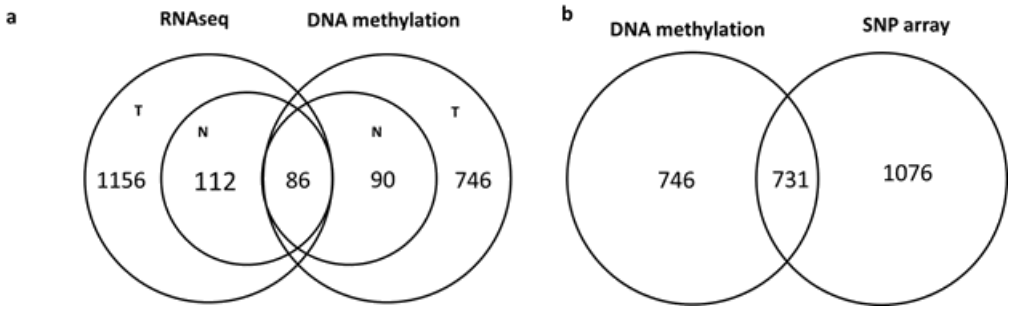

Figure 2.2 Venn diagram details about the BRCA dataset. (a) DNA methylation holds information for 740 tumors (T) and 90 matched normal (N) sample. RNAseq dataset constitutes details for 1156 tumor and 112 matched normal samples. There are 86 samples which share common space in both tumors and normal samples for DNA methylation and RNAseq dataset. (b) DNA methylation and SNP array datasetshare 731 tumor samples in the overlapping zone.

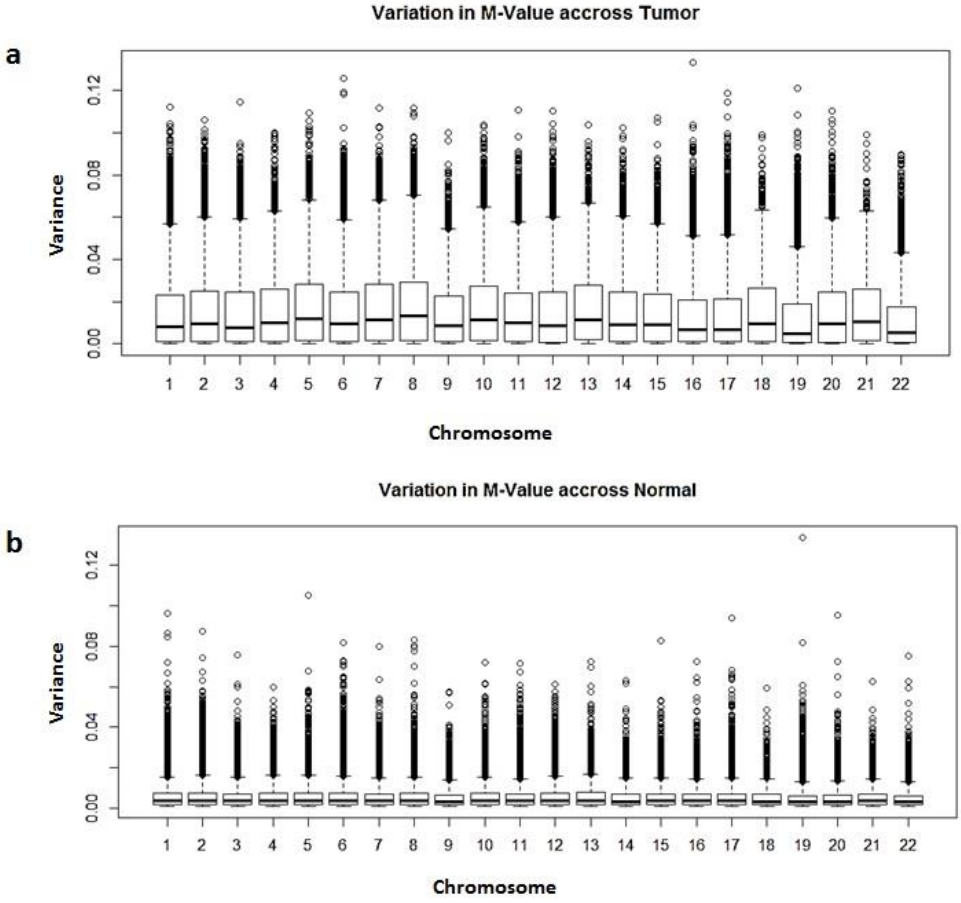

Figure 2.3 Genome-wide variation in methylation pattern associated with each chromosome in (a) Tumor samples b) Matched normal samples. Variation in methylation level has been identified to be comparatively high in tumor in comparison to normal population.

2.3.2 Mapping of significant CpG-SNP pairs in the identification of meQTLs

Screening of epigenomic modification at high resolution has disclosed a direct correlation between the underlying genetic variation and differential methylation pattern, subsequently defining the presence of meQTLs. In an attempt to identify SNPs genetically influencing methylation pattern, we integrated 905,422 SNPs and 372,626 CpG sites using ucsc tool-overlapSelect. Distribution of CpG-SNP pairs around the CpG site were identified within a base interval of 100-bases and the sliding window of 50-bases extending to the maximum boundary of 5000-bases in the upstream and downstream region. Beta value and the genotype associated with each CpG-SNP pairs were mapped across 731 samples sharing a common interface for SNP array and methylation data. An integrated two-dimensional matrix was generated for each CpG-SNP across the samples, and statistically significant CpG-SNP pairs were mapped based upon non-parametric one-way analysis of variance "ANOVA" [340]. There were a few instances in which multiple SNPs were mapped to a single CpG site. Figure 2.4 shows a bell-shaped distribution of CpG-SNP pairs by applying a sliding window. From the figure, it is evident that CpG-SNP density is high across 50-bps upstream and downstream of CpG-site. The overlapselect file constituting 7970 CpG-SNP pairs at 50-bps interval were evaluated for further analysis. The rationale for selecting the loci starting with 50-bases is to minimize the probe effect [341]. Illumina 450K methylation chip is identified to have a "probe effect" i.e SNP within 10bp of the CpG probe may be enriched in methylation quantitative loci (meQTLs). Moreover, DNA methylation locus are primarily associated with promoter regions (besides, inter/intra-genic regions), thus localization of SNP/SNPs may interfere the interaction of DNA methyltransferases enzyme (DNMTs) with CpG loci leading to anomalous DNA methylation [342].

Now considering the presence cis-acting elements mainly the enhancers mostly localized as far 5000bp. Presence of SNPs on enhancer may deregulate its functional property as well as its interaction with the promoter region. Moreover, the presence SNPs in the vicinity of histone marks (H3K4me3, H3K9-14Ac and H3K36me3) associated with active promoters, enhancer and transcriptionally active regions, interferes with the DNA methylation distribution leading to aberrant pattern.

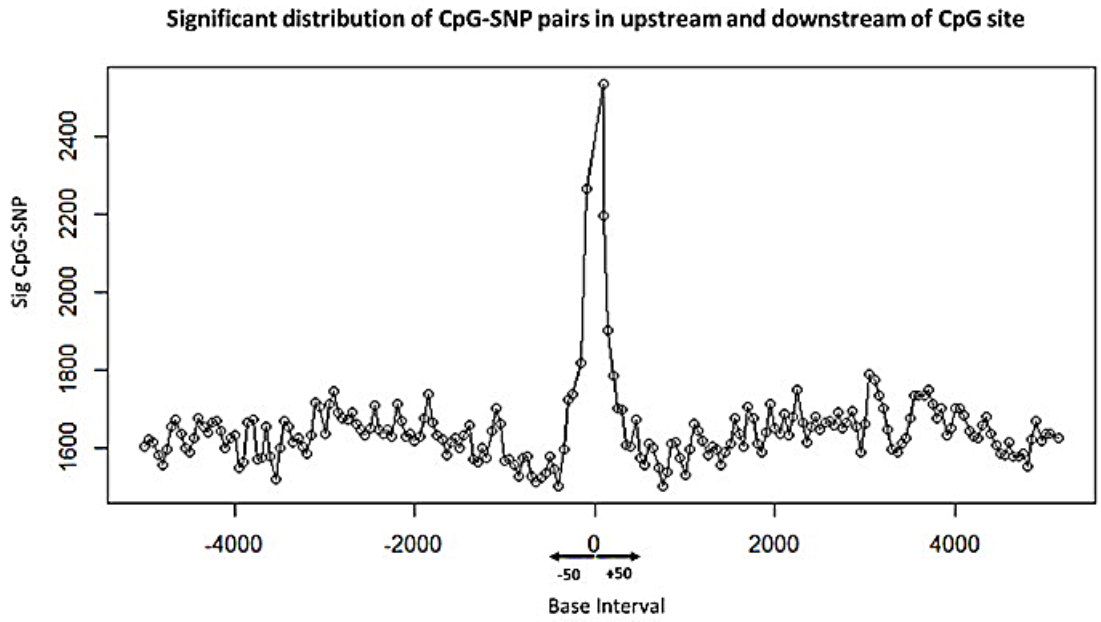

Figure 2.4 Significant distribution of CpG-SNP pairs around a given CpG site. The CpG-SNP density is identified to be high at 50 bases upstream and the downstream region.

2.3.3 Identification of differentially methylated regions in tumor and matched normal samples

Identification of DNA differential methylation distribution between the tumors and the normal cells is a landmark in understanding the processes underlying tumorigenesis. Studying of such differentially methylated region upholds in identification of diagnostic and prognostic marker. Specifically, the aberration in methylation pattern of a particular gene (or group of genes) is beneficial for the early detection of breast cancer and or stratification of tumors into subtypes. Extensive studies have been invested in identifying aberrantly methylated regions and correlating with tumor development or phenotype. However, studies done so far have focused on small sets of loci. Here, we have investigated the genome-wide pattern of differential methylation distribution based on a comprehensive study in breast cancer tissue and matched normal dataset. Localisation of differentially methylated regions (DMRs) have been investigated with respect to the polymorphism (SNPs) associated with each CpG sites. Influence of this local genetic variation in DNA methylation is called as cis-meQTLs. In order to examine the difference in methylation level in tumor with respect to normal, beta values associated with each 7970 significant CpG sites across was analyzed based upon student t-test. The statistical significance was set at threshold *p-value* < 0.05. The significant difference in methylation level across 86 tumor and matched normal sample led to the identification of 997 CpG sites of potential interest. These regions were distributed across all

chromosomes. Significant distribution of 997 differentially methylated CpG sites in order of their chromosomal location and –log 10 *p-values* have been depicted in Manhattan plot [Figure 2.5]. The mounting *p-value* in the plot beyond the threshold led to the identification of those CpG loci holding marked difference in methylation pattern in tumor with respect to normal. This differential methylation pattern was a consequence of the variation in allelic distribution. The quantile-quantile (Q-Q) plot in terms of $-\log_{10}$ (*p*-values) clearly depicts the association between the variable allelic distribution and differential methylation [Figure 2.6]. The major and the minor allelic frequency for each SNP was computed in tumor and normal samples by following HWE and 1000 genome project of population genetics, respectively. We enlisted the top 3 CpG-SNP pairs strongly associated with differential methylation as; cg02058408:rs9891975, cg05388880:rs4421026 and cg25198340:rs17235834 [Table 2.1]. Differential methylation pattern in tumor with respect to normal was a consequence of difference in major and minor allele frequency. While the minor allele frequency associated with SNP rs9891975 and rs4421026 was high, elevation in major allele frequency was seen with respect to SNP rs17235834. In the upcoming section, we elaborate the detailed study of the correlation between differential methylation, allelic distribution and the gene expression.

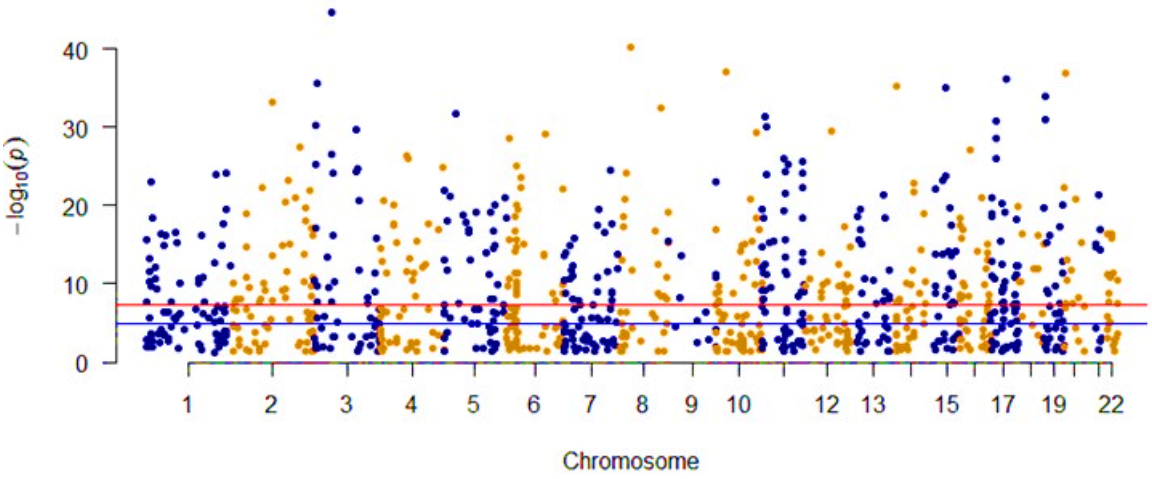

Figure 2.5: Manhattan plot presents the association with $-\log_{10}$ (*P*-values) for each differentially methylated CpG sites (y-axis) in the tumor in comparison to normal samples in the order of chromosomal position. The red and the blue line indicates the threshold $-\log_{10}(1 \times 10^{-4})$ and $-\log_{10}(0.2)$, respectively, for genome-wide statistical significance.

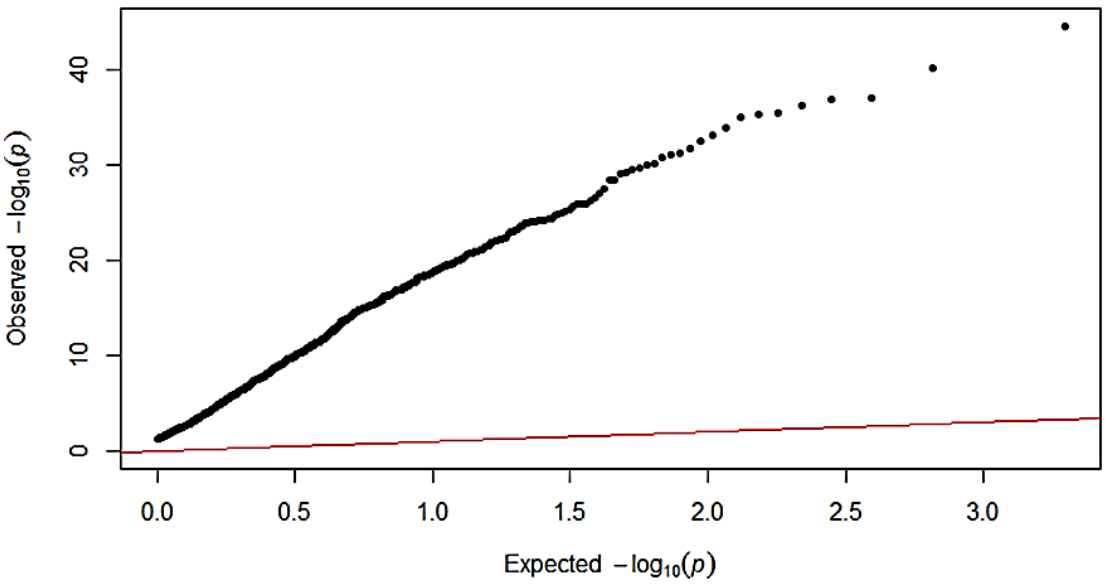

Figure 2.6: Quantile-Quantile (Q-Q) plot with respect to observed versus expected p-values. Q-Q plot of $-\log_{10}$ (*p*-values) depicts the association between differential methylation (DMRs) and the allelic variation associated with SNPs. The observed quantile is higher than the expected value, disobeying the null hypothesis.

Table 2.1 Allelic distribution with respect to top 3-SNPs and its effect on methylation in tumor and normal sample

S.No.	CpG ID	SNP ID	P-value	Mean-value Tumor	Mean-value Normal	Difference	Gene	Allele Frequency Normal		Allele Frequency Breast Cancer	
								Major	Minor	Major	Minor
1	cg02058408	rs9891975	1.29E-31	0.57997	0.16217	0.41865	Intergenic	A 92%	G 8%	A 4 %	G 96%
2	cg05388880	rs4421026	1.18E-25	0.81975	0.51601	0.30373	DCTD	C 99%	T 1%	C 0.6%	T 99.4%
3	cg25198340	rs17235834	2.25E-45	0.84669	0.54799	0.29870	IL17RD	G 98%	A 2%	G 4%	A 96%

2.3.4 Establishing the correlation between allelic distribution, differential methylation and gene expression in the diagnosis of breast cancer

Finally, the significant association between genetic variations and DNA methylation was extended to gene expression by incorporating the RNA-seq dataset. The association between differential methylation and gene expression was measured by Spearmen correlation coefficient [343]. A complete analysis was carried out for 86 samples overlapping with methylation and RNA-seq data set. All data were quantile normalized prior to analysis. Of the total 997 significant CpG-SNP pairs influencing differential methylation, 713 had associated gene information while, the remaining were localized in the intergenic regions. The relationship between DNA methylation and gene expression were analysed in terms of beta values and log2-transformed fold change in gene expression for both tumor and normal samples.

From the conjoint analysis, 16 of total 713 CpG-SNP pairs showed a significant nominal correlation between differential methylation and gene expression. Interestingly, 3 CpG-SNP pairs; cg08710564:rs4929917, cg08306955:rs16890134 and cg14482998:rs9387025 holds high negative correlation with gene expression of *ST5*, *CMAH* and *FYN* genes, respectively. Further ahead, we disintegrated the above analysis in order to have a clear vision of the individual factors (SNP, methylation and gene expression) being correlated. The variable pattern in the allelic frequency distribution is a remarkable feature in understanding the disease etiology. For example; major allele T of SNP rs4929917 was associated with increased methylation level of CpG site cg08710564. Major and the minor allele frequency associated with SNP rs4929917 in normal population was identified to be 5% and 95%, respectively while, in breast cancer the allelic frequency flipped to 96% and 4%, respectively [Figure 2.7a]. The difference in the allelic distribution led to variation in methylation pattern in tumor and normal samples. Methylation distribution with respect to the CpG loci cg08710564 in tumor ranged from 65-85% and was higher than the normal sample [Figure 2.7b]. This differential methylation in tumor led to downregulation of *ST5* gene. Based upon the spearman correlation analysis, we found a significant inverse correlation between differential methylation and fold change in mRNA expression of *ST5* gene in breast cancer (*coefficient* $r = -0.42$, $p < 0.0001$) [Figure 2.8a]. An average fold change in gene expression in tumor with respect to normal is shown Figure 2.8b. In another example, very high frequency of minor allele T associated with SNP rs16890134 led to the differential methylation of the CpG site cg08306955. The frequency of allele “T” is identified to be as high as 99% in the breast cancer patients however, the frequency was low (1%) in the normal population [Figure 2.9a, b]. This differentially methylated CpG

site located at *5'* UTR region was responsible for downregulation of *CMAH* gene, and the correlation coefficient was identified to be -0.44 at p-value < 0.0001 in tumor sample [Figure 2.10a]. Figure 2.10b shows the mean fold change in gene expression of *CMAH* gene in tumor and normal sample. Finally, we also identified that nearly equal distribution of major and minor allele can also affect methylation as well as gene expression. The frequency allele A and G associated with the SNP rs9387025 in breast cancer was found to be 57% and 43%, respectively [Figure 2.11a]. This allelic distribution in breast cancer was linked to the hypermethylation of CpG site cg14482998 associated with the intron variant of *FYN* gene [Figure 2.11b]. Integrated analysis of DNA methylation and *FYN* transcriptome revealed a reverse correlation between differential methylation and mRNA expression (r = - 033, p-value < 0.01) in tumor sample [Figure 2.12a]. Average fold change in gene expression in the tumor and matched normal sample is shown in Figure 2.12b. Moreover, the genotypic distribution of these SNPs (rs4929917, rs16890134, rs9387025) in tumor and matched normal revealed that most of these mutations are germline while, only smaller percentage fall under somatic mutations [Figure 2.13].

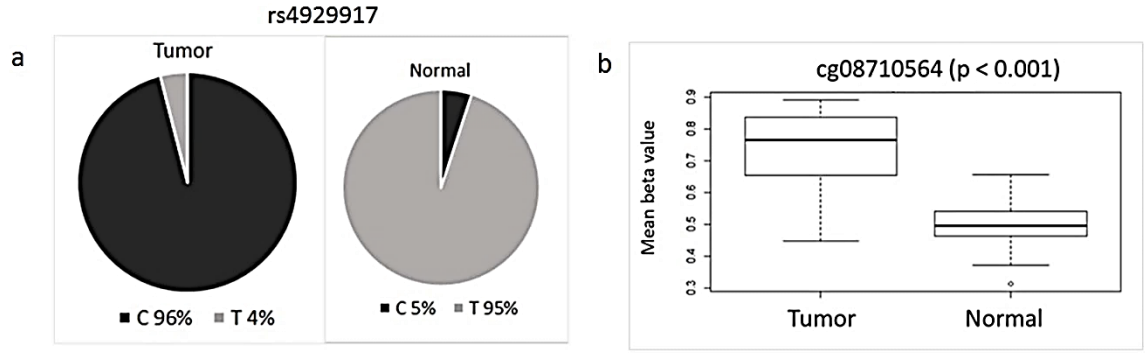

Figure 2.7 (a) Major and minor allele frequency distribution of "C" and "T" associated with SNP rs4929917 in breast cancer and normal population. Major allele frequency is comparatively high in tumor as compared to normal sample. (b) The methylation associated with CpG site cg08710564 in tumor ranges from 65%-82%, while, in normal the distribution ranges from 48-52%.

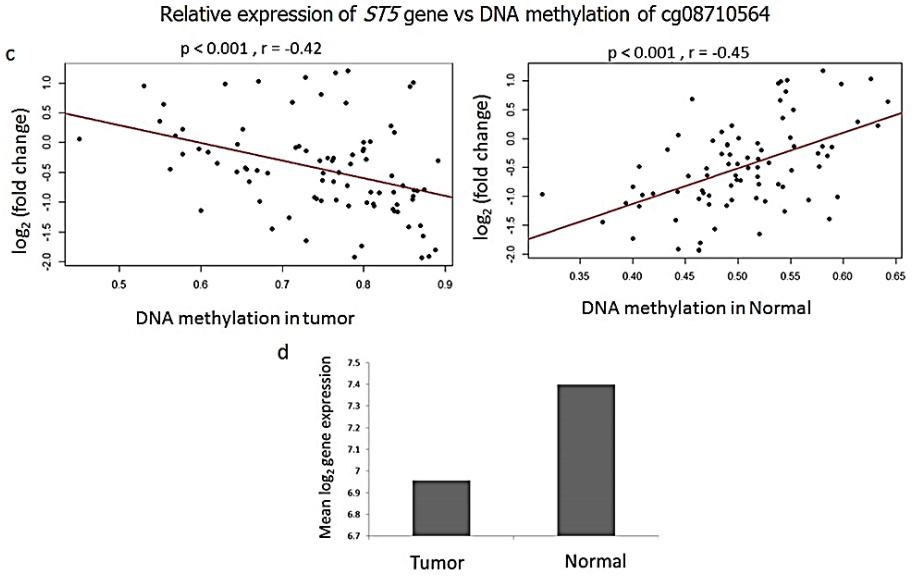

Figure 2.8 Spearmen correlation with respect to fold change in gene expression and DNA methylation in breast cancer in comparison to a normal cell. (a) Fold change in gene expression of *ST5* gene has a negative correlation with respect to DNA methylation (cg08710564) in a breast cancer while, it is positive in normal sample. The correlation coefficients were identified to be -0.42 ($p < 0.0001$) and 0.49 ($p < 0.001$), respectively. (b) Average fold change in gene expression of *ST5* gene in tumor and normal was 6.9 and 7.4, respectively.

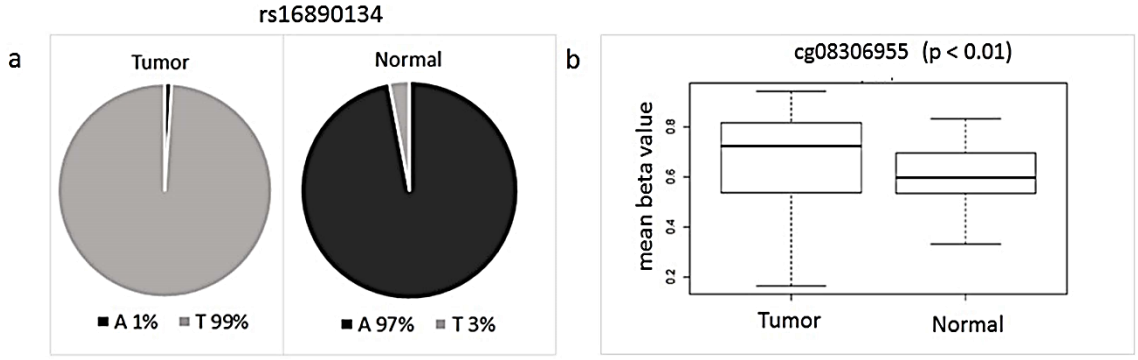

Figure 2.9 Major and minor allele frequency distribution of "A" and "T" associated with SNP rs16890134 in breast cancer and normal population. Minor allele frequency is comparatively high in tumor in comparison to the normal samples. (b) The methylation associated with CpG site cg08306955 in tumor ranges from 55%-80%, while, in normal the distribution ranges from 55-65%.

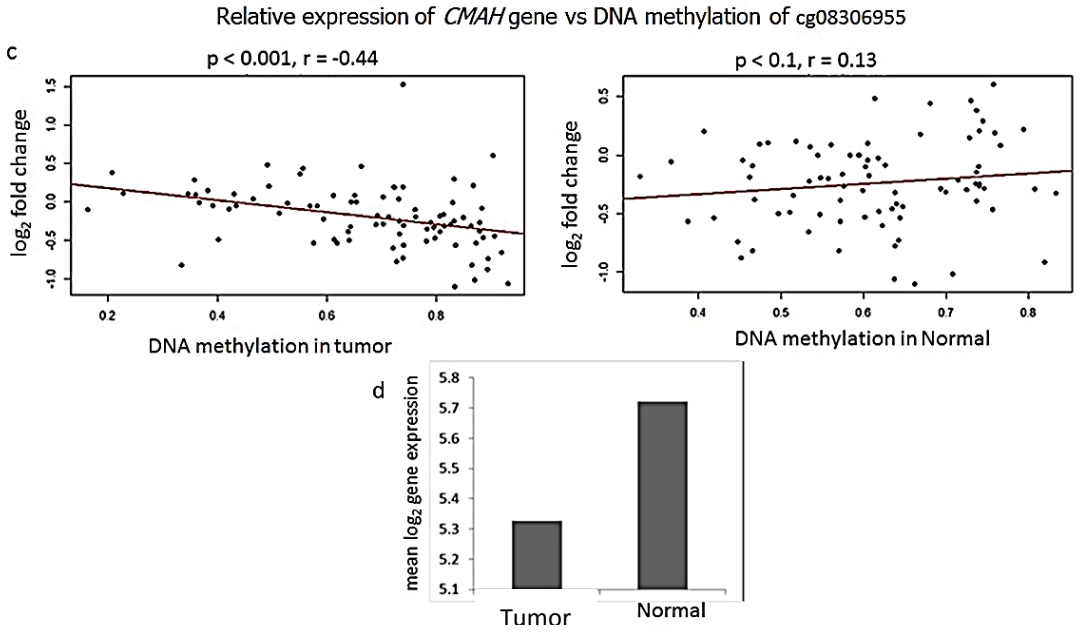

Figure 2.10 Spearman Correlation with respect to fold change in gene expression and DNA methylation in breast cancer in comparison to a normal cell. (a) Fold change in gene expression of *CMAH* gene has a negative correlation with respect to DNA methylation (cg08306955) in a breast cancer while, it is positive correlation in normal sample. The correlation coefficients were identified to be -0.44 ($p < 0.0001$) and 0.13 ($p < 0.1$), respectively. (b) Average fold change in gene expression of *CMAH* gene in tumor and normal is 5.3 and 5.7, respectively.

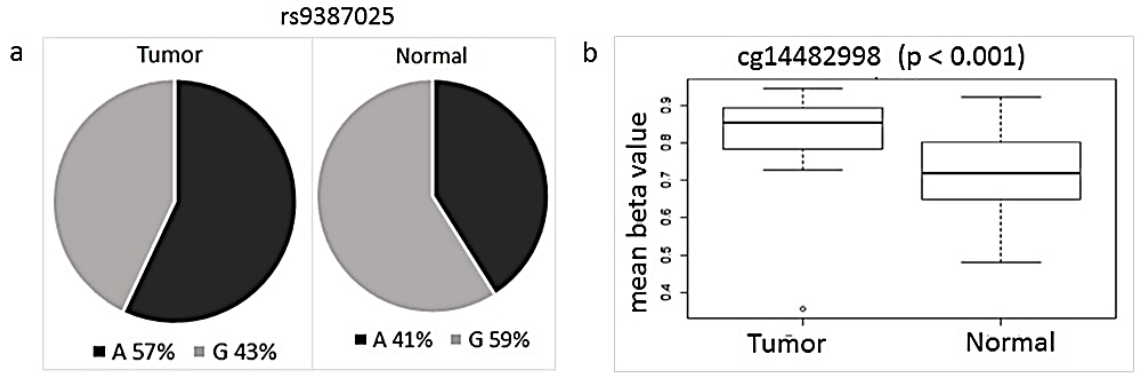

Figure 2.11: Major and minor allele frequency distribution of "A" and "G" associated with SNP rs9387025 in breast cancer and normal population. Both Major allele and Minor allele holds nearly equal frequency in breast cancer and in the normal population. (b) The methylation associated with CpG site cg14482998 in tumor ranges from 78%-88%, while, in normal the distribution ranges from 65-75%.

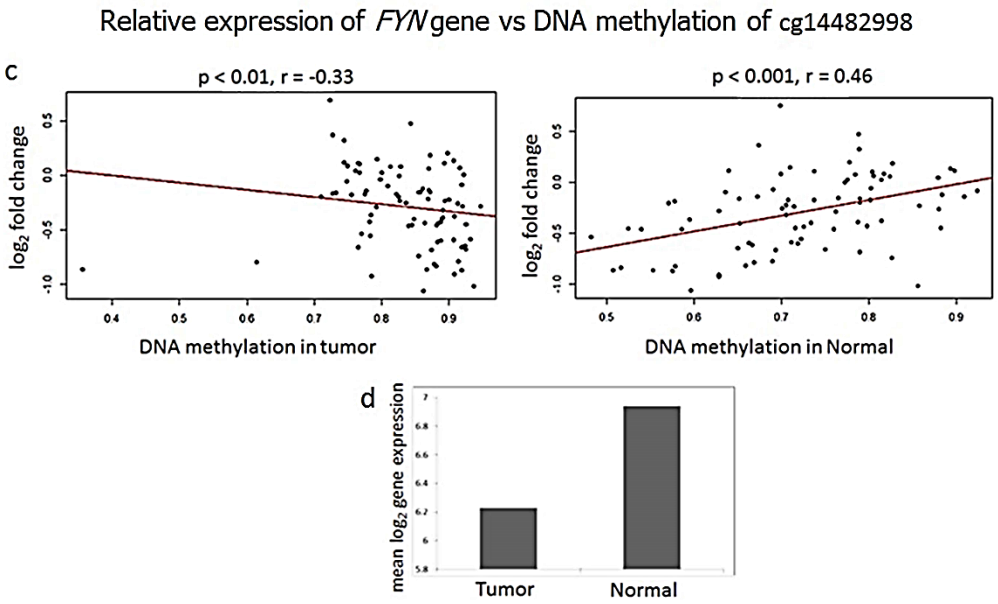

Figure 2.12 Spearman correlation with respect to fold change in gene expression and DNA methylation in breast cancer in comparison to normal cells. (a) Fold change in gene expression of *FYN* gene has a negative correlation with respect to DNA methylation (cg14482998) in breast cancer while, it is positive correlation in normal sample. The correlation coefficients were identified to be -0.33 ($p < 0.01$) and 0.46 ($p < 0.0001$), respectively. (b) Average fold change in gene expression of *FYN* gene in tumor and normal is 6.2 and 6.8, respectively.

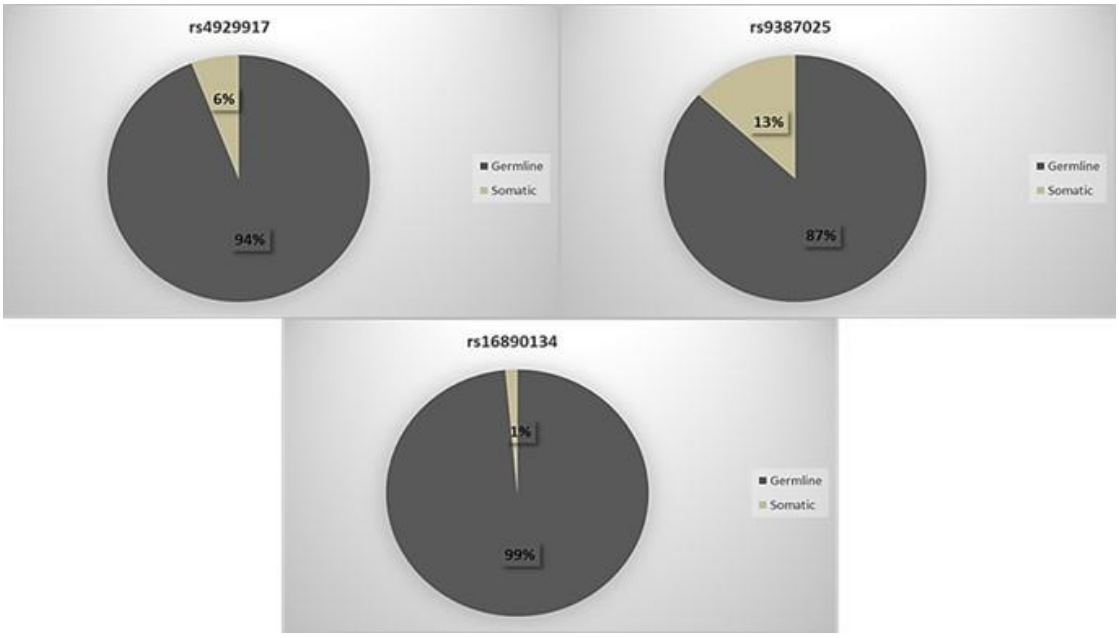

Figure 2.13 Somatic and Germline Mutation: Differential genotypic analysis (Tumor with respect to Normal) of the SNPs (rs4929917, rs9387025, rs16890134) identified from the above studies shows that the mutations are inheritable (germline) in comparison to the somatic.

2.4 Discussion

The inter/intra-tumor heterogeneity in breast cancer possess an important impediment to the targeted therapy [344-346]. Fuelled by Darwinian Theory of evolution regulating the disease, these variability leads to the emergence of resistance in breast cancer cells when being subjected to the selective pressure [347-349]. Indeed, the analysis of a cohort of candidate genes in population-based and in the pedigree analysis would allow in tracing possible clause in the pathway analysis of the specific type of cancer [350, 351]. Identification of common pathway associated with genetic heterogeneity would lead to the identification of novel targets for the early diagnosis of cancer [352-354].

Comprehensive mapping of a genetic variant of the genome between individuals discordant for certain phenotype has revealed a plethora of SNPs having a significant association with the diverse phenotype, including cancer [355, 356]. Despite the success of GWAS in identification of variable loci in disease diagnosis, a substantial proportion of the causality remains inexplicable. On a similar account, epigenetics studies characterized by DNA methylation also provides novel insight into low-frequency drivers of breast cancer [357-359]. DNA methylation reflects phenotypically significant difference in gene transcription making the profile based diagnostic test to be more substantial and reproducing. However, several studies reflect that the diagnostic analyses based on methylations were interrupted by mutations.

Till date, the presence of these mutations were overlooked while, it is surprising in view that more than 1,000 of cell lines recently being analysed reveal the presence of numerous mutations being associated with epigenetic modifiers (DNA methylation) [309]. The fact that pinnacle of the hierarchy of genes is being regulated by the epigenome,, the mutations in these genes will probably affect multiple pathways in relevant to the cancer phenotype. Consequently, we added layer information connecting polymorphism and variance in gene expression. We introduced epigenetic-mediated gene regulation as a potential intermediate connecting genotype-phenotype association. The high-resolution DNA methylation data was integrated with single nucleotide polymorphism resulting in a catalog of genotype-epitype association for identification of diagnostic marker in breast cancer.

Taken together, the massive integration of Affymetrix SNP array data and Illumima 450 DNA methylation data, we identified polymorphic sites potentially regulating CpG methylation in breast cancer. Most of these polymorphic alleles (SNPs) are predominantly located in a noncoding region depicting their close association with epigenetic modification primarily DNA methylation and aids in understanding its functional significance in etiology of breast cancer. In our present study, we have mainly observed the cis-regulation of CpG site by the genetic variant. The genome-wide study of

methylation pattern across 37, 2626 CpG sites showed an extensive range of variability in methylation distribution in the order of the chromosome. This increased variability across tumor sample was striking feature as it largely distinguished cancer from the normal cells. The variable pattern of methylation was analyzed across 740 tumors, and 90 matched normal sample obtained from TCGA cohort. Of the total 485, 5512 CpG site, only 37, 2626 methylated site displayed significance variance across the data set, was considered for further analysis. This variance in methylation was explained by associated SNPs in the vicinity of CpG site. Thus, allelic polymorphism (SNPs) having significant influence DNA methylation is called DNA methylation quantitative trait loci (meQTLs). The presence of statistically significant CpG-SNP pair around a given CpG site was interrogated at a base interval of 100nt with an overlapping window of 50nt extending to 5000nt in the upstream and downstream regions. The maximum density of CpG-SNP pair hovered around 50-bps upstream and downstream region of each CpG site. Of the total, 7970 CpG-SNP pairs were significantly associated with CpG site at a base interval of 50-bps both in upstream and downstream regions. These 7970 CpG sites were further evaluated to identify how many of them are associated with differential methylation level in tumor with respect to matched normal samples. Out of the total, 997 CpG sites loci exhibited remarkable difference in methylation pattern. This differential methylation was explained by the variable allelic distribution associated with each SNPs in the vicinity of CpG loci. The flipping of major and minor allele frequency in breast cancer and the normal population was an incredible feature that explained underlying differential methylation in tumor and normal sample. We enlisted the top 3 SNPs; rs9891975, rs4421026 and rs17235834, strongly regulated methylation level of CpG sites cg02058408, cg05388880, and cg25198340, respectively. However, these differential methylated CpG sites did not have significant effect on gene expression. Therefore, in our subsequent analysis, we extended our study to see the effect of differential methylation on gene expression in tumor and normal samples. Taking SNP, DNA methylation, and gene expression together, we identified 16 genes being influenced by difference in the methylation level. However, 3 genes showed a strong negative correlation with increase in DNA methylation in tumor sample. In particular, we identified 3 major class of allelic distribution which could regulate methylation pattern which in turn affected gene expression. In the first case, we have the example of SNP rs4929917, where the increase in frequency of major allele “C” was associated with increased methylation of CpG site cg08710564 in tumor sample. This increase in methylation resulted into the decreased expression of *ST5* gene. However, the increase in methylation level was associated with increased expression of *ST5* gene in normal sample. Suppression of tumorigenicity 5 (*ST5*) gene is located on chromosome 11 and has the ability to suppress tumor of Hela cells in nude mice [360]. This gene encoded a protein such that its C-terminal shares

homology with Rab 3 family of small GTP-binding proteins. This protein preferentially binds to SH3 domain of the c-Ablkinase and acts as a regulator of the MAPK1/ERK2 kinase, contributes in reducing the tumorigenic phenotype in cells [361]. From the previous studies, it has been reported that aberrant silencing of this gene is of great risk in breast, lung and cervical cancer development [360-362]. In the second case, the elevation in minor allele frequency has been identified to be of high risk in breast cancer. The high frequency of minor allele "A" associated with SNP rs16890134, located in the downstream region causes hypermethylation of CpG site cg08306955, subsequently leads to down-regulation of *CMAH* gene in the tumor sample. Cytidine monophosphate-N-acetylneuraminic acid hydroxylase (*CMAH*) gene having loci on chromosome 6 encodes for the sialic acid which a component of carbohydrate chains of glycol-conjugates and actively participates in ligand-receptor and cell to cell interactions [363, 364]. The carbohydrate is actively synthesized and secreted by oral and mammary carcinoma cells promoting to malignancy [365-368]. Finally we also identified that nearly equal distribution of major and minor alele can have signficant effect of differential methylation pattern in tumro wiith repsect to normal sampele. SNP rs9387025 having nearly equal allele frequency for "A" (57%) and "G" (43%) causes increase methylation level of CpG site cg14482998 in tumor sample. The increased methylation level resulted in decreased expression of FYN (tyrosine kinase) gene. The gene holds dual property of oncogene and tumor suppressor gene. *FYN* tyrosine kinase gene located on chromosome 6 has been reported new candidate tumor suppressor in prostate cancer and gastric cancer [369-371]. The SNP mentioned so far have been identified to be germline in comparison to somatic mutation.

The positive correlation between DNA methylation and gene expression in the normal sample is condition-specific [342, 372, 373]. Several studies have established the concept of negative correlation between DNA methylation and gene expression at transcription start sites (TSSs). However, the explained concept cannot be extrapolated for CpGs located in the intergenic and intragenic regions. DNA methylation and gene transcription follows a non-linear equation. In general the DNA methylation has been identified to block the initiation of the transcription but not the elongation. In fact several inter/intragenic nucleosomes are associated with histone tri-methylation marks H3K36me3 which recruits DNMTs facilitating the methylation of inter/intragenic regions [374]. Besides, DNA methylation also regulates gene transcription by incorporating molecular mechanism through alternative promoter, enhancer and non-coding RNA. More recently, several studies have observed positive correlation between inter/intragenic DNA methylation and gene expression in context of cellular development, differentiation and in cancer cells [67, 372, 375-378].

Our results reveal the new findings based upon genetic variability contribute to differential methylation and gene expression in breast cancer. Differential expression of *ST5*, *CMAH* and *FYN* gene and the associated CpG-SNP pair will contribute to the major finding in the early diagnosis of breast cancer. These results will lead to the discovery of a novel mechanism that determine gene-specific DNA methylation and the functional effects of polymorphism on disease phenotypes including cancer.

Chapter 3

To decipher how single nucleotide polymorphisms affect DNA methylation at nearby CpGs and impact breast cancer prognosis among individuals

3.1 Introduction

The variation in the gene expression transforms the cellular programming from normal to a diseased state. The multiple genetic circuits within a cell creates a characteristic signature profile of gene expression endorsing each cell a unique identity. The gene-expression-based signatures have been successfully implemented in classifying the breast cancer into different subtypes [379, 380]. Similarly, approaches based upon genome-wide DNA methylation profiling identified breast-cancer-specific methylation signatures that correlate with specific clinical outcomes [44]. In addition to the diagnostic potential, aberrations in DNA methylation profile regulates gene expression dictating tumor recurrence and overall survival in breast cancer and their subtypes [229, 381-385]. The prognostic potential of genes mainly *FLRT2* and *SFRP1* have been identified to be regulated by DNA methylation and are enriched in ER1/luminal B of breast cancer. However, the expression of specific genes linked to immune function such as *CD3D, CD79B, CD6, HCLS1, HLA-A* and *lAX1* have been identified to be consistently associated with recurrence-free survival (RFS) and overall survival (OS) in ER2/HER22 subtypes of breast cancer [386-388]. Further ahead the combination of methylated genes such as *GSTP1, FOXC1,* and *ABCB1* has been correlated with respect to the survival of the patients [389]. The downregulation of DNA methylation have been significantly correlated with the expression of *BCAP31* and *OGG1* genes and have shown significant association with the survival in a large cohort of breast cancer patients [390]. Besides, the differential methylation of CpG islands proximal to the genes regulating cell cycle and proliferation (*HDAC4, KIF2C, Ki-67*, and *UBE2C*), angiogenesis (*BTG1, KLF5, VEGF*)

and cell fate determination (*LHX2*, *LXH2*, *OLIG2*, *SPRY1*) possess significant prognostic values independent of subtypes and clinical features [391].

GWAS have identified a large number of genomic variants associated with complex diseases, including breast cancer [356, 392, 393]. However, most of the disease-associated genomic variants that have been reported in the literature so far are predominantly located in the intergenic or intronic regions of the genome [394]. Furthermore, numerous studies have noted that GWAS haplotypes are enriched in regulatory elements that are concordant with the disease phenotypes [395]. Therefore, it is highly likely that most of the disease-causing genomic variations act by altering gene regulation, such as transcription factor binding and DNA methylation, rather than directly affecting protein function.

Despite the advances in sequencing and availability of multi–omics datasets [332, 396], finding causative and prognostic genetic variants for complex diseases, such as breast cancer, remains challenging. Thus, a robust method of associating genomic variants, such as SNPs, in regulatory regions, such as CpG islands, with corresponding DNA methylation alterations is required [397]. The influence of these genetic variants on DNA methylation level was referred to as cis-methylation quantitative trait loci (cis-meQTLs) [342, 398]. Here, we report the joint effect of meQTLs on clinicopathological variables for identification of prognostic biomarkers, their clinical validity and the extent to which they capture the pathological difference between breast cancer prognostic groups using these external independent studies.

3.2 Materials and Methods

Details for Illumina 450K methylation, SNP array and RNAseq TCGA dataset incorporated in the present study have been detailed in the previous section of the diagnostic analysis.

3.2.1 Clinical data

A central premise of cancer treatment resides in deciphering the genotypic information into phenotypic expression. The clinicopathological data from TCGA aids in investigating the risk associated with polymorphism and associated phenotypic aberration. These clinical data have been collected by Biospecimen Core Resource (BRC) with respect to the participant sample. The data are available in XML and flat file biotab format. We obtained the biotab file for BRCA that spanned the detail for 1035 tumor patients. The clinical data recapitulates; age, gender, menopause status, race and ethnicity, history of neoadjuvant treatment, histopathological subtypes and tumor stage corresponding to each patient as major details. Length of survival of each patient was measured from the date of treatment to the date of last follow-up or death. Vital statistics enumerated the death status

of breast cancer patients. Patients alive on the last follow-up date were considered as censored.

3.2.2 Procedure for the identification of CpG-SNP pair associated with the prognosis in breast cancer

Figure 3.1 shows an outline of the procedure for identification of regulatory CpG-SNP pair involved in the risk associated with the survival of breast cancer patient. We describe the details in the following steps.

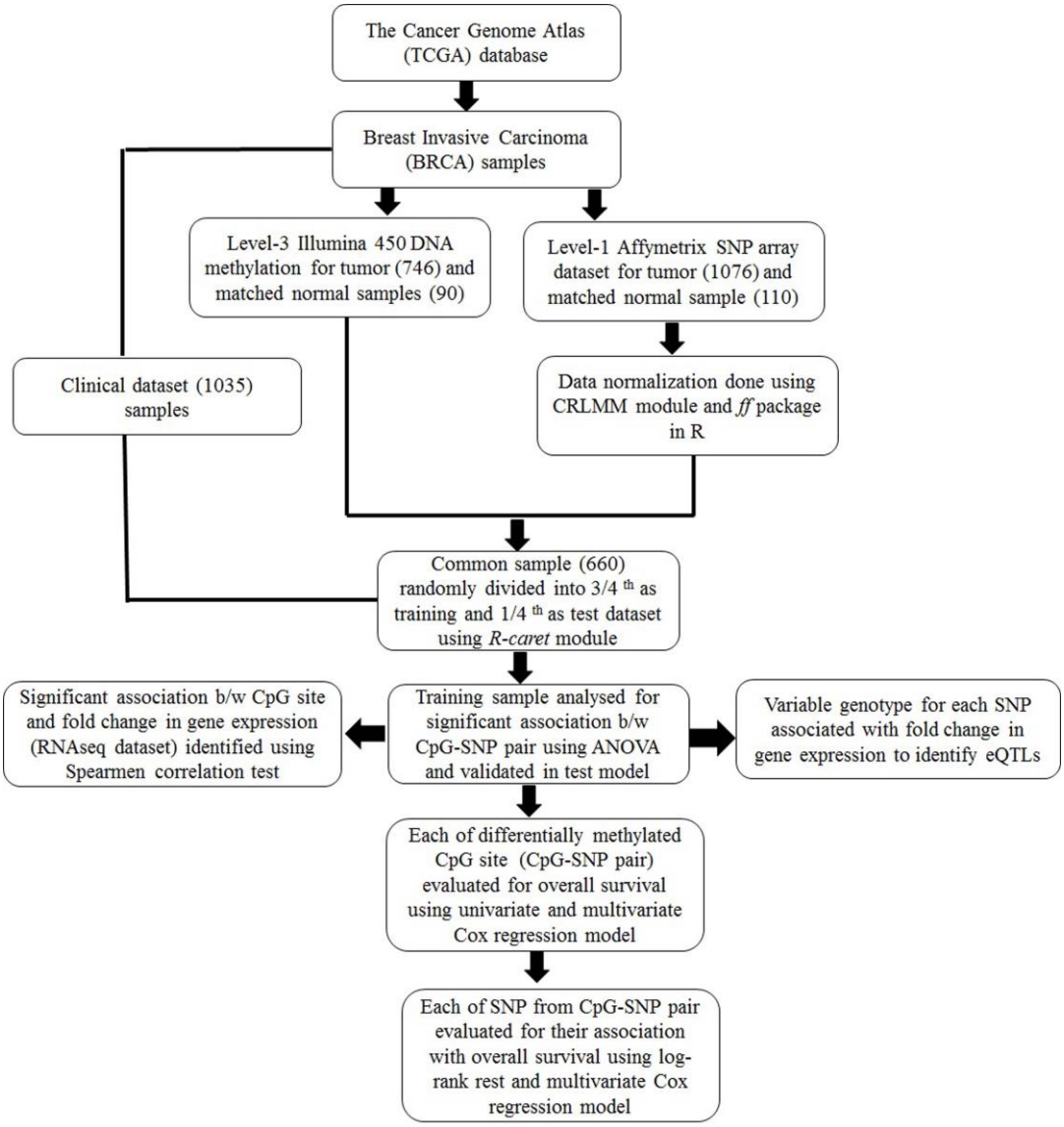

Figure 3.1 Detailed outline for identifying significant effect of CpG-SNP pair on the overall survival. It also includes in finding the candidate risk SNPs in the breast cancer prognosis. The individual CpG sites and SNPs have also been correlated with the gene expression. This process utilizes DNA methylation, SNP-array, RNAseq and clinical data.

Step 1: In order to study the synergistic effect of methylation and the associated polymorphism in regulating the survival of the breast cancer patients, 660 samples sharing the common space between DNA methylation, SNPs and clinical dataset were randomly split into training and test model [Figure 3.2 a]. The caret package of R (http://caret.r-forge.r-project.org/) was implemented to group the ¾ of the samples (486) into training and ¼ (164) as testing based on the vital status of the patients from the clinical data [Figure 3.2 b].

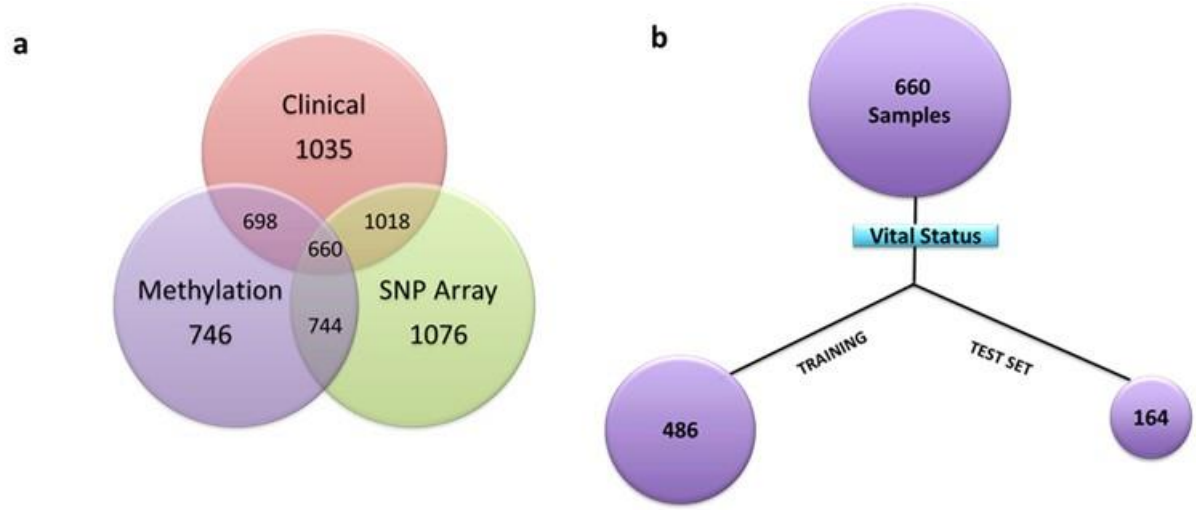

Figure 3.2 (a) Venn diagram details about the DNA methylation, SNP array and clinical samples across the tumor patients. (b) The tumor sample overlapping across the three datasets is grouped into the 75% training, and 25% test set based on the vital status.

Step 2: For each of the 7970 CpG-SNP pair located at 50-nt upstream of downstream of CpG site (as described in step 1 of diagnostic section), the training model was built across 486 samples. The beta value associated with each CpG site and the variable genotype (AA, AB, BB) with respect to each SNP were selected as the features in the training model. The findings in the training model were validated across 164 samples in an exclusively independent test model.

Step 3: The significant association between the beta-values or the proportion of methylation at each CpG site and the variable genotype associated with each SNP was computed based upon non-parametric one-way analysis of variance (ANOVA). Here the β-values were modeled as a linear function with respect to alleles (AA, AB, BB). The complete analysis was carried out at *R-interface* at threshold p-value of 0.05. Each of the SNP having a significant association between DNA methylation was labeled as meQTLs. The finding of these meQTLs in training model was validated in the test model.

Step 4: For each of the significant CpG-SNP pairs from step 3, we analyzed the significant association between beta-value with respect to each CpG and the gene expression. We extracted the corresponding beta-values with respect to each CpGs and log2-tranformed fold change in gene expression (tumor w.r.t normal) for 86 patient samples. The association between the DNA methylation and the fold change in gene expression was analysed based upon Spearman-correlation test.

Step 5: For each of the CpG-SNP pair from step 3, we also studied the significant effect of the allelic polymorphism on the gene expression. We extracted the variable genotype (AA, AB, BB) associated with each SNP and log_2-tranformed fold change in gene expression. We then applied ANOVA to assess the statistical significance between each SNP genotype and its neighboring gene expression [399]. Moreover, the mean fold change in gene expression was calculated with respect to the genotype associated with each SNP. This association between the differential gene expressions with respect to allele was labeled as expression quantitative trait loci (eQTLs).

Step 6: For each of the significant CpG-SNP pair (test set) from step 3, the differentially methylated CpG sites were assessed for the risk associated with the survival of the breast cancer patients. The complete analysis was based on the univariate and multivariate Cox Proportional Hazard (PH) model [400-402]. It is a regression model which describes the relation between the event incidence expressed as hazard function and a set of covariates. The hazard parameter is denoted by h (t) or λ (t) and is defined as the risk associated with the survival of the diagnosed cancer patient in a given time *t.* Mathematically the Cox Model is represented as;

$$h\,(t) = \; h_0(t) \;\; \times \exp\{b_1 x_1 + \; b_2 x_2 + \cdots b_n x_n\}$$

where, the hazard function $h\,(t)$ is determined by a set of *n* covariates $(x_1, x_2 \ldots\ldots x_n)$ and the impact of each variable is measured by the respective coefficients $(b_1, b_1 \ldots\ldots b_n)$. Here in the equation h_0 is the baseline hazard, while, $h\,(t)$ is the hazard function variable over time *t*. The Kaplan-Meir survival curve was plotted to classify the patients into high and low risk, respectively.

Step 7: Besides, the SNPs were also analysed to study their effect on overall survival. We extracted the variable genotype details associated with each SNP and clinical details including the vital status (patient alive or dead) and the date of the last follow-up. The training model was built for each of the SNP across 486 patients. The findings in the training model were validated in the test model. A complete analysis was carried out based on the log-rank test [403, 404]. All the significant SNPs identified in the test model

were subjected to multivariate Cox regression analysis to visualize their cumulative effect on overall survival.

3.3 Results

3.3.1 Identification of methylated probes or loci differing in genotypes

In the previous section of our study, we have described the polymorphism linked to the CpG loci results in differential DNA methylation in tumor versus normal cells. However, the accumulation of genetic variations on certain chromosome remains dormant and needs to be excavated for identification of meQTLs linked to disease progression. In our present analysis, we mainly elaborate the pattern of polymorphic allele distribution (*AA*, *AB*, and *BB*) and their influence of differential methylation exclusively in breast cancer patients. Considering the close proximity between the genetic variability and DNA methylation, the comprehensive analysis of the overlapping layers expands our knowledge in understanding the association of genetic variability with disease etiology. Realizing the fact that a large portion of cancer-related SNPs ispositioned in the noncoding region holds substantial functional impact, the coaxial analysis of genotype-epitype interactions will facilitate identification of novel prognostic markers.

In order to determine the association of genotype-epitype interactions comprehensively, we integrated the high-resolution Affymetrix SNP array and Illumina 450k DNA methylation platforms, analyzing 905,422 SNPs and 485,512 CpG sites. The training data set comprising of 486 samples was constructed across the CpG-SNP pairs. For each of the benchmark data set, its training and test were used as exclusive subsets. The predictive model was built in training data set and validated in the test data set bearing 164 samples. Based on the analysis carried out by overlapSelect tool, a total of 7970 CpG-SNP pairs were identified at a base interval of 50bps upstream and downstream across given CpG loci. Of the total 7970 CpG-SNP pairs, 1820 CpG loci were identified to be influenced by the variable genotype resulting into differential methylation patterns in the predictive training model. These loci are called as methylation quantitative trait loci (meQTLs) and have influence the methylation pattern across the extended genomic regions. Out of the total 1820 meQTLs in the training model, 489 polymorphic alleles were identified to be significantly associated with differential methylation in the test data set ($P < 0.05$). However, only 392 and 243 SNPs were detected to be significantly associated with differential methylation pattern at astringency of 0.01 and 0.001, respectively. The majority of these meQTLs were mapped to the intronic regions (50-60%) though a limited number were associated with synonymous (1.2-1.7%) or non-synonymous coding SNPs (3-4%). Some of these SNPs being associated with one or more CpG loci suggest that they not only influence the methylation status to the associated CpG loci but also affect the surroundings at very close distance.

Genome-wide localization of meQTLs identified in test model ($p < 0.05$) and their loci on the respective chromosome have been depicted by Manhattan plot [Figure 3.3].

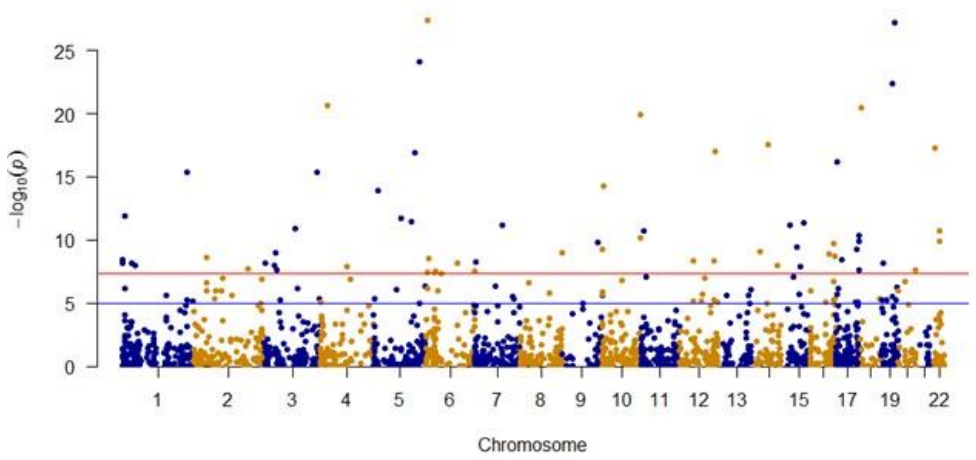

Figure 3.3 Dots within the Manhattan plot displays the identification of significant SNPs in the vicinity of CpG site leading to the meQTLs in the test model; the x-axis represents genomic position of SNPs, while the y-axis represents the –log p-value of the association between the SNPs and CpG site. The red and the blue line indicates the threshold $-\log_{10}(1 \times 10^{-4})$ and $-\log_{10}(0.2)$, respectively, for genome-wide statistical significance.

In particular, the association of breast risk alleles, rs1570056 and rs11154883 with DNA methylation levels (cg18287222) of *MAP3K5* gene ($p < 0.001$), is an interesting case because the gene encodes for mitogen-activated protein kinase protein that activates signalling cascade. The downstream protein kinases that are activated include MAPK or extracellular signal-regulated kinase (ERK), MAPK kinase (MKK or MEK), and MAPK kinase (MAPKKK). These kinases are highly conserved, and the homologs exist in yeast, Drosophila, and mammalian cells [405]. While, the differential distribution of major (T) and minor allele (C) (SNP: rs1570056) regulates the DNA methylation of the CpG site cg18287222 [Figure 3.4 a], the mutation in the allele G -> A associated with SNP rs1154883, simultaneously regulates the same CpG loci. These alleles influenced DNA methylation at a p-value of 5.8×10^{-5} (< 0.001) and 0.0002, respectively [Figure 3.4 a, b]. Thus, it presents an interesting fact that the alleles of the respective SNP act in a differential manner in regulating DNA methylation. Finally, we examined the overlap in regulatory variation affecting both methylation and gene expression based on RNAseq data.

The differentially methylated CpG site was identified to be negatively associated ($r = -0.53$) with the expression of *MAP3K5* gene at p-value < 0.01 [Figure 3.5]. We also tested the association of these SNPs with the expression level of the gene. The variable

allele associated with each SNP regulated the quantitative expression of *MAP3K5* gene at *p-value* of 0.028 and 0.012 for rs1570056 and rs11154883 SNP, respectively [Figure 3.6 a, b]. The polymorphism associated with differential mRNA expression level is referred to as expression quantitative trait loci (eQTLs). In summary, our result clearly demonstrates that the genetic variants (SNPs) significantly overlay with both meQTLs and eQTLs.

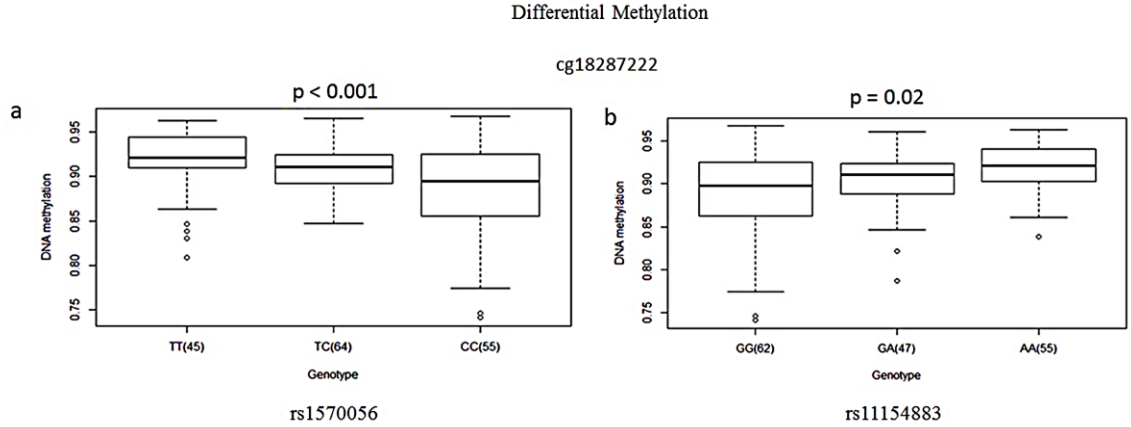

Figure 3.4 Breast cancer risk SNP rs1570056 and rs11154883 is associated with differential CpG Methylation. Cis-association between the SNPs (a) rs1570056 (b) rs11154883 regulates the methylation of CpG site cg18287222. These SNPs have loci as an intron-variant of *MAP3K5* gene. The box plots show the distribution of the methylation levels with respect to each genotype category with error bars representing the 25 and 75% quantiles.

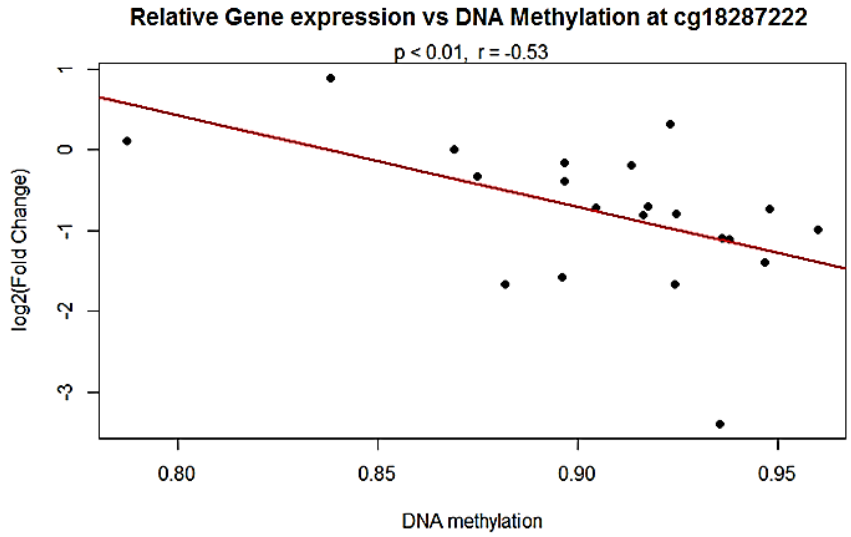

Figure 3.5 Spearman Correlation with respect to fold change in gene expression and DNA methylation in breast cancer. (a) DNA methylation residuals at loci cg18287222 is negatively associated (*r = -0.53*) with *MAP3K5* expression in breast cancer patients at *p-value < 0.01*. The

regression line (red line) depicts the linear association between DNA methylation residuals and gene expression residuals.

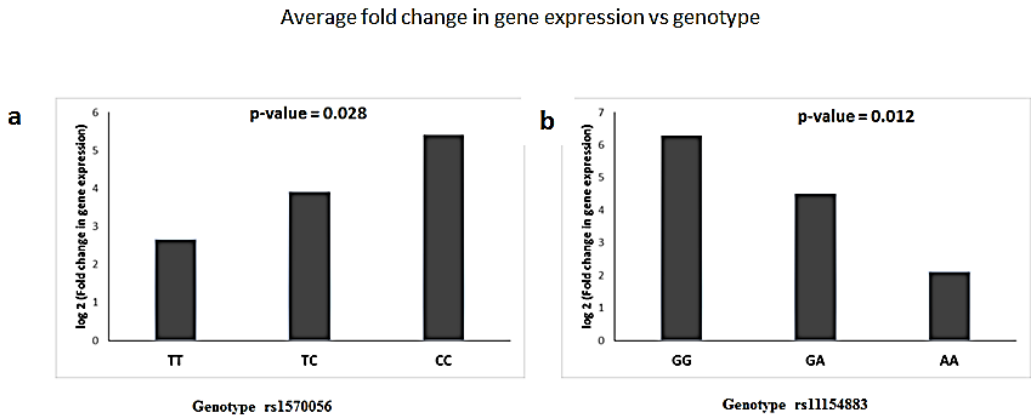

Figure 3.6 Fold changes in gene expression with respect to variable genotype associated with SNPs was identified to be significant at *p-value* of 0.028 and 0.012, respectively. (a) Fold change in gene expression was evaluated in the presence of SNP rs1570056. Homozygous dominant allele "TT" causes comparatively more downregulation in gene expression in comparison to heterozygous (TC) and homozygous recessive (CC) allele. (b) Fold change in gene expression is evaluated in the presence of SNP rs11154883. Homozygous recessive allele AA causes results in more downregulation in comparison to heterozygous (GA) and homozygous dominant (GG) allele.

3.3.2 Prognostic potential of differentially methylated CpG sites on survival of breast cancer patients

Breast cancer has displayed an increasing incidence and more importantly, the steady mortality rate in a past decade. While the clinical screening has attributed to the enhanced survival of breast cancer patient, still improvised markers are required to assess accurately patient prognosis at the time of diagnosis. The disease heterogeneity, limited specificity and the clinicopathological variables are being used in prognostication and staging of breast cancer. Thus, the development of complementary biomarkers with more specific prognostic potential will allow assessing the risk of developing recurrent and/or metastatic disease. We report for the first time the association between the differentially methylated CpG site and overall survival of breast cancer patients. Univariate and multivariate Cox PH regression analysis have been implemented to establish the prognostic potential of differentially methylated CpG sites. Of the total 1820 meQTLs, 489 differentially methylated CpGs were identified to be significantly associated with the survival of breast cancer patients in the training model. The prognostic potential of these differentially methylated CpGs were validated in test model of 164 patients.

To test the association of risk in 164 breast cancer patients for overall survival, we first began our analysis using univariate Cox PH model. On evaluating 489 differentially methylated CpGs (training model) based upon the clinicopathological variables of vital status and last follow-up days, 18 covariates were found to be significantly associated with overall survival of breast cancer patients in the test model. The most significant association with overall survival were observed for cg04003327 on chromosome 2q37.3 (HR= 0.01, p = 0.003), cg14033170 on chromosome 7p15.1 near *CREB5* gene (HR = 158.94, p = 0.004) and cg00902464 on chromosome 1p21.2 (HR = 0.02, p = 0.016) [Table 3.1]. The risk allele associated with CpG sites cg11340537, cg00956490, cg04586622, and cg14033170 have already been identified in GWAS phenotypes. The genotypic variation associated with SNP rs2640785 has been identified to regulate the differential methylation of CpG site cg11340537 located in the exonic region of the *EXPH5* gene. The missense variation (GAG -> GTG) associated with this risk allele is of greater significance as it is conjointly associated with differential methylation, gene expression and survival of breast cancer patient. A similar explanation can be associated with synonymous risk variant rs940453 (ATA -> ATC) that regulates methylation of CpG site cg00956490 and simultaneously influences *ZNF775* gene expression and overall survival. However, the risk allele rs2384061 is an intron variant that is associated with CpG site cg0458662 and regulates the expression of *ADCY3* gene. The SNP rs2230576 mapped to the 3'-UTR variant is correlated with differential methylation of CpG site cg05370838 and gene expression of *ADMA8* gene. The differentially methylated CpG site holds significance in regulating the overall survival of breast cancer patients (HR= 0.008, p = 0.049).

Table 3.1. Univariate analysis of differentially methylated CpGs sites, and theit associations with the overall survival in test model: HR: Hazard Ratio; CI: Confidence Interval for the hazard ratio

CpG ID	SNP ID	GENE	HR	95% of CI	P-value
cg04003327	rs1054641	ESPNL; SCLY	0.011948	0.00076 – 0.18	0.0032
cg14033170	rs177595	CREB5	158.9545	3.10816 – 8129.1	0.0038
cg00902464	rs17403618	LOC100128787	0.023795	0.00178 - 0.32	0.0167
cg03383184	rs6988652	Intergenic	52.99806	1.4918 - 1882.7	0.0170
cg00101629	rs6660333	KIAA1026	0.050101	0.00378 - 0.667	0.0173
cg03521812	rs4620521	Intergenic	0.023186	0.00107 - 0.498	0.0177
cg17378966	rs2431663	DUSP1	13.45278	1.2033 - 150.39	0.0262
cg08937612	rs12409375	VSIG8	0.003215	2.63E-05 - 0.39	0.0270
cg26901096	rs17444979	LOC254312	13.55016	1.32054 - 139.1	0.0292

cg13558682	rs9424283	LRRC47	0.024577	0.001227 - 0.49	0.0366
cg16774160	rs3088007	HSPA12B	0.000191	2.23E-07 – 0.16	0.0384
cg06099459	rs10505956	C12orf77	0.002645	1.64E-05 – 0.426	0.0416
cg05370838	rs2230576	ADAM8	0.008903	0.000201 – 0.395	0.0498
cg11340537	rs2640785	EXPH5	0.031486	0.00135 - 0.733	0.0528
cg00956490	rs940453	ZNF775	0.001156	3.58E-06 – 0.373	0.0645
cg04586622	rs2384061	ADCY3	0.008966	0.000116 – 0.693	0.0648
cg00889709	rs16923085	FAM110B	0.061646	0.00400 - 0.948	0.0652
cg14798310	rs738806	SLC2A11 ,MIF	0.000387	2.06E-07 - 0.725	0.0793

The univariate analysis was followed by the multivariate regression model to assess the risk associated with 18 co-variables obtained from the univariate study. This logistic regression analysis led to the identification 8 differentially methylated CpGs having a significant association with overall survival of the breast cancer patient [Table 3.2]. Among these, the most substantial findings were observed for cg04003327 (HR= 0.016; 95% of CI = 0.0003-0.86; P = 0.04), cg11340537 (HR = 0.28; 95% of CI = 0.005-14.49; P = 0.05) and cg00956490 (HR = 0.0005; 95% of CI = 1.36 x 10^{-7}-2.44; P = 0.08). These 8 covariates showed the clear demarcation of the patient into high (84 patients) and low risk (84 patients), respectively at a significant *p-value* of 0.04 [Figure 3.7]. Beside these differentially methylated CpGs, the exclusive effects of SNPs were also evaluated for both direct and indirect effect on overall survival of breast cancer patients. In the next section of our study, we explain the variable allele distribution and its association with survival of breast cancer patients.

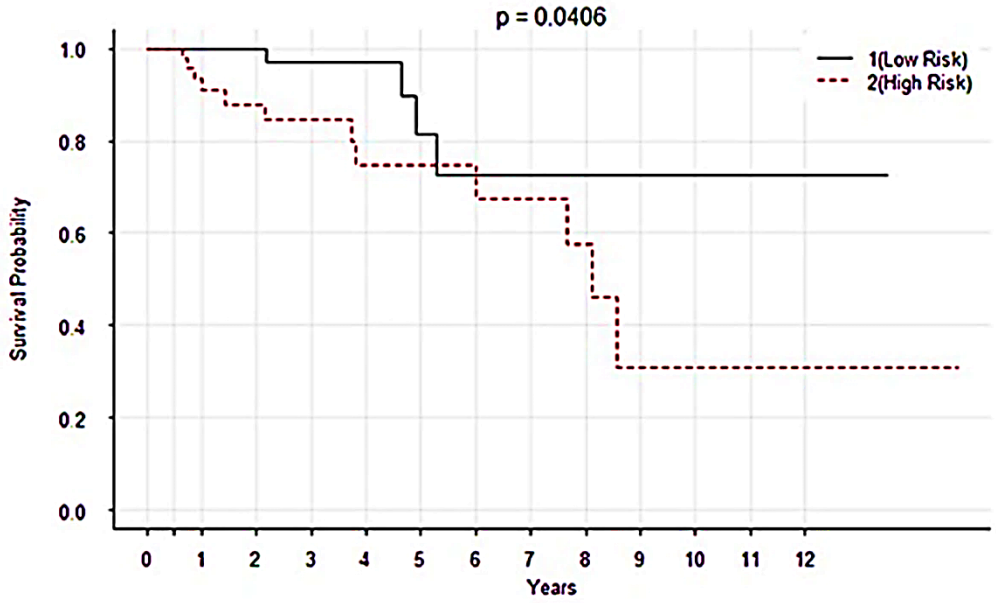

Figure 3.7: Kaplan-Meir plot associated with differnentially methylated CpGs sites. These differentially methylated sites could successfully classify 164 tumor patients (test set) into high (84 patients) and low risk (84 patients), respectively, at *p-value = 0.041*.

Table 3.2. Summary for univariate and multivariate analysis of differentially methylated CpGs and the associations with overall risk based upon Cox proportional hazard model in test dataset. HR: Hazard Ratio; CI: Confidence interval for the hazard ratio.

SNP ID	CpG ID	Gene	Locus	Univariate			Multivariate		
				HR	95% of CI	p	HR	95% of CI	p
rs1054641	cg04003327	ESPNL; SCLY	2q37.3	0.012	(0.001-0.18)	0.003	0.016	(0.0003-0.86)	0.04
rs2640785	cg11340537	EXPH5	11q22.3	0.031	(0.001-0.73)	0.05	0.28	(0.005-14.49)	0.05
rs940453	cg00956490	ZNF775	7q36.1	0.001	(3.58E-06-0.37)	0.06	0.0005	(1.36E-07-2.44)	0.08
rs2230576	cg05370838	ADAM8	10q26.3	0.0008	(0.002- 0.39)	0 .049	0.028	(0.0001-4.5)	0.17
rs6660333	cg00101629	KIAA1026	1p36.21	0.05	(0.003-0.66)	0.17	0.88	(0.02-37.57)	0.95
rs177595	cg14033170	CREB5	7p15.1	158.94	(3.1-8129.07)	0.003	213	(1.7-25740)	0.028
rs4620521	cg03521812	Intergenic	1q31.2	0.02	(0.001-0.49)	0.018	0.04	(0.001-1.8)	0.098
rs9424283	cg13558682	LRRC47	1p36.32	0.024	(0.001-0.49)	0.036	0.336	(0.001- 101.1)	0.71

3.3.3 Probing the association SNPs on the survival of breast cancer patients

Genetic variation characterized by single nucleotide polymorphism offers promising surrogate biomarker to predict therapeutic response and prognosis in breast cancer patients. In the present study, we investigated the risk associated with the individual SNP and in cumulative fashion on the overall survival. We developed a probabilistic framework for predicting and prioritizing the candidate SNPs in the training data set and validated across test set constituting 164 samples. The complete survival analysis was based upon the homozygous dominant and recessive allele and heterozygous allele distribution available for each SNP.

The univariate survival analysis associated with individual SNP was based upon the log-rank test at threshold *p-value* of 0.05. Of the total 7970 CpG-SNP pair, 492 SNPs were significantly associated with the overall survival in the training set of breast cancer patient. Each individual SNP were validated in the test model. Of the total significant SNPs in the training set, 23 were substantially associated with survival and their respective *p-value* ranged from $\leq 0.0001 - \leq 0.05$ [Table 3.3]. These SNPs had a variable distribution across the genome. Of the total significant SNPs in the test set, 7 SNPs (rs2880556, rs17006586, rs876701, rs41470747, rs2967798,rs11804125, rs1548373) were present as an intro variant, 6 SNPs (rs12085531,rs12653167, rs12591432, rs940482, rs1532272) were present in the intergenic region, 3 SNPs (rs16943263, rs9325443,1538146) were localised in the upstream region, each of 2 SNPs were associated with non-coding transcript variant (rs7117026, rs10101376) and synonymous variant (rs17142291, rs140679) and remaining one SNP (rs1862372) was associated with 5'UTR variant. Moreover, the SNPs highlighted in the table are already mentioned in GWAS study in relevance to cancer and other diseases.

The Kaplan-Meir plot for the significant SNPs having nearly equal genotypic frequency is displayed in Figure 3.8. While the presence of heterozygous allele “GA” associated with SNP rs10101376 is detrimental, the homozygous dominant allele “CC” and “TT” concomitant with SNP rs140679 and rs1538146 affects the survival of the breast cancer patient at threshold *p-value* of 0.05. The homozygous dominant allele “TT” (rs1538146) is located in the upstream of the TRPC4 gene. The transient receptor potential cation channel (TRCP4) gene encodes a member of a canonical subfamily of transient receptor potential cation channels. This encoded protein forms a non-selective calcium-permeable cation channel that is activated by a Gq-coupled receptor and tyrosine kinase. The polymorphism associated with TRCP4 gene is deleterious, as it is conjointly linked with gene expression and regulates the overall survival.Similarly, the allele CC

associated with SNP rs1538146 regulates the expression of gamma-aminobutyric acid (GABA)-A receptor gene and is detrimental to breast cancer patients.

Table 3.3. Summary of SNPs associated with overall survival of breast cancer patients using log-rank in test dataset. AA: Reference allele, AB: Heterozygous allele, BB: Alternate allele.

CpG_ID	SNP_ID	P-value	GENE	A	B	AA	AB	BB
cg11929693	rs2880556	2.29E-24	LOC340073	G	T	153	9	2
cg09939673	rs7117026	2.55E-12	DQ592890	A	T	1	10	153
cg00067528	rs17006586	1.47E-05	ATP6V1B1	C	T	140	21	3
cg01711124	**rs12085531**	9.05E-05	Intergenic	C	T	4	24	136
cg09573435	**rs1862372**	0.000594	SEMA6A	C	T	111	43	10
cg22675791	**rs876701**	0.000627	DGKZ	A	G	6	36	122
cg20705812	**rs2286218**	0.001795	DLGAP2	A	G	143	16	5
cg08980697	**rs41470747**	0.006462	RASGEF1B	C	A	1	12	151
cg14584565	**rs16943263**	0.006649	LOC283761	G	C	152	8	4
cg04513214	rs12653167	0.008100	Intergenic	T	G	162	1	1
cg22422090	**rs2967798**	0.008121	KLHL3	T	A	102	44	18
cg24310780	**rs11804125**	0.008351	LMX1A	G	T	122	30	12
cg03339247	**rs1548373**	0.013806	ZFHX3	C	T	106	38	20
cg25203310	**rs10101376**	0.014656	LOC286083	G	A	59	47	58
cg20214734	rs17142291	0.016161	ASB13	G	A	4	9	151
cg15179472	**rs12591432**	0.018465	Intergenic	C	T	123	33	8
cg15461663	**rs940482**	0.029081	Intergenic	C	T	99	53	12
cg22514112	**rs1532272**	0.031189	Intergenic	A	G	94	52	18
cg04966682	**rs140679**	0.033337	GABRG3	C	T	57	67	40
cg02576753	rs140679	0.033336	GABRG3	C	T	57	67	40
cg20896197	**rs9325443**	0.037904	KIF20B	A	C	91	59	14
cg24540569	**rs574095**	0.041499	Intergenic	A	G	3	26	135
cg15398976	**rs1538146**	0.049488	TRPC4	G	T	65	54	45

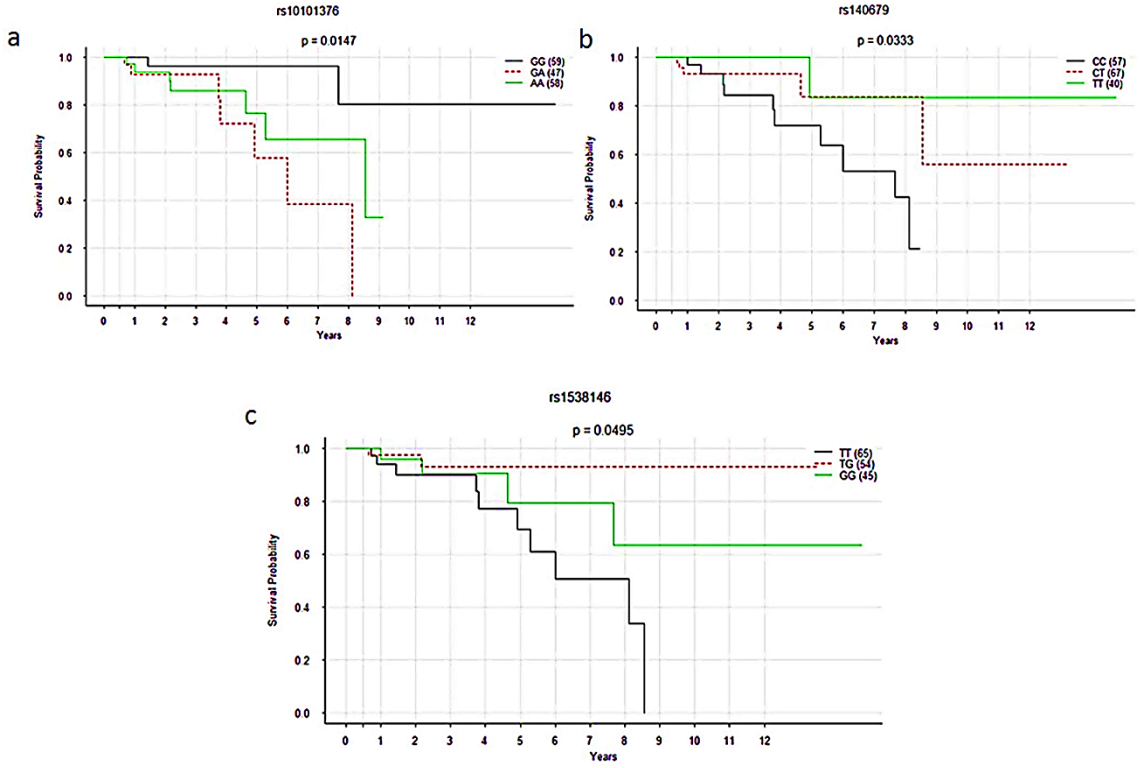

Figure 3.8 Kaplan-Meier survival plot for SNPs: (a) rs10101376 (b) rs140679 and (c) rs1538146. The survival analysis has been done such that the solid black line represents homozygous dominant, red dotted line: heterozygous allele and solid green line for homozygous recessive genotype. The findings are based upon test-dataset at threshold *p-value* of 0.05.

Beside the log-rank test, these 23 significant SNPs were also subjected to univariate Cox PH regression analysis. The most significant association in the univariate model for survival was observed for rs7117026 located on chromosome 11p11.2 (HR= 0.109, $p < 0.001$) as non-coding transcript variant of DQ582890 gene, rs1548373 at chromosome 16q22.3 (HR = 2.35 and p = 0.0096) as an intron variant of ZFHX3 gene, rs140679 on chromosome 15q12 (HR= 0.359, p = 0.016) as non-synonymous variant of GABRG3 gene, rs876701 on chromosome 11p11.2 (HR= 0.371, p = 0.038) as a intron variant of DGKZ gene and rs41470747 at chromosome 4q21.21 (HR = 0.357, p = 0.039) as an intron variant of RASGEF1B gene. Additionally borderline associated risk variants included rs574095, rs12653167, rs2286218 and rs1538145 at threshold of p = 0.1. Besides SNPs rs16943263 associated with CpG loci cg14584565 (HR = 2.44 and p = 0.17) is also identified in classifying the patients in high and low risk [Table 3.4].

Finally, we performed conjoint analysis by including 23 SNPs in order to assess the cumulative effect of the genetic variant on overall survival. We performed the multivariate Cox PH regression analysis between the SNPs and clinical variance constituting vital status and last follow-up days. Of the total 23 variables from the log-

rank test study, 16 SNPs were identified to be significant (test model) in grouping the patent into high and low risk based upon the multivariate model at threshold p-value of 0.05 [Figure 3.9a]. However, top 9 SNPs presents clear demarcation of patients into high and low risk at a *p-value* of 0.005 [Figure 3.9b]. The delineation was such that 84 patients (Test sample) survived for a longer duration while the remaining 84 were prone to poor prognosis and had survival probability for only $8^{1/2}$ years. Most of these genetic variants are germline and have shown significant association with overall survival. Thus, the Cox proportional model conjointly with clinicopathological features suggests the association between the genetic variants and the risk in the survival of breast cancer patients which may also modulate the cancer prognosis.

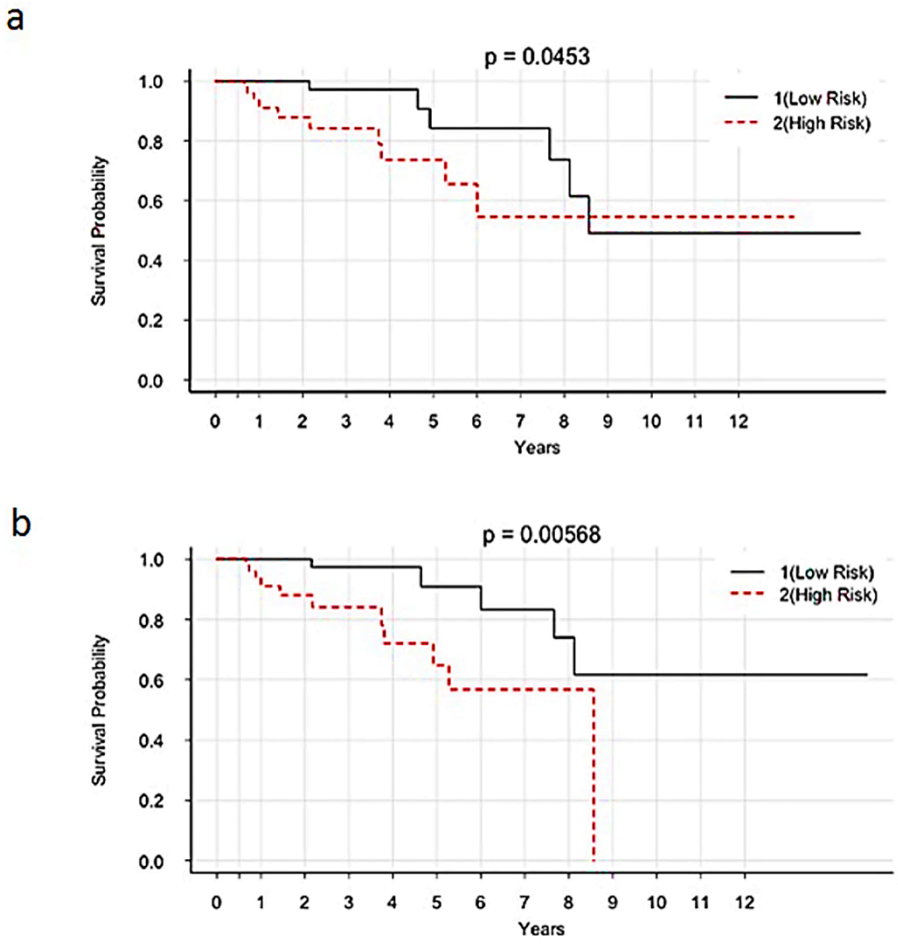

Figure 3.9 Kaplan-Meir curve associated with (a) top 16 SNPs, (b) top 9 SNPs (listed table 3.4) could classify 164 tumor patients into high (84) and low risk (84 patients) at threshold of $p < 0.05$ in the test dataset.

Table 3.4. Summary of univariate and multivariate analysis of SNPs associations with overall risk based upon Cox proportional hazard model in test dataset. HR: Hazard Ratio; CI: Confidence interval for the hazard ratio.

SNP ID	CpG ID	Gene	Locus	Univariate			Multivariate		
				HR	95% of CI	p	HR	95% of CI	p
rs1862372	cg09573435	SEMA6A	5q23.1	1.66	(0.66-4.15)	0.28	1.15	(0.2-6.3)	0.87
rs2880556	cg11929693	LOC340073	5q31.3	2.23	(0.5-9.8)	0.29	59.7	(2.52-14.12)	0.011
rs1548373	cg03339247	ZFHX3	16q22.3	2.35	(1.23-4.17)	0.0096	5.99	(1.84-19.49)	0.0029
rs12591432	cg15179472	Intergenic	15q23	1.32	(0.48-3.6)	0.59	4.50	(0.67-30.21)	0.12
rs12653167	cg04513214	Intergenic	5p15.1	2.96	(0.95-9.24)	0.062	4.85	(0.45-52.28)	0.19
rs16943263	cg14584565	LOC283761	15q26.1	2.44	(0.68-8.6)	0.17	0.06	(0.0007-6.6)	0.25
rs12085531	cg01711124	Intergenic	1p36.12	0.52	(0.20-1.3)	0.17	1.4	(0.44-4.87)	0.53
rs1538145	cg15398976	TRPC4	13q13.3	0.558	(0.29-1.07)	0.081	0.58	(0.21-1.54)	0.28
rs41470747	cg08980697	RASGEF1B	4q21.21	0.35	(0.13-0.94)	0.039	0.38	(0.047-3.07)	0.37
rs140679	cg04966682	GABRG3	15q12	0.359	(0.15-0.82)	0.016	0.11	(0.018-0.7)	0.019
rs17142291	cg20214734	ASB13	10p15.1	0.56	(0.15-2.0)	0.38	11.80	(0.47-291.84)	0.13
rs11804125	cg24310780	LMX1A	1q23.3	1.12	(0.56-2.2)	0.75	1.38	(0.34-5.47)	0.64
rs7117026	cg09939673	DQ5982890	11p11.2	0.109	(0.03-0.3)	3 .00 X 10^{-4}	0.00198	(0.00003-0.12)	0.0034
rs876701	cg22675791	DGKZ	11p11.2	0.371	(0.14-0.94)	0.038	0.29	(0.07-1.27)	0.1

rs574095	cg24540569	Intergenic	1p31.3	0.445 (0.19-1.0) 0.058	0.25 (0.036-1.70) 0.16
rs2286218	cg20705812	DLGAP2	8p23.3	2.76 (0.88-8.5) 0.08	0.97 (0.05- 19.20) 0.99

3.4 Discussions

Molecular understanding of inter-tumor heterogeneity is key to effective cancer treatment and personalized medicine. Analysis of high-throughput molecular profiling data has revealed the extent of inter-tumour heterogeneity in breast cancer. The identification of diverse levels (sub-types) of tumor heterogeneity and the most-appropriate treatment strategies for each sub-type is expected to radically improve the treatment practices for the optimal clinical management of breast cancers [406].

Genome-wide association, studies have led to the identification of a large number of genetic variants that confer susceptibility to different types of cancers. However, the risk conferred by individual variant is not sufficient to uphold the individual risk prediction. Assessing the genetic variability by incorporating multiple SNPs into a predictive model could achieve improved risk discrimination that may be useful for prognostic stratification of breast cancer patients [407, 408]. It is often a challenge to assess the functional impact of non-coding genetic variants, for example, the effect of SNPs transcriptional activity, and the associated disease risk. It is more likely that such variation may indirectly influence the epigenetic regulation located in a nearby position (*cis*) or distant loci (*trans*).

Here, we have investigated the relationship between genetic variation, DNA methylation and gene expression, and their potential utility as prognostic biomarkers of breast cancer. Numerous studies have discovered the association of genetic variants with variation in gene expression [409-411]. To our knowledge, this is the first study where we have investigated the relationship between methylation quantitative trait loci (meQTLs) and single nucleotide polymorphism (SNPs), and their combined effect on breast cancer prognosis. Questions still remain for the prognostic biomarkers identified for cancer. The first question is that there is little overlap among numerous prognostic signatures generated from different studies. Another question is that most signatures generated do not have clear biological meanings as why these prognostic genes may affect patient outcome, which lead to the clinical application of such signatures still under debate. In this study, we developed a novel method to identify prognostic gene signatures for breast cancer by integrating genomic and epigenomic data. This is based on the hypothesis that multiple sources of evidence pointing to the same gene or pathway are likely to lead to reduced false positives. We also apply random resampling to reduce overfiiting noise by dividing samples into training and test datasets. In the current analysis, TCGA breast invasive carcinoma (BRCA) overlapping dataset between DNA methylation, Affymetrix SNP array, and clinical samples were randomly divided into two subsets based on thevital status obtained from clinical data. The Caret module was implemented in the random classification of $3/4^{th}$ (486) of sample into training and $1/4^{th}$ (164) into test subset at R-

interface. The predictive model was trained based on certain features mainly the beta values and genotype associated with methylation and SNP, respectively. The robustness of the features were evaluated statistically in the training subset and were validated in an exclusive and independent test subset. The significant association between methylation and genotype was calculated based on one-way ANOVA at threshold *p-value* of 0.05. Each SNP encoded for variable homozygous and heterozygous genotypic (allele) frequency across the breast cancer samples. Localization of each SNP was interrogated at 50-bps upstream and downstream of each CpG site. Thus, for a window size of 50-bps we investigated CpG-SNP pairs to enlist their statistical significance such that minimum of one SNP is associated with one CpG loci. This evidence of a correlation between genetic variant at specific loci and DNA methylation led to the identification of meQTLs. Of the total distribution of 7970 CpG-SNP pair in the window size of 50-bps, 1820 SNPs were significantly associated with differential methylation in the predictive training model. Out of these 1820 CpG-SNP pairs, 489 SNP were significantly correlated with differential methylation leading to the identification of meQTLs in the test set. These CpG-SNP pairs enlighten on the plausible mechanism through which SNPs have an influence on the phenotype. In one of the scenario, presence of SNP in the vicinity of CpG loci prevents the binding of CpG methyl binding proteins as a consequence of which affects DNA methylation [412]. In another scenario, these SNPs may affect the transcriptional silencing via differential DNA methylation. Indeed, it has also been reported that DNA methylation plays a significant role in the regulation of splicing and aids in distinguishing exons from introns [413, 414]. Thus, genetic variant characterized by the presence of SNPs in the intronic region, causes differential methylation and leads to a different, set of spliceosome [415]. Interestingly, we have identified CpG loci (cg18287222) that constitute two SNPs (rs1570056, rs11154883) located in the intronic region and affects the function of *MAP3K5* gene. These genotypic variation associated with these SNPs regulates the methylation pattern in contrary manner. While the homozygous dominant allele TT with respect to SNP rs1570056 is responsible for hypermethylation, the homozygous recessive allele AA associated with rs11154883 causes increase in methylation level for the same CpG loci. These differential distribution landmarks the presence of specific meQTL.Beside overlap with meQTL, these SNPs also leads to eQTLs in the cis-regulatory region. Thus, the meQTLs have been identified to be enriched in eQTLs. Moreover, *MAP3K5* (Mitogen-activated protein kinase (MAPK) is an essential component of MAP kinase signal transduction pathway and plays a crucial role in the apoptosis [416, 417]. Characterizing the genetic control of methylation and its association with the regulation of *MAP3K5* gene expression presents signature marks that can resolve in understanding the underlying biology behind the complex phenotype in breast cancer.

The differentially methylated CpG sites obtained from above study was further evaluated for their association with overall survival of breast cancer patients. The high mortality rate associated with metastasis in breast cancer urge for the development of more personalized prognostic algorithms that will complement the general, clinical predictors. We have systematically investigated the risk associated with host-related breast invasive carcinoma traits that may serve as a biomarker for disease prognosis. In this study, we have implemented model selection framework composed of linear statistical techniques of univariate analysis based on log-rank test and multivariate Cox proportional regression model. Of the total 1820 significant CpG-SNP pair, we identified a comprehensive panel of 489 differentially methylated CpGs to be associated with overall survival in the training set based upon the the univariate regression model. However, 18 differentially methylated CpGs were identified as the landmark risk loci for overall survival in the test set. The conjoint multivariate regression analysis of these differentially methylated CpG sites led to the identification of 8 differentially methylated CpGs as promising candidates having significant prognostic potential. These noteworthy biomarkers clearly demarcated 164 breast cancer patients of the test sample into high and low risk, respectively. The most interesting fact is that the SNPs (rs2640785, rs940453, and rs9424283) associated with the differentially methylated CpG sites (cg11340537, cg00956490, and cg04586622) have been already reported in GWAS phenotypes. We explored the potential mechanism by which differentially methylation CpG site cg11340537 directs overall survival in breast cancer patients. The missense variant (GAG -> GTG) associated with SNP rs2640785 dictates differential methylation of CpG site cg11340537 and mRNA expression of EXPH5 (Exophilin 5) gene. *EXPH5* gene shares homology with Rab-GTPase and plays a significant role in vesicle trafficking [418, 419]. The active participation of this gene has been reported in colorectal cancer [420]. The differential methylation associated with the CpG site cg14033170 also holds greater significance. SNP rs177595, an intron variant located in the vicinity of CpG site cg14033170 regulates the differential methylation and subsequently deregulates *CREB5* gene expression. *CREB5* gene encodes for cAMP responsive element binding protein 5. Previous studies have suggested that *CREB5* gene play a fundamental role in a metastatic signal network in colorectal cancer [421]. Moreover, it has been reported that eQTL associated with *CREB5* gene causes colorectal, prostate and nasopharyngeal cancer [422-424]. On a similar account, differentially methylation associated with CpG cg00956490 holds prognostic significance. The risk variant rs940453 linked to CpG loci regulates the mRNA expression of *ZNF775* gene. The gene encodes for zinc finger protein 775 [425]. It has been identified to be involved in transcriptional regulation. SNP rs2230576 is a 3'-UTR variant that has been mapped to the vicinity of differential methylated CpG site cg05370838 and ADAM metallopeptidase domain 8 (*ADAM8*) gene. The differentially

methylated CpG site is associated with high risk in breast cancer patients. *ADMA8* gene localised in the vicinity of the CpG site encodes for membrane-anchored protein that have been implicated in several biological process including cell-cell interactions, cell-matrix interactions and neurogenesis [426]. It has been reported that *ADMA8* is aberrantly expressed in breast tumours, especially in triple-negative breast cancer (TNBCs). The aberrant expression of *ADAM8* gene has been correlated with poor prognosis in breast cancer patients and concomitantly with increased number of circulating tumour cells and metastasis [427]. The anomalous expression of the *ADAM8* gene is also associated with poor survival in colorectal, lung, gastric, pancreatic cancer, hepatocellular, gastrointestinal carcinoma and gliomas [428-431].

Studies have been done so far correlate the conjoint effect of significant CpG-SNP pair regulating the differential methylation and overall survival of breast cancer patients. Recent studies have illustrated the upshot of genetic variants in regulating the overall risk associated with breast cancer patients. However, the cumulative effect is still to be disclosed. In the next section, we detailed about the prognostic potential of individual SNPs and their cumulative action. In our study, we have comprehensively analyzed the TCGA SNP array data mapped to methylated loci and concomitantly evaluated its association with the breast cancer survival. Of the total 7970 CpG-SNP pair, 492 SNPs were predicted to be significantly associated with overall survival in the training set. However, the univariate analysis based upon log-rank test mapped 23 SNPs to be significant across the test data set. Most of these SNPs have been highlighted in GWAS-studies. In this study, we have mainly displayed Kaplan-Meir plot for the SNP having higher and nearly equal allelic distribution in breast cancer population. The heterozygous allele "GA" associated with SNP rs10101376 is detrimental and is related to poor prognosis. Similarly, the homozygous dominant allele "TT" linked to rs140679 SNP disrupts the mRNA expression of gamma-aminobutyric acid A receptor (GABRG3) [432], subsequently deteriorates survival probability in breast cancer patients. The homozygous genetic variant "TT" of SNP rs1538146 mapped to 1349 bps upstream of TRPC4 gene (transient receptor potential cation channel, subfamily C) reduces the overall survival and has a significant prognostic determinant. The canonical transient receptor potential (TRPC) channels are permeable to Ca^{2+} cationic channels and regulate Ca2+ influx in response to G protein-coupled receptor [433]. Overexpression of TRPC4 gene resulting in anomalous cell proliferation have been reported in the prostate, ovarian, lung cancer and renal cell carcinoma [434-437]. Our findings have demonstrated the potential importance of assessing prognosis in breast cancer based upon the univariate model of SNP distribution. Finally, we assembled these SNPs to construct logistic regression model and evaluated their cumulative effect on overall survival of breast cancer. Of the total 23 SNPs, 18 SNPs had significant prognostic potential and could classify 164 breast cancer

patient into poor prognostic (high risk) and higher prognostic group (low risk). However, the conjoint effect of 9 SNPs holds more clear vision on demarcation.

In summary, the comprehensive assessment of CpG-SNP pairs has led to the identification of loci that holds the risk to the overall survival of breast cancer patients. The novel findings are highly promising and strongly support the identification of these loci in the clinical visualization of breast cancer progression. Such prognostic scans at the genome-wide level will likely be beneficial not only for identification of novel prognostic biomarkers, but also will open a new horizon to the novel pathways involved in breast cancer progression, directing to the potential targets for more efficient treatment strategies.

Chapter 4

To identify novel inhibitor(s) targeting DNA methyltransferases for therapeutic intervention of breast cancer

4.1 Introduction

The intra-tumor heterogeneity in breast cancer characterized by extended molecular diversity poses an important impediment [438]. Fuelled by the Darwinian evolutionary dynamic governance of the disease, the selective pressure of targeted therapeutics in breast cancer inevitably leads to the emergence of resistance in the tumor cell [439]. Interrogation of molecular mechanisms mediating high resistance rationally guides our choice to combinatorial therapeutics. The systemic efforts of interrogation based upon synergetic profiling of genetic and epigenetic aberration will provide guidance for rationally selecting therapeutic strategies [440]. In the studies done so far, we have already predicted the close association of genetic polymorphism and DNA methylation in diagnosis and prognosis of breast cancer. The presence of SNPs in the vicinity of CpG loci largely influences the distribution of methylation pattern. This differential methylation distribution is mainly associated with hypermethylation of tumor suppressor genes in breast cancer [129]. However, the reversal of this hypermethylation by small molecule or inhibitors may provide novel cancer therapeutic strategies [441].

Cellular DNA methylation is established and maintained by the complex interplay of a family of dedicated enzymes called DNA methyltransferases (DNMTs). Along with regional hypermethylation and overall hypomethylation of the genome in many cancers, it has been reported that the expression and activity of DNMTs are very high [442]. This gives a clue that DNMTs may have oncogenic potential apart from its DNA methylation activity thus, it has emerged as a budding anticancer drug development target. Targeting inhibitors to the catalytic domain of DNMTs is essential for therapeutic interventions [443]. Currently, the available inhibitors of DNMTs are classified into two broad groups, known as nucleoside and non-nucleoside analogs. The archetypal nucleoside inhibitors are derivatives of the cytidine nucleoside, and they inhibit DNMT activity only after getting incorporated into newly synthesized DNA strands and trapping the DNMTs by forming DNA-protein adduct [21, 443].

This chapter is based upon published research article [444].

Concomitantly, the cellular levels of DNMTs are rapidly depleted leading to DNA demethylation and continued DNA replication. The prototypical nucleoside inhibitor 5-azacytidine (Vidaza) is an FDA-approved drug largely used in the treatment of cancer [445]. 5-azacytidine being a ribose nucleoside is chemically modified to a deoxyribose sugar to get incorporated into DNA [446].

However, a portion of ribose sugar gets incorporated into RNA, affecting diverse RNA functions including ribosome biogenesis. 5-aza-2'-deoxycytidine (i.e., decitabine), deoxyribose analog of 5-azacytidine was identified as new potent inhibitor as it directly gets incorporated into DNA [447, 448]. It was found to be effective against myelodysplastic syndrome, acute myelogenous leukemia, and chronic myelogenous leukemia. However, the substantial toxic effect of these nucleoside inhibitors offers limitation to their usage in higher dose against the treatment of cancer [449].

Another class of inhibitor constituting the non-nucleoside group directly blocks DNA methyltransferase activity, and it does not possess the inherent toxic property as that of the nucleoside inhibitors. One of such DNMT inhibitor is (–)-epigallocatechin-3-gallate (EGCG), a tea polyphenol. However, the degradation of EGCG produces a substantial amount of hydrogen peroxide. H_2O_2 being strong oxidizing agent causes oxidation of DNA methyltransferases and other associated protein [450]. Other group of phytochemicals which belong to the category of non-nucleoside inhibitors in the treatment of cancer are: mahanine, a carbazole alkaloid [451]; and curcumin, a component of the Indian spice turmeric [452]. Many others are effective against non-cancerous diseases. Some of them are hydralazine, antihypertensive drug [453]; procaine, local anesthetic [454] and procainamide, an antiarrhythmic drug [455]. SAH, the end product of DNA methylation reaction, and its analogs apparently have also been reported as selective inhibitors towards inhibition of DNA methylation [456, 457]. However, these inhibitors have some limitations in lieu of their specificity in inhibiting DNMTs.

In this investigation, we report for the first time a detail view of the elementary interactions between non-nucleoside inhibitors and DNMTs in the active site terminal and thus, find out the best inhibitors. To explore the interactions between the enlisted non-nucleoside inhibitors and DNMTs, we performed docking and simulation studies [Figure 4.1]. Application of docking based on varying algorithms affirms the binding pattern and profile of a ligand. The relative binding site for all the compounds/inhibitors is chosen at the SAH-binding pocket of DNMT1 and DNMT3A/and of human and mouse. Molecular dynamics (MD) simulations analysis of the potential complexes from docking gives the final clue about the stability of the DNMT-inhibitor complexes. The dynamic picture of the complexes is determined using the time-dependent evolution of the system during the simulation.

Changes in free energies for binding (-ΔG) were determined and total energy was decomposed on the basis of per residue contribution. The non-covalent interactions constituting hydrogen bonding, van-der-Waals, and electrostatic occupancies were monitored throughout the docking and simulations. Moreover, the efficacy of the best-found inhibitor is tested by *in-vitro* studies in invasive breast cancer cell line, MDA-MB-231.

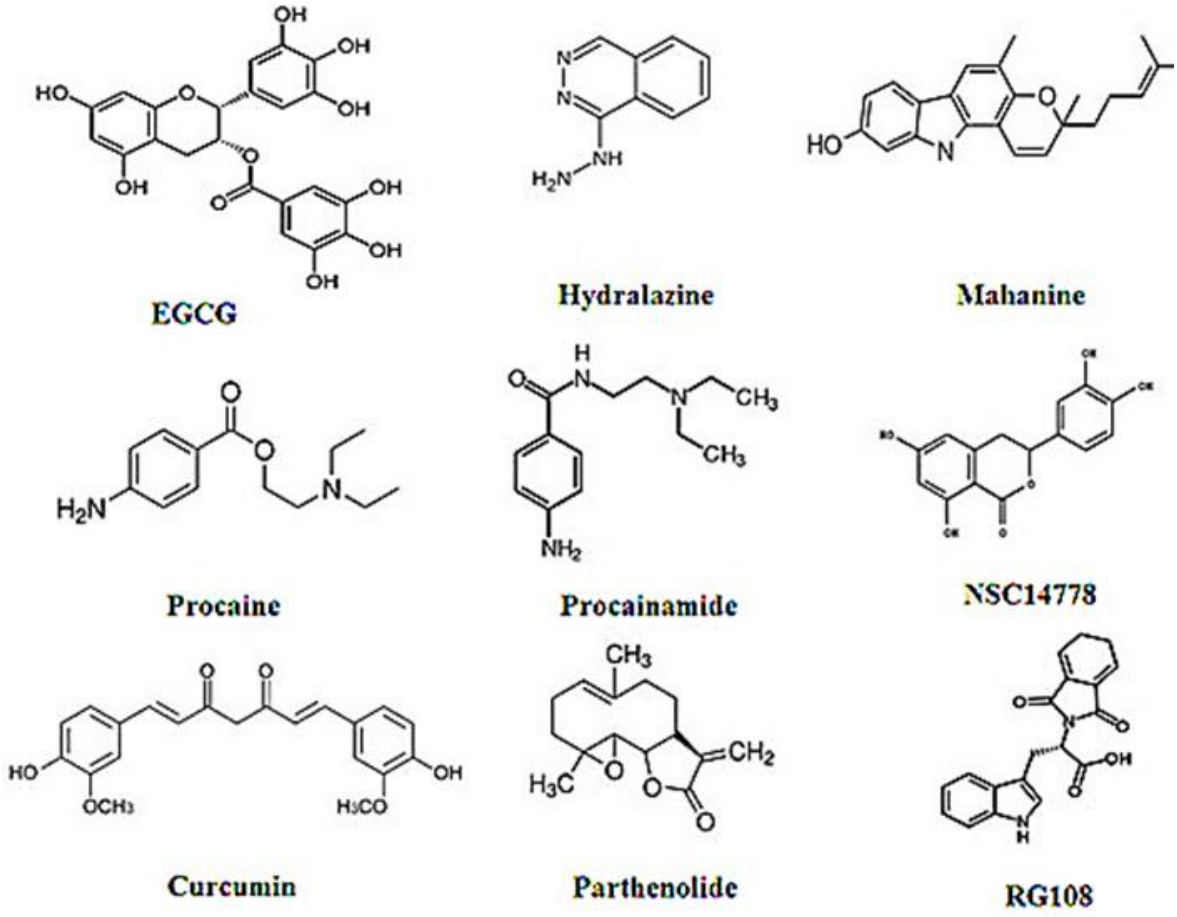

Figure 4.1 chemical structures of non-nucleoside inhibitors against DNMTs known till date

4.2 Materials and Methods

4.2.1 In-silico dataset preparation, molecular docking and simulation studies

4.2.1.1 Preparation of protein structure and ligand

The X-ray crystallographic structures of human DNMT1 (PDB id: 3PTA) [458] and DNMT3a (PDB id: 2QRV) [459] co-crystallized with SAH at resolutions of 3.6 Å and 2.89 Å, respectively, were retrieved from RCSB Protein Data Bank [460]. Subsequently the structure of mouse DNMT1 of 3.25Å (PDB: 3AV5) [461] co-crystallized with DNA and SAH was obtained from protein data bank. Mouse DNMT1 was used for the *in-silico*

study because it is used as an experimental model for the *in-vivo* study. The heteroatoms including SAH and zinc ion other than those present in the active site were edited using chimera software. The double-stranded DNA attached to human and mouse DNMT1 were also clipped off. The energy minimization of protein structures was done using steepest descent and conjugate gradient algorithm of 100 steps and step size of 0.02Å. It followed the addition of polar hydrogen and Gasteiger charges. The set of non-nucleoside inhibitors as described in Figure 4.1 was retrieved from PubChem (http://pubchem.ncbi.nlm.nih.gov/) and ChEBI database [462] (www.ebi.ac.uk/chebi). The ligand preparation was done using "Prepare ligands" protocol at Discovery Studio 2.5. The preparation of ligand involved removal of duplicate structure, generation of the tautomer, isomers, Lipinski filter, change of ionization state and generating 3D structure.

4.2.1.2 Multiple sequence alignment of DNMTs nucleotide sequence

Multiple sequence alignments were carried out for DNMT3A/a DNMT3B/b amino acid sequences of human and mouse to determine the conserved amino sequence of the active site residues. Amino acid sequences of UniProt accession number Q9Y6K1**,** Q9UBC3, O88508, and O88509 were retrieved from Uniprot database [463] located at (http://www.uniprot.org/). The sequence were aligned with a window interface of CLUSTALW [464]. The BLOSUM62 substitution matrix [465] was used, with a gap start penalty of 10 and a gap extend penalty of 0.2.

4.2.1.3 Docking protocols

In order to generalize the ligand and protein conformation on binding accuracy, variable docking algorithms were employed. Each algorithm constitutes an alternative way of scoring and treating ligand flexibility keeping protein structure in a rigid state in order to reduce conformation search space. The various algorithms used for handling ligand flexibility constituted "Lamarckian Genetic Algorithm (LGA) (Autodock) [466] , monte-Carlo conformational search (LigandFit) [467], descriptor matching (Glide_XP) [468] and molecular dynamics simulated annealing (CDOCKER) [469].

- **Autodock**

AutoDock 3.05 is freely available software availed from Scripps Research Institute. Here, the inhibitors are treated as flexible ligands by modifying their rotatable torsions while the protein template is considered to be a rigid receptor. The minimised protein structures 3PTA, 2QRV and 3AV5 were used as target structures. Grid maps were prepared using auto grid to fix the active site of protein having specific co-ordinates and dimension of 40 × 40 × 40 and a resolution of 0.375 Å. Docking parameters were set as: number of individuals in the population (set to 150), maximum number of energy

evaluations (set to 2500000), maximum number of generations (set to 27000), and number of hybrid GA-LS runs (set to 100).

- **Glide_XP**

The protein structure was prepared using the protein preparation module of Schrödinger software. The co-crystallized water molecules were removed. All the selected ligands were assigned an appropriate bond order using the LigPrep 2.4.107 script and converted to .mae format (Maestro, Schrödinger, Inc.) and optimization was carried out by means of the OPLS_2005 force field. The Protein ligand docking studies were performed using Maestro 9.1.107. Parameters having default values were selected and docking was carried out using Glide Extra Precision (XP Glide), version 4.5.19. After the complete preparation of protein and ligand for docking, receptor-grid files were generated. Here van-der Waal radii were scaled of receptor atoms by 1Å with partial atomic charge of 0.25 for running the grid generation module.

- **CDOCKER**

CDOCKER is an in-house docking protocol of "Accelyrs Discovery studio". For initial stage MD a softcore potential is used. Each of the structures from the MD run are then located and fully minimized. The solutions are then clustered according to position and conformation and ranked by energy. CHARMm charges are used for the protein structure, i.e., the param19/toph19 parameter set38 using only polar hydrogens. CDOCKER only allows for flexible ligand treatment. Here the docking model constitutes receptor in its rigid state and static protein conformation of biding site is described using 1.0 Å grid and for every point grid, interaction energies of 20 types of probe atoms are calculated. The three dimensional grid is calculated such that radius of 8Å extend in all directions from any atom in the ligand. Subsequent to simulated annealing conformational search of the flexible ligand, the grid is removed minimization of all atoms of protein-ligand is performed by fixing the coordinates of the protein using the standard all atom potential function with a distance dependent dielectric (RDIE). This interaction energy is taken as the score for the final ligand pose.

- **LigandFit**

LigandFit is another docking programme of Accelyrs Discovery Studio. It is based on protein minimization using steepest descent (gradient <0.1) and conjugate gradient algorithms (gradient <0.01) of CHARMm force field. The active site determination includes within 10Å radius from the centre of the bound ligand. Docking was performed with monte-carlo simulations using the CFF95 force field. The grid resolution was set to 0.5Å (default), and the ligand accessible grid was defined such that the minimum distance between a grid point and the protein is 2.0Å for hydrogen and 2.5Å for heavy atoms. The

grid extends from the defined active site to a distance of 5Å in all directions. The top 10 conformations were saved after rigid body minimizations of 1,000 steps. The scoring was performed using set of scoring functions (including Dock_score, -PMF, -PLP1 and –PLP2) implemented in LigandFit module. The combination of consensus scoring functions was employed to obtain the most preferable output conformation.

4.2.1.4 Molecular dynamics simulation analysis

The molecular dynamics of the protein–inhibitor complex provides understanding to the flexibility associated with ligand conformational change and thus provides an insight into the molecular basis for the inhibition. All of the simulations were carried out using the GROMACS 4.5.5 package with an identical protocol [470]. The best orientation obtained out of docking of protein-ligand complex was used for simulation. We performed the simulation for ligands SAH, EGCG and procyanidin B2 and their respective binding with human DNMT1, DNMT3a and mouse DNMT1. Here we separated the ligands from protein in order to prepare protein and ligand topology file separately. We used the GROMOS96 43a1 force field to generate topology file for protein. The ligand topology file was generated using PRODRG server employing GROMOS96.1 force field [471]. The protein was solvated in a dodecahedron box with edges 1 nm in length using the explicit solvent–simple point charge model (SPC216 water molecules), which generated the water box. The next step followed the 5000 steps of steepest descent minimization and position restrained dynamics to distribute water molecule throughout in 100 ps. The simulation was carried out at 300 k of constant temperature and pressure of 5000 steps for 100 ps using Nose–Hoover method (nvt) and the Parrinello–Rahman method respectively. Once the system was equilibrated with desired temperature and pressure, the final step was to release the position restraints and run production of 500000 steps for 1000ps for data collection.

4.2.1.4 Evaluation of free binding energy by MM-PBSA method

Free energy of binding was calculated using the molecular mechanics-Poisson-Boltzmann surface area (MM-PBSA) method implemented in Amber12 [472]. For each complex, a total number of 40 snapshots were taken from an interval of 50 ps from the final 5000 ps of the MD simulation. The MM-PBSA method and nmode module of Amber 12 were implemented to calculate the binding free energy of the inhibitor and the detail is summarized as:

$$\Delta G_b = \Delta E_{MM} + \Delta G_{sol} - T\Delta S, \quad (1)$$

where, ΔG_b is the binding free energy in solution consisting of the molecular mechanics free energy (ΔE_{MM}) and the conformational entropy effect to binding ($-T\Delta S$) in the gas phase, and the solvation free energy (ΔGsol). ΔE_{MM} was evaluated as:

$$\Delta E_{MM} = \Delta E_{vdw} + \Delta E_{ele}, \quad (2)$$

where, ΔE_{vdw} and ΔE_{ele}stand for van der Waals and electrostatic interactions in the gas phase, respectively. The solvation free energy (ΔGsol) was calculated in two steps:

$$\Delta G_{sol} = \Delta G_{pol} + \Delta G_{nonpol} \quad (3)$$

where, ΔG_{pol} and ΔG_{nonpol} are polar and nonpolar components of the salvation free energy, respectively. The ΔG_{sol} was calculated with the PBSA module of Amber 12 suite. The nonpolar contribution of the solvation free energy is calculated as a function of the solvent-accessible surface area (SAS), as follows:

$$\Delta G_{nonpol} = \gamma\,(SAS) + \beta, \quad (4)$$

where, the values of empirical constants γ and β were set to 0.00542 kcal, (molÅ2) and 0.92 kcal, mol, respectively. The contributions of entropy (TΔS) to binding free energy can be evaluated as the sum of change in the translational, rotational, and vibrational degrees of freedom, as follows:

$$\Delta S = \Delta S_{translational} + \Delta S_{rotational} + \Delta S_{vibrational} \quad (5)$$

TΔS was calculated using classical statistical thermodynamics and normal-mode analysis.

4.2.1.5 Residue-inhibitor interaction decomposition

The interaction between inhibitors and each residue of hDNMT1, DNMT3A, and mDNMT1 was calculated using molecular mechanics of Generalized Born Surface Area (MM-GBSA) decomposition process module of Amber 12 [473, 474]. The binding interaction of each inhibitor-residue pair was evaluated in terms of electrostatic (ΔE_{ele}) contribution, van der Waals (ΔE_{vdw}) contribution in the gas phase, polar solvation (ΔG_{pol}) and nonpolar solvation (ΔG_{nopol}) contributions.

$$\Delta G_{\text{ inhibitor-residue}} = \Delta E_{ele} + \Delta E_{vdw} + \Delta G_{pol} + \Delta G_{nopol} \quad (6)$$

The polar contribution (ΔGpol) to solvation energy was calculated by using the GB (Generalized Born) module.

4.2.2 In-vitro analysis of gene expression, DNMT activity, and toxicity

4.2.2.1 Reagents

3-(4,5-dimethylthiazol-2-yl)-2,5-diphenyltetrazolium (MTT), dimethylsulfoxide (DMSO), epigallocathechin-3-gallate (EGCG), procyanidin B2-3, 3'-di-O-gallate (procyanidin B2), S-Adenosyl-L-homocysteine (SAH) and trypsin were purchased from Sigma-Aldrich (St Louis, MO, USA). Dulbecco's modified Eagle medium (DMEM), Fetal bovine serum (FBS) (sterile-filtered, South American origin) were purchased from Invitrogen (Carlsbad, CA, USA).

4.2.2.2 Cell culture

Human invasive breast cancer cell line MDA-MB-231 and the immortalized human keratinocyte cells HaCaT were obtained from National Centre for Cell Science, Pune. MDA-MB-231 and HaCaT cells were maintained in DMEM supplemented with 10%

FBS and 100 U Penicillin and 0.1 mg Streptomycin at 37 °C in a humidified incubator (5% CO_2).

4.2.2.3 DNMT inhibition assay

Cultured MDA-MB 231 cells were harvested to prepare nuclear extract according to standard protocol [475]. After quantification of protein by Bradford method, nuclear extract having 7.5 µg of protein was used to measure total DNMT activity using the EpiQuik DNA Methyltransferase Activity Assay Kit (Epigentek, Inc.) according to the manufacturer's protocol. As per protocol, the nuclear extracts were added to the pre-coated substrate and then AdoMet was added followed by inhibitors (EGCG and procyanidin B2) at varying concentrations and incubated for 2 h at 37°C. The above incubated nuclear extracts were exposed to capture antibody against 5-methyl cytosine for 1 h and the detection antibody for 30 min at room temperature. Finally, developer solution was added, and absorbance was recorded using microplate reader spectrophotometer (Perkin-Elmer, Waltham, MA, USA) at 450 nm with an optional reference wavelength of 655 nm. The assay was conducted to identify the IC_{50} for the inhibitors against DNMT activity. The log dose response curve was plotted, and IC_{50} was calculated using the following equation.

$$y = A_1 + \frac{A_2 - A_1}{1 + 10^{(LOGx0 - x)p}}$$

$$\mathrm{IC}_{50} = 10^{LOGx0}$$

Here A_1 and A_2 are bottom and top asymptote, p is the hill slope, and LOG_{x0} is the center of the curve. The graph for log dose response was plotted using GraphPad Prism software.

4.2.2.4 Quantitative reverse transcription PCR (qRT-PCR) of DNMT target genes

MDA-MB-231 cells were treated with EGCG and procyanidin B2 for 24 h at their respective sub-lethal concentration. Total cellular RNA was isolated with Tri Reagent (Sigma) according to the manufacturer's instructions. Reverse transcriptase reactions were performed using Revert-Aid First Strand cDNA Synthesis Kit (Thermo Scientific) with 1 µg of RNA. qRT-PCR was performed using SYBR® Green JumpStart™ Taq Readymix in the Realplex4Eppendorf system. The primer sequences of DNMT target and the three DNMT genes are enlisted in Table 4.1. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) gene was used as an internal control.

Table 4.1 Primer sequences of DNMT target and DNMT genes.

Gene name	Primers sequence	Amplicon size (bp)
E-cadherin	F 5'-CGAGAGCTACACGTTCACGG-3' R 5'-GGGTGTCGAGGGAAAAATAGG-3'	119
Maspin	F 5'-GGAATGTCAGAGACCAAGGGA-3' R 5'-GGTCAGCATTCAATTCATCCCTT-3'	139
BRCA1	F 5'-ACAGCTGTGTGGTGCTTCTGTG-3' R 5'-CATTGTCCTCTGTCCAGGCATC-3'	107
DNMT1	F 5′-GGCTGAGATGAGGCAAAAAG-3′ R 5′-ACCAACTCGGTACAGGATGC-3′	112
DNMT3A	F 5′-TATTGATGAGCGCACAAGAGAGC-3′ R 5′-GGGTGTTCCAGGGTAACATTGAG-3′	111
DNMT3B	F 5′-AATGTGAATCCAGCCAGGAAAGGC-3′ R 5′-ACTGGATTACACTCCAGGAACCGT-3′	191
GAPDH	F 5′-GGAGCGAGATCCCTCCAAAAT-3′ R 5′-GGCTGTTGTCATACTTCTCATGG-3′	197

4.2.2.5 Evaluation of cytotoxicity of SAH, EGCG and procyanidin B2

The cytotoxicity of SAH, EGCG and procyanidin B2 was evaluated in both MDA-MB-231 and HaCaT cells by colorimetric MTT assay. In brief, the cells in the logarithmic phase were plated in 96-well flat-bottom culture plates at a density of 4000 cells/well and treated with SAH, EGCG and procyanidin B2 at six different concentrations for 24 h. The cytotoxic effect of SAH, EGCG and procyanidin B2 was determined by measuring the absorbance intensity of formed formazan solution at 595 nm by using microplate reader spectrophotometer (Perkin-Elmer, Waltham, MA, USA). Water in case of SAH and DMSO (0.01%) in case of EGCG and procyanidin B2 was used in the control treatment. All the experiments were done in triplicate, and the cell viability was determined by percentage at varying concentration of drugs.

4.2.2.6 Statistical analysis

The statistical significance of the above result was analyzed using Student's t-test by SPSS software. Data are expressed as a mean ± standard deviation. The significant difference in IC_{50}, LC_{50} and gene expression between two groups (EGCG and procyanidin B2 treatment) was computed using one-way ANOVA, and the p-value was evaluated at the threshold of 0.05.

4.3 Results

Till date, many non-nucleoside inhibitors of DNMTs have been identified. We have enlisted some of them through a literature survey and taken into consideration for this study [Figure 4.1]. Several experimental analyzes depict the interaction of specific non-nucleoside inhibitor of the target DNMT enzyme; however, the comparative analysis of the known non-nucleoside inhibitors was not done prior to our present work. We investigated to identify which among these inhibitors is best in inhibiting DNMTs activity, including both DNMT1 and DNMT3a. Our results obtained byanalyzing the existing inhibitors and their analogsare supporting procyanidin B2, a novel phytochemical to be the best effective in reducing DNMT activity. The efficiency of ligands has been analyzed by in-silico and in-vitro experiments. *In-silico* analyzes involve the ligand interaction with DNMTs at both static and dynamic conditions. The in-vitro study includes cell viability assay, relative gene expression study on the application of different drugs of varying concentration and identification of drug concentration at which it reduces DNMT activity to 50%. A detailed report is presented below.

4.3.1 Comparison of active site loop of DNMT3A/a and DNMT3B/b

Figure 4.2a shows the sequence alignment of the catalytic domain of human DNMT3A, DNMT3B and mouse DNMT3a, and DNMT3b. It is apparent that the catalytic domains superimpose well due to their high sequence similarity. The active-site loop of human DNMT3A (2QRV) (residues 708–729) superimposes well with the mouse DNMT3a (residues 704-725). These DNMT3A/a active site residues also depict significant overlapping with human DNMT3B (residues 649-670) and mouse DNMT3b (residues 655-976) with an exception of amino acid Ile (isoleucine) residue substituted by Asn (aspargine) (marked yellow). The key amino acid residues for the catalysis and cofactor binding are found to be conserved. Thus, it may be assumed that the inhibitors will have a similar effect on both DNMT3A/a and DNMT3B/b. So, for our further studies we have focused on the detailed analysis of the interaction of non-nucleoside inhibitors on hDNMT1 (3PTA), mDNMT1 (3AV5) and DNMT3A (2QRV). The non-nucleoside inhibitors were docked at the SAH-binding pocket present in the active site of 3PTA, 2QRV and 3AV5 [Figure 4.2, b-d].

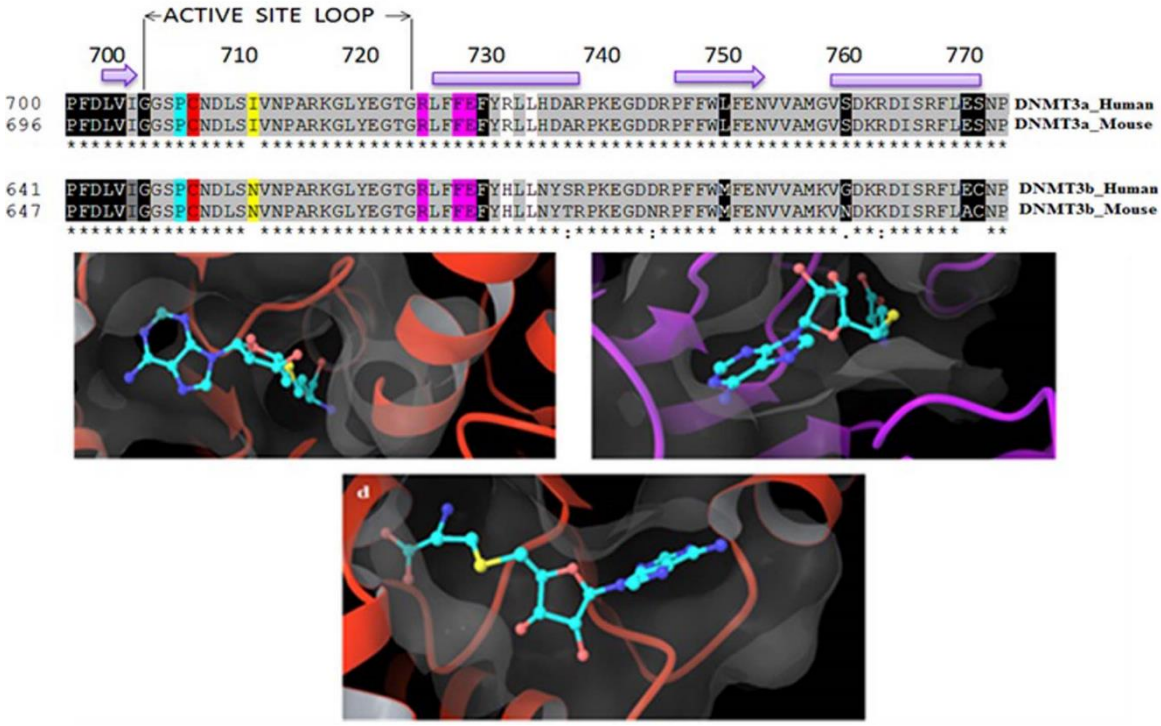

Figure 4.2 (a) Depiction of conserved active site regions of DNMT3A/a and DNMT3B/b in human and mouse. The numberingof the sequences corresponds to the mouse orthologs. Grey highlighted regions were conserved sites along DNMT3a and DNMT3b while red highlighted regions depict amino acid Cysteine (C), a nucleophilic group. Positions highlighted in yellow exhibit unconserved residues and pink highlighted regions exhibit common interacting regions between DNMT3A/a-DNMT3L and DNMT3B/b-DNMT3L. Neighbour sites of the active site with differences in amino acids are marked as green. Binding of SAH to active site of pocket of (b) hDNMT1 (3PTA), (c) DNMT3A (2QRV), (d) mouse DNMT1 (3AV5) as obtained by modelling using Maestro 9.1.107 of Schrodinger.

4.3.2 Interactions of DNMTs with non-nucleoside inhibitors

The enlisted non-nucleoside inhibitors in Figure 4.1 constituting both synthetic (hydralazine, RG108, procaine, procainamide) and natural compounds (curcumin, EGCG, parthenolide, mahanine) are identified to inhibit DNMT activity through different mechanisms. The binding affinity of non-nucleoside phytochemicals were analyzed with respect to synthetic inhibitors. These inhibitors were docked at SAH-binding pocket of DNMTs, and their binding affinity was analyzed by different algorithms of glide_XP, autodock, cdocker and LigandFit [Figure 4.3].

Among all non-nucleoside inhibitors, EGCG binds with the highest efficacy exhibiting glide score of -10.56, -10.13 and -11.0 kcal/mol, when bound to catalytic domain of 3PTA, 2QRV and 3AV5, respectively [Figure 4.3 a]. However, curcumin binds only to 3PTA showing glide score of -10.52 kcal/mol. The binding energy analysis was

further confirmed with other algorithms [Figure 4.3 b-d]. The more negative binding energy (ΔG) reflects the stability of a complex and is a consequence of non-covalent interactions, mainly hydrogen bonding, van der Waals and electrostatic forces by the residues of the binding pocket. EGCG interacts with 3PTA via 5 hydrogen bonds. Six hydrogen bonds are formed on an interaction with 2QRV and 3AV5 active site residues, respectively. The binding residues and hydrogen bond donor and acceptor groups and their respective bond length have been depicted in Figure 4.4 (a-c).

Thus, among the enlisted non-nucleoside inhibitors, EGCG have been identified as the most potent inhibitor to diminish DNMTs activity. Thereafter, we examined a novel set of other phytochemicals which would be better than EGCG to reduce DNA methylation density and cellular toxicity.

4.3.3 Interaction of DNMTs with novel set of phytochemicals/compounds

In quest of identification of novel inhibitors, thirty-two EGCG analogs were retrieved from PubChem database (http://pubchem.ncbi.nlm.nih.gov/). It mainly constituted the polyphenolic groups of phytochemicals. Some of them are eryvarinol A, mangiferin, isomagniferin, 3, 4'- 5-trihydroxystilbene, theaflavin-di-gallate, procyanidin B2-3, 3'-di-O-gallate and others. Extensive chemoinformatic analyzes of these compounds were done to expand the medicinally relevant chemical space. Compounds selected were docked into the active site pocket of the crystallographic structure of DNMTs using Glide_XP protocol. According to the Glide score, procyanidin B2-3,3'-di-O-gallate (Prc) ranked, first of all, the analogs, indicating that it may possess higher inhibition potential against DNMTs. The score obtained was as high as -13.95, -11.53 and -14.9 kcal/mol when docked to 3PTA, 2QRV, and 3AV5, respectively. The comparative analysis of binding energy, hydrogen bond donor-acceptor groups and the bond length for SAH, EGCG and procyanidin B2 with respective DNMTs is depicted in Table 4.2. The binding energy of procyanidin B2 was comparatively higher than EGCG and SAH. The increased binding energy is characterized by increased non-covalent interactions, mainly, involving hydrogen bonding, van der Waals, and electrostatic forces. Procyanidin B2 on an interaction with 3PTA exhibit -7.54, -6.25 and -1.85 kcal/mol of hydrogen bonding, hydrophobic and electrostatic interactions, respectively. Similarly, the respective non-covalent interactions are identified to be -5.73, -3.59 and -1.65 kcal/mol on interaction with 2QRV, while -8.04, -5.42 and -1.37 kcal/mol with 3AV5.

The detailed binding residues of procyanidin B2 with DNMTs (3PTA, 2QRV and 3AV5), hydrogen bond donor and acceptor groups, their respective bond lengths and pi (π)-cation interactions have been depicted in Figure 4.5 (a-c). The principle pi (π)-cation interaction is identified between the phenolic ring of procyanidin B2 and NH1, NH2

group at R1312 of 3PTA. Similarly, it also forms pi (π)-cation interaction with NH1 and NH2 group at R684 and R887 of 2QRV.

Finally, DNMT-inhibitor (SAH, EGCG, and procyanidin B2) complexes exhibiting higher binding score were selected and subjected to molecular dynamics simulations in explicit aqueous solution.

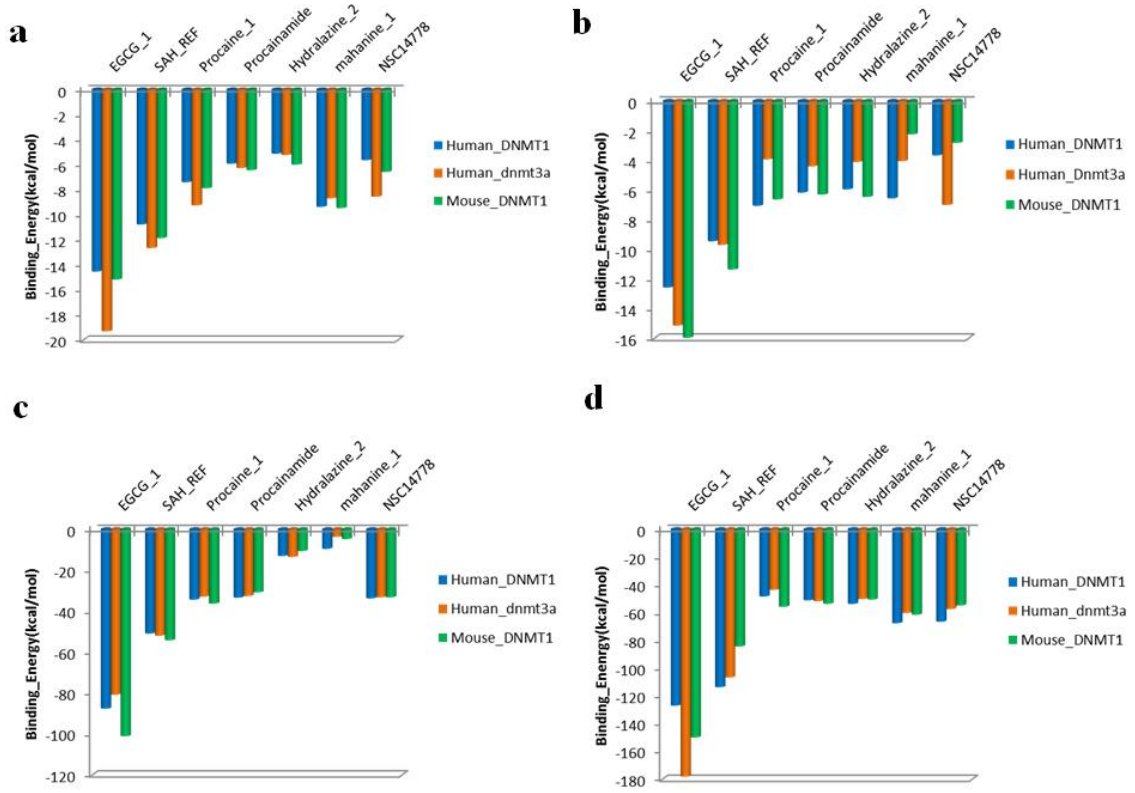

Figure 4.3 Docking of non-nucleoside inhibitors with 3PTA (blue), 2QRV (orange) and 3AV5 (green) with (a) Glide_XP, (b) Autodock, (c) CDOCKER and (d) LigandFit. The binding energy was determined in terms of kcal/mol. Among all known non-nucleoside inhibitors, EGCG was identified to have higher binding energy at SAH-binding pocket of DNMT.

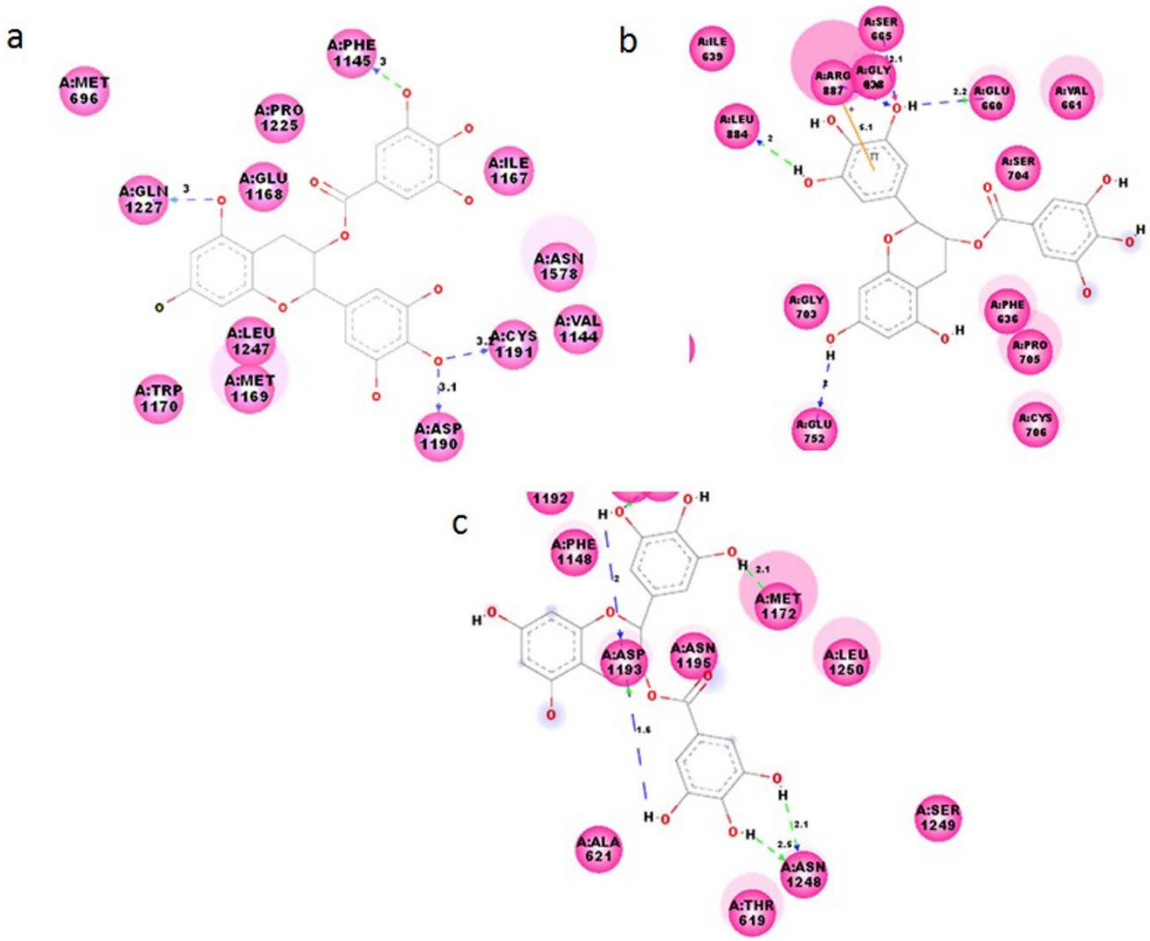

Figure 4.4: Depiction of the interaction of EGCG with (a) 3PTA, (b) 2QRV and (c) 3AV5 via hydrogen bonds of specific length. The hydrogen bonds and their respective bond lengths in Å have been shown. The π-cation interactions have been displayed by a solid orange line. The figure is produced by Accelrys discovery studio.

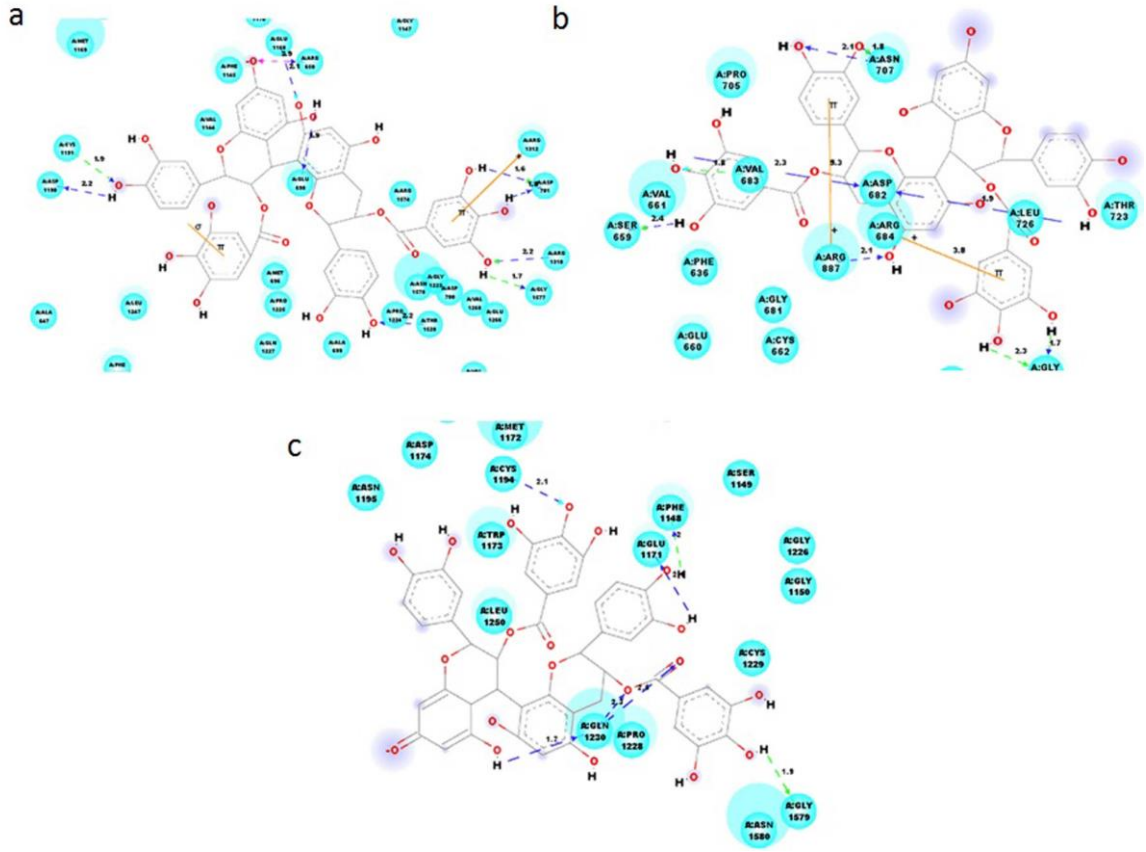

Figure 4.5 Depiction of interaction of Procyanidin B2 with (a) 3PTA), (b) 2QRV, (c) 3AV5 via hydrogen bonds defining specific length. The hydrogen bonds and their respective bond lengths in Å have been shown. The π-cation interactions have been shown by a solid orange line. The figure is produced by Accelrys discovery studio.

Table 4.2. Detailed study of interaction of SAH, EGCG and Procyanidin B2 with DNMTs

Protein (PDB)	Ligand	Glide Score (kcal/mol)	Residues at binding site	H-bond donor-acceptor groups	H-bond distance (Å)
Human_DNMT1 (3PTA)	SAH	-9.44	G1223,S1146,G1222, L1151,E1266,G1147, E1168,M1169,I1167 D1190,E1189,C1191, P1225,F1145	L1151:H->SAH: O E1168: HE2->SAH:O3 E1168: HE2->SAH:O2 C1191: H->SAH:N1 SAH: HN3->E1266:OE1 SAH: HN3>E1266:OE2 SAH: HN1>D1190:OD2	1.8 2.1 2.1 2.3 1.9 2.1 1.8
	EGCG	-10.56	E1168,L1247,C1191,M 1169,D1190,T1170,Q1 227,M696,P1225, E698,F1145,G1147, L1247,R650,E1168, T1170,F1146,V1144, I1167,M1169,C1191, D1190,L1247,	C1191: H->EGCG:O7 Q1227: HE22>EGCG:O3 EGCG: H6->M696:SD EGCG: H14>D1190:OD2 EGCG: H16->F1145:O	1.9 2 2.4 2.3 2.1
	Procyanidin B2 (Prc)	-13.95	P1225,M696,Q1227 E698,A699,P1224, T1528,N1578,G1223,D 700,V1268,E1266, G1577,R1310,D701, R1312, G1147	E1168:HE2->Prc:O5 C1191:H->Prc:O11 R1310:HH22->Prc:O11 T1528:HG1->Prc:O14 Prc:H16->E698:OE1 Prc:H23->D1190:OD2 Prc:H31->G1577:O	2.9 1.9 2.2 2.2 2.1 2.2 1.7

				Prc:H32->D701:OD1 Prc:H33->D701:OD1	1.8 1.6
Human DNMT3a(2QRV)	SAH	-9.69	E660,F636,D637, W889, T641, S888, S659, R887, P705, V683, D682,V661, L726	T641:HG1->SAH:OXT V863:H->SAH:N1 W889:H->SAH:OXT SAH:HO->E660:OE1 SAH:H1->E660:OE2 SAH:H3->D682:OD1	2A 2.4 1.9 2A 2.1 2.4
	EGCG	-10.13	E660,S665,G638, R887,L884,G703, E752,F636,P705, V683,L726,S704, V661	S665:HG->EGCG:O8 R887:HH22>EGCG:O8 EGCG:H5->E752:OE1 EGCG:H5->E660:OE2 EGCG: H13->L884:O EGCG:H15->E660:OE1 V683:H>Prcd:O20	2.1 2.1 2 2.2 2 2.2 1.8
	Procyanidin B2 (Prc)	-11.53	N707,P705,V683, V661,S659,F636, G681,R887, R684, D682,R883,G722, L726,T723	N707:HD22->Prc:O12 N707:HD22>Prc:O14 R887:HH21->Prc:O7 Prc:H23->D682:OD1 Prc:H30->G722:O Prc:H31->G722:O Prc:H33->S659:OG Prc:H34->D682:OD2	2.1 2.1 2.3 1.7 2.3 2.4 2.3 2.2
Mouse_DNMT1 (3AV5)	SAH	-10.35	E1171,N1580,G1225,G1150,F1148,S1149,A1581,D1146,V1582,L115	G1152:H->SAH:O G1153:H->SAH:O L1154:H->SAH:OXT	2.1 2 2.1

			4,E1269,G1153,G1152,	C1194:H->SAH:N1	2.2
			C1151,V1147,I1170,P1	N1580:HD21->SAH:O3	2.1
			228,C1194,N1195,E119	V1582:H->SAH:OXT	2.1
			2,D1193,L1250, M1172	SAH:HN2->S1149:O	2.2
				SAH:HO->N1171:OE1	2.1
				SAH:HO->N1171:OE2	2.1
				SAH:HN1->D1143:OD1	3.6
				SAH:HN1->D1193:OD1	1.8
				SAH:H1->F1148:O	2.4
			C1194,P1228,W1173,F	M1172: H->EGCG:O6	2.1
	EGCG	-11.0	1148,E1192,A621,N12	C1194: H->EGCG:O8	1.9
			48,D1193,N1195,L125	EGCG: H15>D1193:OD1	2
			0,M1172	EGCG: H17->N1248:O	2.1
				EGCG: H17- N1248: OD1	2.5
				EGCG:H18>D1193:OD2	1.6
				C1194:H->Prc:O17	2.1
	Procyanidin	-14.9	C1194,W1173,L1250,	Q1230:HE11->Prc:O4	2.3
	(Prc)		D1174,N1196,Q123,	Q1230:HE22->Prc:O13	2.5
			P1228,N1580,G1579,	Prc:H16->Q1230: OE1	1.7
			C1229,G1150,G1226,	Prc:H24->E1171:OE1	2.1
			S1149,E1171,F1148	Prc:H27->F1149:O	2
				Prc:H33->G1579:O	1.9

4.3.4 Molecular dynamics simulation of DNMT-inhibitor complexes

The molecular dynamics simulations were implemented to authenticate the docking results and decipher efficacy inhibiting DNMTs activity. In order to maintain proper orientation of ligand distance, restraints were applied to inhibitors in the initial few picoseconds (ps) and then whole complexes were allowed to move freely. The docked conformations were analysed by examining their relative total energy scores, protein backbone root mean square deviation (RMSD), total hydrogen bonds, van-der-Waals interaction, electrostatic interaction and root mean square fluctuation (RMSF) of active site residues.

Here, the stability of DNMT-inhibitor complex was determined in terms of total energy (kinetic Ek+ potential Ep) at the given temperature of 310K and pressure of 1atm. We plotted the fluctuation in total energy as a function of constant time gaps. The average total energy (kJ/mol) for EGCG and procyanidin B2 was very high at SAH-binding pocket of 3PTA, 2QRV, and 3AV5. In Figure 4.6 it is evident that both the inhibitors oscillated with a nearly same frequency of -1.718×10^6, -6.868×10^5 and -4.071×10^6 kJ/mol.

The RMSD of each protein relative to binding of respective inhibitors (SAH, EGCG, and procyanidin B2) was calculated to monitor the stability of each trajectory. Here, we mainly focussed on the active site residues of DNMTs. The stability of protein-inhibitor complex was analyzed by aligning heavy atoms of the complex to the crystal structures of proteins using the mass-weighted least square fitting method. The RMSD plot exhibited the structures which are stable during the course of MD simulations. The RMSD plot unveiled an increase in deviation at first 200ps of the production phase. This is because the equilibration phase was performed with restraints on complex, while the restraints in production phase were released. From Figure 4.7 one can easily watch that the inhibitor procyanidin B2 attains equilibrium after 500ps and on average comparatively lesser RMSD of 3.0 Å, 1.6 Å, and 2.4 Å is attained on interaction with 3PTA, 2QRV, and 3AV5 respectively. Thus from this observation it's evident that procyanidin B2 forms more stable complex than SAH and EGCG with DNMTs. The stability of an enzyme-ligand complex is characterized by ligand binding mode inside the active site pocket of the enzyme, and the binding force includes strong hydrogen bonds, electrostatic and van der Waals interactions. The intermolecular hydrogen bond plots between enzyme and inhibitor are shown in Figure 4.8.

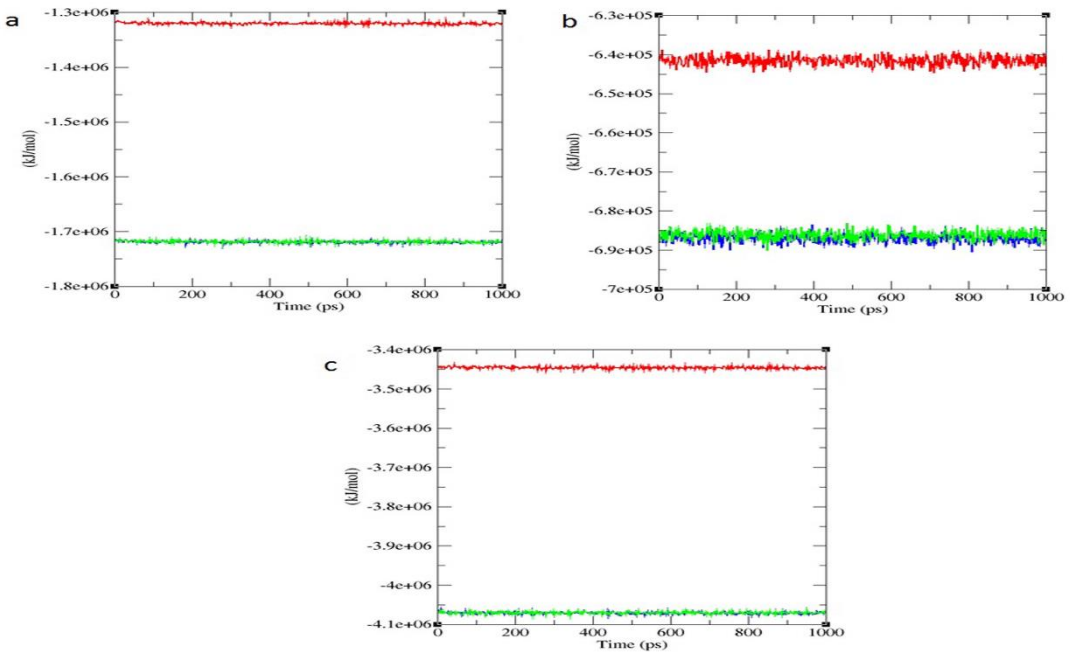

Figure 4.6 Total energy at each ps on interaction of SAH (red), EGCG (blue) and Procyanidin B2 (green) with (a) 3PTA, (b) 2QRV, (c) 3AV5 in kJ/mol. Total energy has been identified to be elevated for EGCG and procyanidin B2 on interaction with DNMT1 in human and mouse and DNMT3A human.

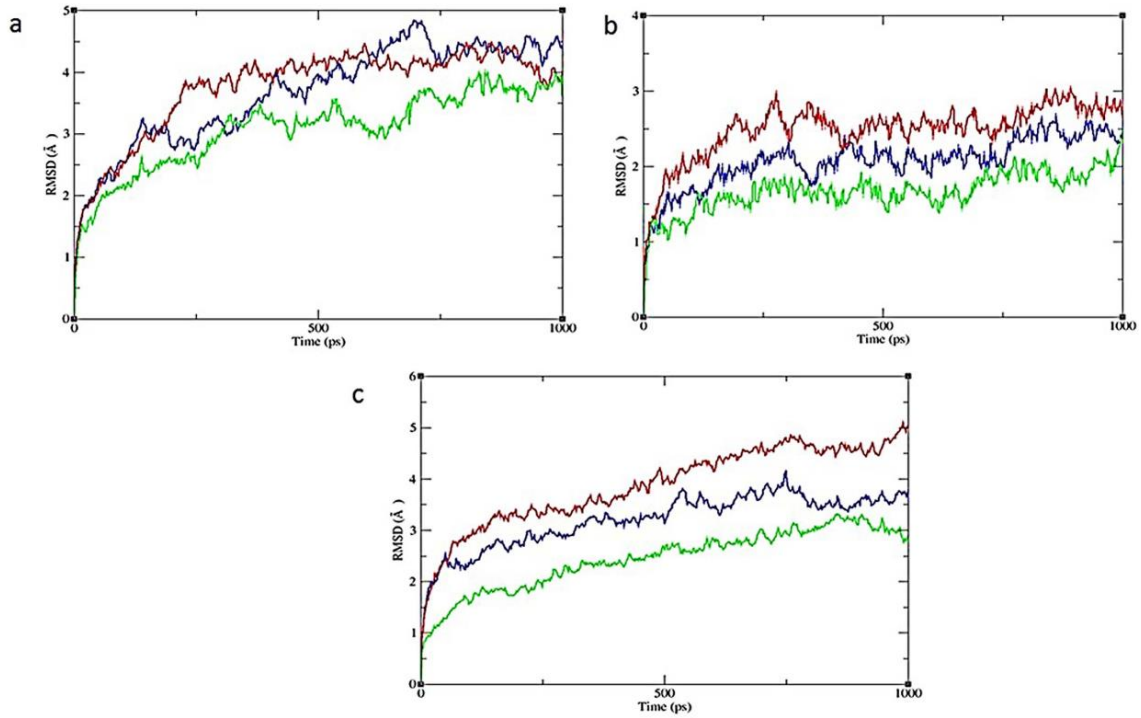

Figure 4.7 RMSD plot with respect to time in ps on binding of SAH (red), EGCG (blue) and Procyanidin B2 (green) with (a) 3PTA, (b) 2QRV and (c) 3AV5 in Å. RMSD is calculated for heavy atoms with reference to their respective orientation in the crystal structures. Procyanidin B2

exhibits the least deviation on interaction with DNMT and forms most stable complex as compared to SAH and EGCG.

On an average procyanidin B2 forms a higher number of hydrogen bonds than SAH and EGCG with the cognate donor/receptor in the vicinity of the binding pocket of DNMTs. It forms an average of 6.5 on interaction with 3PTA while 5.2 hydrogen bonds are formed when ligated to 2QRV and 3AV5. Further, non-bonding interaction constituting Columbic function and Lennard–Jones potential function were employed to calculate electrostatic and van-der-Waals interactions, respectively with a cut-off distance of 9Å. On the evaluation of both the parameters, it has been identified that van-der-Waals interaction take over the electrostatic interaction, thus favoring the interaction of inhibitors mainly to procyanidin B2 at the active site pocket. The average van-der-Waals energy (ΔEvdw) is identified to -367.743, -296.80 and -330.101 kJ/mol on interaction with 3PTA, 2QRV, and 3AV5. Thus from the above findings it can be inferred that hydrogen bonds and van der Waals energy dominates in the total binding energy profile for stabilizing the protein-inhibitor complex.

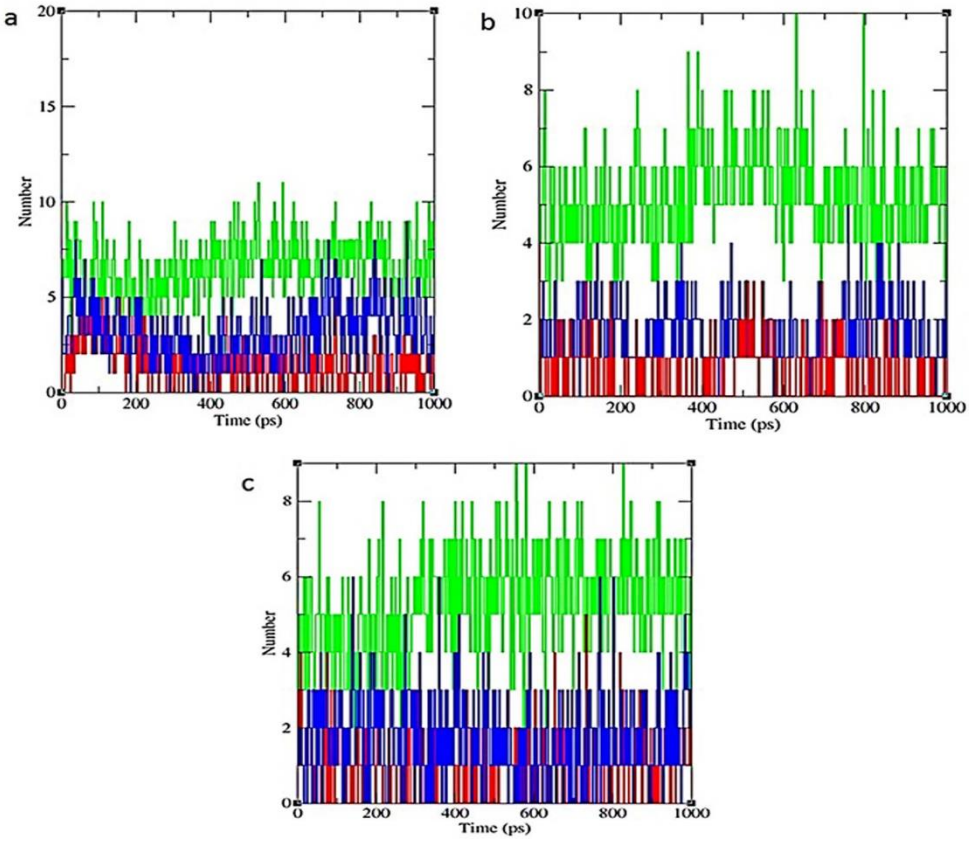

Figure 4.8 Observation of intermolecular hydrogen bond in Å between SAH (red), EGCG (blue), Procyanidin B2 (green) with active site residues of (a) 3PTA, (b) 2QRV and (c) 3AV5. On an average Procyanidin B2 forms a higher number of hydrogen bonds with the respective enzymes.

The residue flexibility of bound DNMT-inhibitor complexes was examined by analyzing the RMSF of the Cα atoms of each residue. The RMSF plot for different protein complexes has shown the flexible regions of the systems; however plot clearly displayed minimum fluctuation around the active site residues [Figure 4.9 a-c]. The amino acid residues exhibit an average fluctuation of 0.98, 0.69, and 1.1 Å in case of procyanidin B2 while 1.2, 0.8 and 1.13 Å in case of EGCG, on interaction with 3PTA, 2QRV, and 3AV5, respectively. The average fluctuation of amino acid residues in the interaction of SAH with these proteins has been found to be 1.3, 1.0 and 1.5 Å, respectively. From the above findings, we can infer that procyanidin B2 forms more stable complex because the residues fluctuation is comparatively lesser than others. Thus, stability of procyanidin B2-DNMT complexes has been established by total energy, RMSD of the protein backbone, non-covalent interactions and RMSF of active site residues measurements.

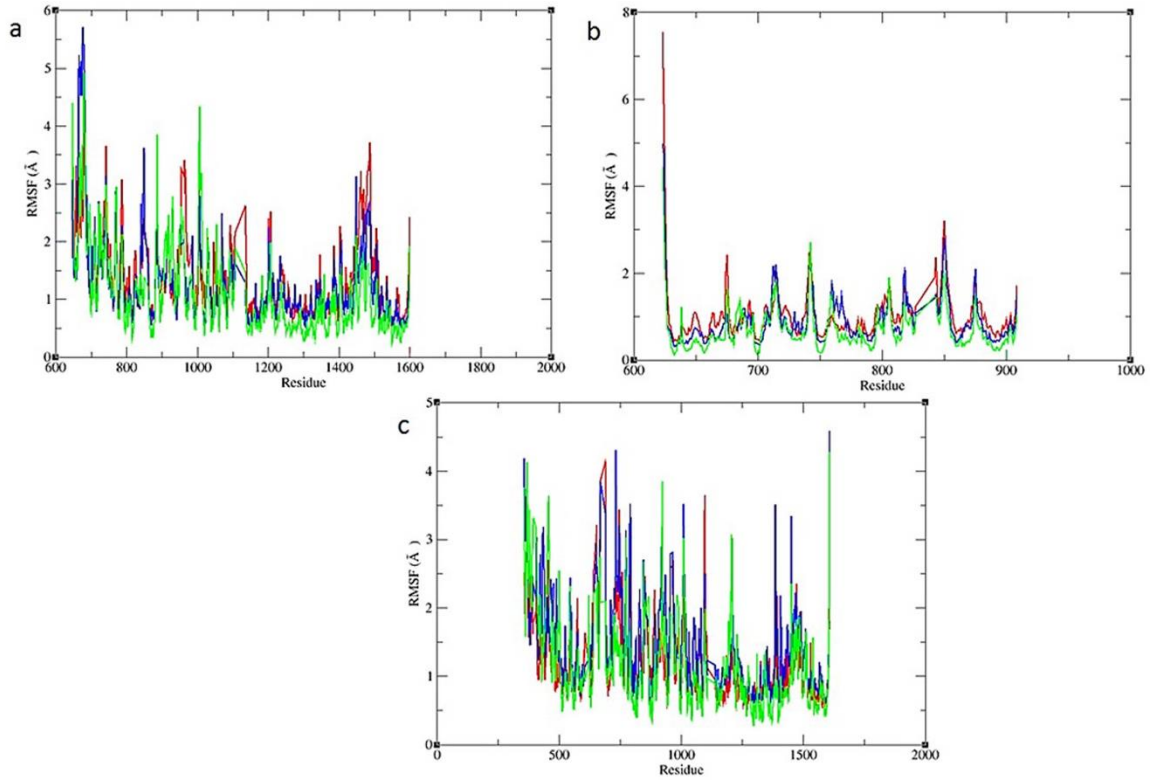

Figure 4.9 Root-mean-squared fluctuations (RMSF) of backbone atoms (Å) of (a) 3PTA, (b) 2QRV and (c) 3AV5 on binding to SAH (red), EGCG (blue) and Procyanidin B2. Active site residues for all protein-ligand complexes exhibit the least fluctuation, exhibiting the stability of the complexes.

4.3.5 Thermodynamic evaluation of DNMT-inhibitor complexes

Absolute free energies of binding were evaluated using MM-PBSA method in order to gain insight into the continuous spectrum of binding energy of hDNMT1, DNMT3A, and mDNMT1 with respect to SAH, EGCG and procyanidin B2. In this method the interaction and solvation energy is computed for complex, receptor and ligand in order to investigate average binding free energy. The detailed binding energy of protein-inhibitor complexes have been depicted in Figure 4.10 (a-c). The free energy of binding has been evaluated with respect to SAH at active site pocket of DNMTs. It is evident from Figure 10 that procyanidin B2 has highest binding efficiency for DNMTs (3PTA, 2QRV, and 3AV5). The binding energies of the 3PTA, 2QRV and 3AV5 with respect to procyanidin B2 are -16.64, -15.06 and -17.29 kcal/mol respectively. The binding efficiency of EGCG for the respective proteins has also been identified to be greater than SAH. The discrepancy in binding energy of protein-inhibitor complexes implicates that there are various mode of interactions From Figure 4.10, it is evident that the highest binding energy for procyanidin B2-DNMT complexes are consequence of gaseous phase electrostatic and van der Waals interactions. The nonpolar solvation energy (ΔGnp) also favors binding affinity in active site pockets of the enzyme. The relatively lower value of non-polar solvation energies indicates the closeness and integrity of packing of cavity regions. Moreover, the parameters like entropy ($-T\Delta S$) and polar solvation energy which is unfavorable for binding of inhibitors is identified to comparatively less for procyanidin B2 as compared to SAH and EGCG. Thus, the calculations of the free energy of binding of DNMT-inhibitor complexes elucidate that procyanidin B2 may be a novel inhibitor against DNMTs. Thereafter, the detailed mechanism of interaction was decomposed into inhibitor-residue pairs in order create a spectrum of inhibitor-residue interaction.

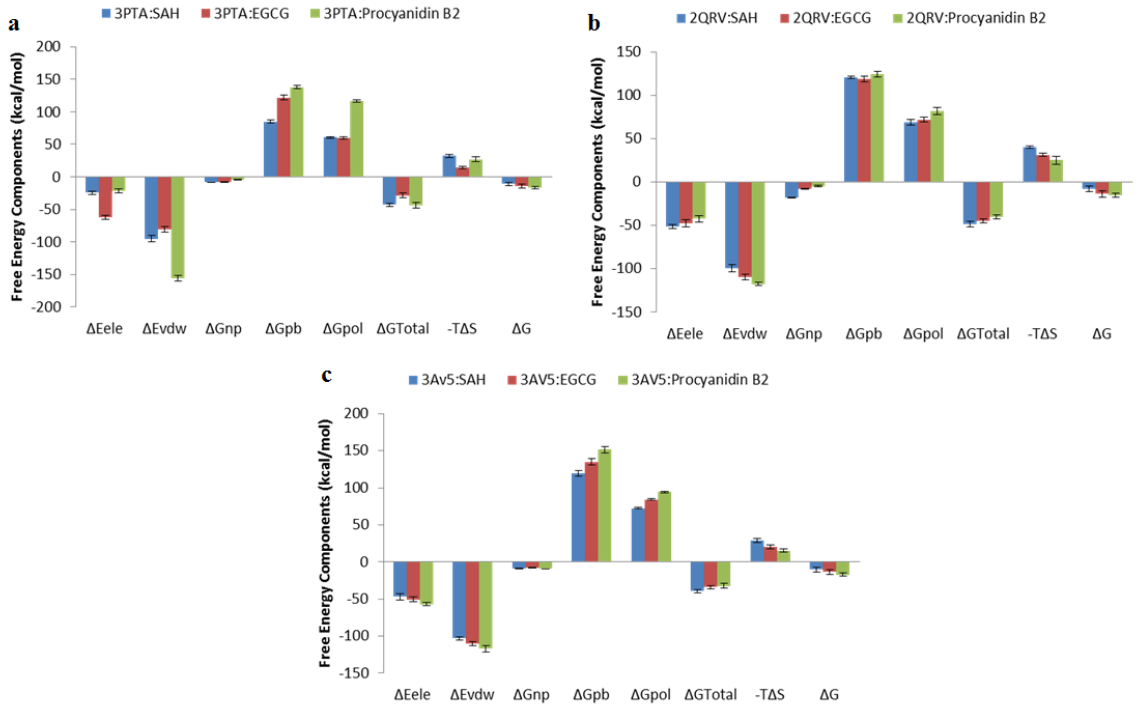

Figure 4.10 Energy components (kcal/mol) for the binding of SAH (blue), EGCG (brown) and procyanidin B2 (green) at binding pocket of (a) 3PTA (b) 2QRV and (c) 3AV5. ΔE_{ele}, electrostatic energy in the gas phase; ΔE_{vdw}, van der Waals energy; ΔG_{np}, nonpolar solvation energy; ΔG_{pb}, polar solvation energy, $T\Delta S$, total entropy contribution; $\Delta G_{total} = \Delta E_{ele} + \Delta E_{vdw} + \Delta E_{int} + \Delta G_{pb}$; $\Delta G = \Delta G_{total} - T\Delta S$. Error bars in indicates the difference.

4.3.6 Binding spectrum of residues at active site pocket of DNMTs

In order to obtain the detailed thermodynamic description of contribution from amino acid residue to the free energy of binding, the interaction energies were further decomposed to contributions of individual residues through MM-GBSA script in the AMBER 12 suite. The method of residue decomposition aids in understanding the atomistic detail of the mechanism of residue-inhibitor interactions. The detailed study of the contribution of residues to the binding energy for SAH, EGCG, and procyanidin B2 with respect to 3PTA, 2QRV and 3AV5 has been depicted in Figure 4.11 (a-c). Overall the major interaction at the active site of the pocket of 3PTA is the contribution of F1145, E1168, M1169, E1189, and C1191. The average binding energy of these residues is greater than -1 kcal/mol. Similarly, in case of mouse DNMT1 (3AV5), the residues F1148, E1171, M1172, C1194, and C1248 contributes to the elevation of binding energy. From the Figures 4.11a and 4.11c, it is evident that procyanidin B2 has higher interaction with these residues as compared to SAH and EGCG. Moreover, among all, the cysteine residue offers highest binding energy on interaction with procyanidin B2. C1191 and C1194 contribute to the binding energy of -5.10 and -4.21 kcal/mol to the interaction of

procyanidin B2 with 3PTA and 3AV5 respectively. In case of DNMT3A, the residues F636, D637, E660, D684, and W889 imparts to the elevation in the binding of inhibitors in the catalytic pocket of the enzyme. The D684 residue has a higher contribution to the interaction with procyanidin B2, and the average binding energy is identified to be -4.01 kcal/mol. This decomposition of free binding energy (ΔG) per residue is a consequence of van der Waals (ΔEvdw) energy, the sum of electrostatic and the polar solvation energy and the non-polar solvation (ΔGnp) energy. The change of entropy parameter is not included. Further ahead, the in-vitro studies were carried out to affirm the role of procyanidin B2 as a potent DNMT inhibitor.

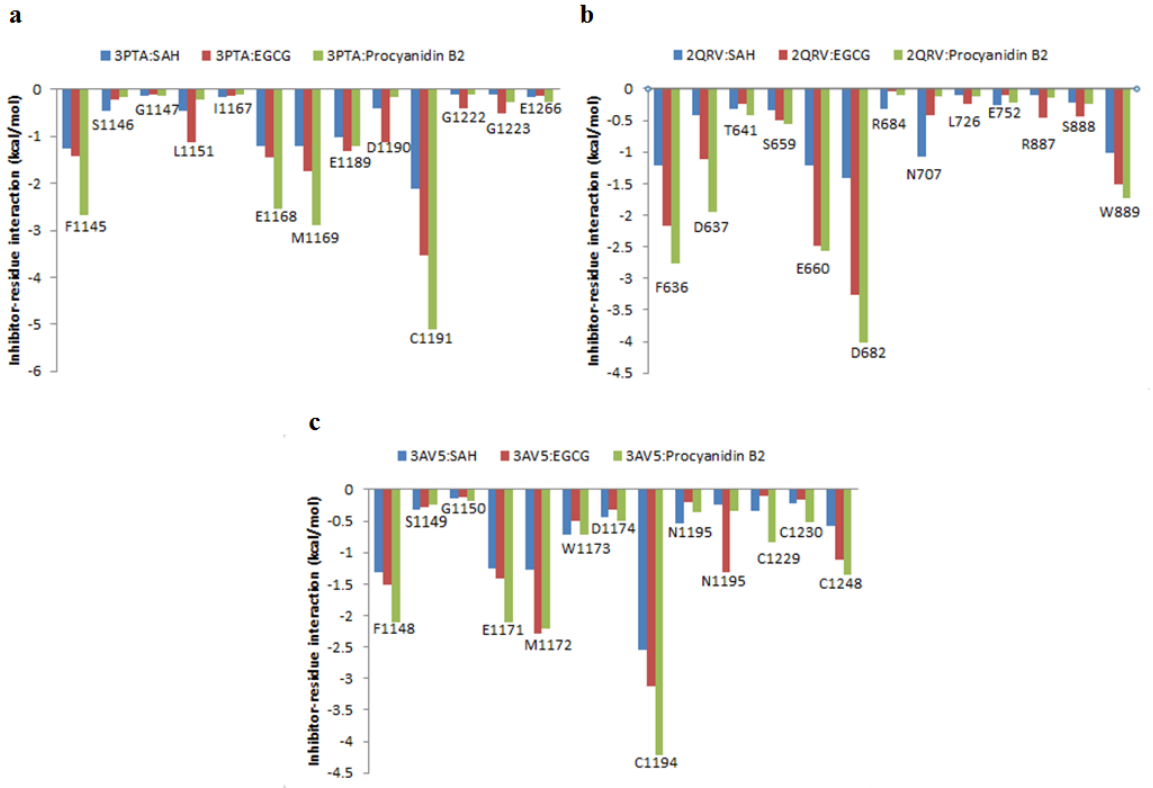

Figure 4.11 Decomposition of ΔG on a per-residue basis for the SAH (blue), EGCG (brown) and procyanidin B2 (green) at the binding pocket of (a) 3PTA (b) 2QRV and (c) 3AV5.

4.3.7 Effect of EGCG and procyanidin B2 on DNMTs activity

The inhibitors, EGCG and procyanidin B2, were used to construct a dose-response curve in terms of IC50 value at varying concentration of drugs. The nuclear extracts from MDA-MB-231 cells were incubated with increasing concentrations of EGCG and procyanidin B2 (1, 2.5, 5, 10 and 15 µM). We observed dose-dependent growth inhibition of DNMTs on incubation with various concentration EGCG and procyanidin B2. The IC_{50} was determined under identical assay conditions. The IC50 of EGCG is identified to be 9.36±1.02 µM while, that of procyanidin B2 is 6.88±0.64 µM [Figure 4.12]. Thus, from above analysis it can be concluded that procyanidin B2 is effective successfully inhibiting

DNMTs at a lower dose of drug concentration as compared to EGCG. The difference in IC50 between EGCG and procyanidin B2 was identified to be significant (n=3, mean ± S.D.). The p-value was found to be significant at 0.05 (p=0.023).

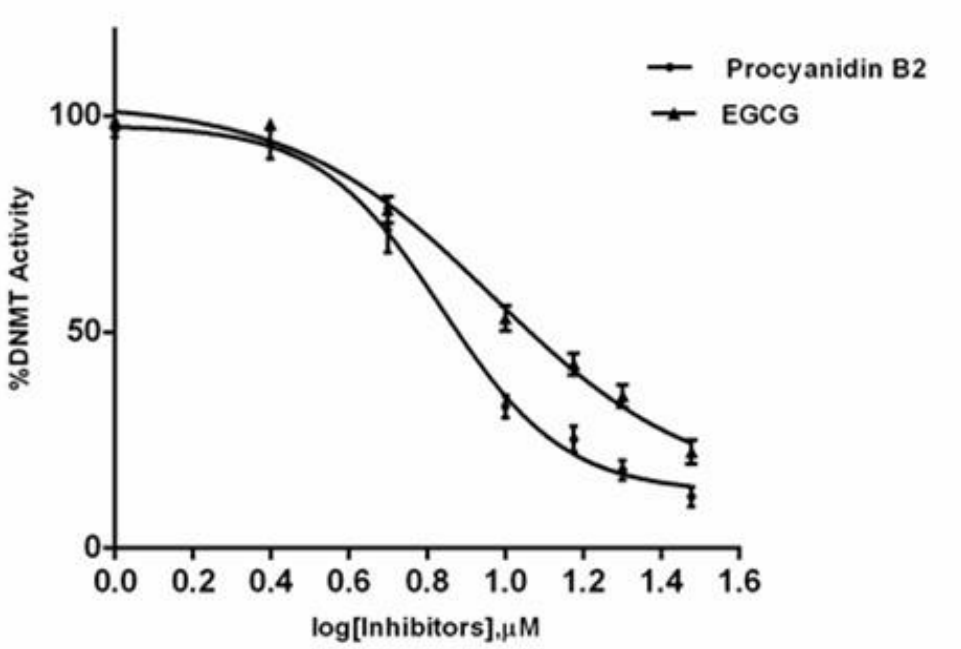

Figure 4.12 Depiction of the log of dose-response plot in terms of percentage decrease in DNMT activity against increasing log of the concentration of procyanidin B2 and EGCG. The IC50 of procyanidin B2 and EGCG are found to be 6.88±0.647 μM and 9.36±1.02, respectively. This clearly demonstrates that procyanidin B2 is more active in inhibiting DNMTs. Data are expressed as mean ± S.D., n=3, $p < 0.05$.

4.3.8 Upregulation of DNMT target and DNMTs genes by EGCG and Procyanidin B2

To further validate the DNMT inhibitory activity, we examined the effect of procyanidinB2 and EGCG on the expression of DNMT target genes (E-cadherin, Maspin, and BRCA1) in MDA-MB-231 cells. Our result indicates that the procyanidin B2 treatment more efficiently upregulates the expression of E-cadherin, Maspin, and BRCA1 as compared to EGCG. This apparently reveals that procyanidin B2 inhibition of DNMTs causes upregulation of these genes [Figure 4.13a]. Moreover, the expression of the DNMT genes; DNMT1, DNMT3A, and DNMT3B were also enhanced by treatment with these polyphenols (Figure 4.13a). The mRNA levels of these genes were identified to be significant (n=3, mean ± S.D.) between EGCG and procyanidin B2 treated groups. The p-value was significant at 0.05.

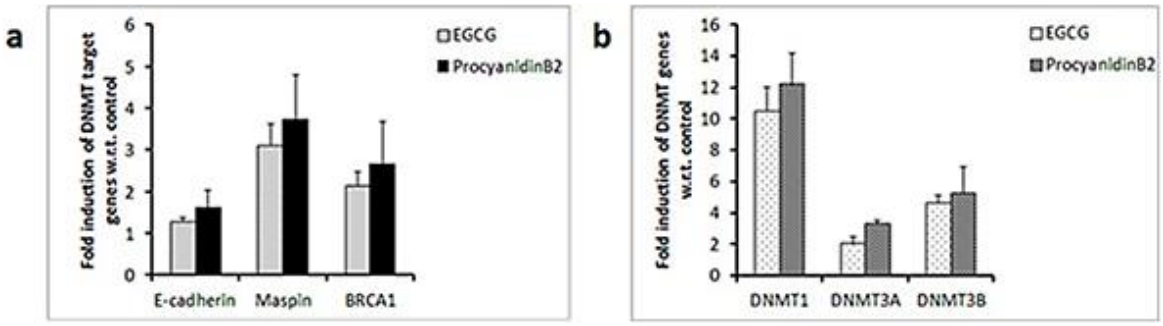

Figure 4.13 Real-time RT-PCR analyses: (a) E-cadherin, Maspin, BRCA1, and (b) DNMT1, DNMT3A, DNMT3B gene expression after treatment with EGCG and procyanidin B2. The mRNA level of both DNMT target and DNMT genes are upregulated more in case of procyanidin B2 than EGCG. Data are expressed as mean ± S.D., n=3, $p < 0.05$.

4.3.9 EGCG and procyanidin B2 are non-toxic for normal cells

The cytotoxicity analysis examined the toxic nature of EGCG and ProcyanidineB2 towards normal keratinocytes (HaCaT) and triple negative breast cancer cells (MDA-MB231) in terms of percentage of cell viability. From the figure 4.14, it is evident that a significant decrease in cell viability was seen with an increasing concentration of EGCG and procyanidin B2 in MDA-MB231 cells. However, these inhibitors did not elicit any lethal effect on normal cells. The sub-lethal concentration (LC_{50}) of EGCG and procyanidin B2 was determined to be 200 and 150µM, respectively, in MDA-MB-231 cells. In contrast, at LC50 of EGCG and procyanidin B2, the HaCaT cells were found to induce no such cytotoxic phenomena. SAH, being an endogenous biochemical product of the methylation reaction has a minimal cytotoxic effect on MDA-MB 231 cells. The difference in LC_{50} between EGCG and procyanidin B2 was identified to be significant (n=3, mean ± S.D.). The p-value was found to be significant at 0.05 (*p=0.01*).

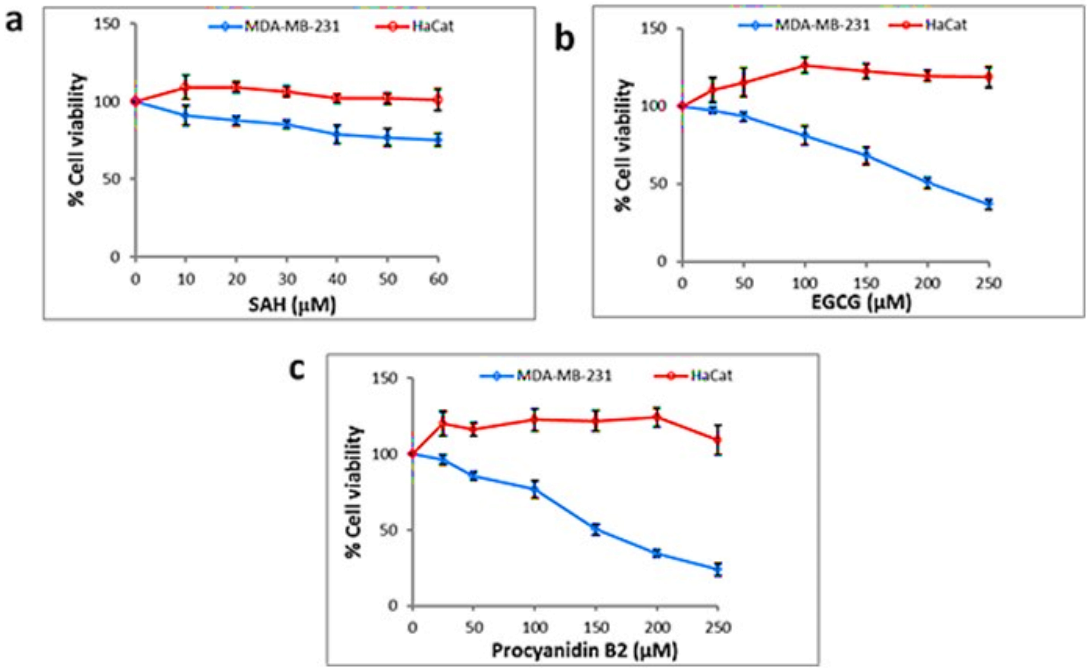

Figure 4.14 Effect of (a) SAH, (b) EGCG and (c) Procyanidin B2 on cell viability and growth. MDA-MB231 and HaCaT cells were treated with indicated concentrations of (a) SAH, (b) EGCG and (c) Procyanidin B2 in µM for 24 h. Cell viability was determined by MTT assay. Data are represented as the mean±SD of three different observations. EGCG and Procyanidin B2 exhibits LC_{50} of 200 µM and 150 µM, respectively, while SAH has been identified having the almost negligible cell growth inhibitory effect of MDA MB 231. However, ECGC and Procyanidin B2 show no cytotoxic effect on HaCaT cells. Data are expressed as mean ± S.D., n=3, $p < 0.05$.

4.4 Discussion

Several studies have documented the apoptosis-inducing effect of polyphenols like catechins and procyanidins indicating their anti-cancer potential and chemotherapeutic application [476, 477]. However, the mystery behind the molecular targets of cell killing effect of these polyphenolic groups or their effect on epigenetic molecular marks, like DNA methylation and manipulators, like DNMTs was not resolved prior to this work. We are acquainted with the fact that DNMTs are responsible for hypermethylation of various genes, imposing a genotoxic effect including tumor suppressor gene during tumor development and cancer progression [478]. Thus, we sought to explore the factors dictating the selectivity of inhibitors binding to DNMT1 and DNMT3A/a. Docking, molecular dynamics simulation, and free energy analyzes were carried out in order to unravel the potential of non-nucleoside inhibitors known till date. We found that the binding affinity of EGCG was highest among all, successfully inhibiting DNMTs in both human and mouse. The docking and simulation analysis illustrate that EGCG-DNMT complex is energetically favored, and the finding is apparently consistent with an in-vitro

analysis. Further ahead, EGCG analogs, mainly polyphenolic groups were screened in quest of identification of novel inhibitors. In this second category, procyanidin B2-3, 3-di-O-gallate moiety presented itself valid, and the most promising inhibitor is having a strong correlation with the active site residues of DNMTs. The selection of procyanidin B2 is of higher negative binding energy and greater selectivity towards DNMTs. Our results demonstrate that binding geometry, driven by hydrogen bond and van-der-Waals energy dominates in the enhancement of total binding energy. Electrostatic interaction occupancy is comparatively feeble along the interface of protein-inhibitor complex. Molecular dynamics simulation of protein-ligand complexes fortified the docking analysis. The stability of the inhibitor inside the binding pocket has been substantiated by total energy, RMSD and RMSF data. Furthermore, the analysis of free energy of binding by MM-PBSA for SAH, EGCG, and procyanidin B2 with DNMTs established that procyanidin B2 has the highest efficacy for the catalytic pocket. Moreover, the detailed thermodynamic description of residue contribution to the free energy of binding affirms the intimate interaction with active site residues.

In addition to the results rationalizing *in-silico* observed selectivity, in vitro experimental analyzes also revealed the potential of procyanidin B2 as an effective inhibitor for diminishing DNMTs activity. The effect of procyanidin B2 in inhibiting DNMTs was evaluated following direct as well as an indirect approach. The indirect approach, procyanidin B2 was directly used as an inhibitor against DNMTs in the nuclear extract of the cells and DNMT activity was noted to be declined with respect to control. Moreover, procyanidin B2 elicits a greater reduction of DNMT activity at a concentration of 6.88±0.647 μM as compared to EGCG. The indirect approach deals with reactivation of DNMT target genes on the application of procyanidin B2. Previously, it has been documented that E-cadherin [479], Maspin [480] and BRCA1 [481] are epigenetically inactivated in breast cancer due to aberrant cytosine methylation in their promoter regions. Evidence also report that inhibition of DNA methylation by 5-Aza-2-deoxycytidine (AZA), could restore the expression the E-cadherin and Maspin (gene) in this cell line [482]. Moreover, in prostate cancer cells it has been reported that BRCA1 can be reactivated by treatment with AZA. Based on this fact, we sought to examine whether EGCG and procyanidin B2 could be able to restore the expression of these DNMT target genes as that of AZA. Corroborating our in silico analysis and in vitro DNMT activity inhibition measurement, we found that both EGCG and procyanidin B2 lead to reactivation of E-cadherin, Maspin and BRCA1 gene at the transcription level. Of note, procyanidin B2 could more efficiently upregulate the DNMT target gene expression relative to EGCG. This might be due to more affinity of procyanidin B2 for DNMTs, thus inhibiting DNMTs to a greater extent than EGCG. However, this enzymatic inhibition impels the elevated expression of DNMTs. Consequently, among the non-nucleoside

inhibitors procyanidin B2 can be considered to be more effective in reducing the DNMT activity and hence can be used to decrease the methylation level.

One of the foremost concerns for clinical application of any anticancer drug relies on its clinical toxicity. So, after searching a potential DNMT inhibitor as procyanidin B2, our next attempt was to examine its cytotoxic nature towards normal cells in contrast to breast cancer cells. Both EGCG and procyanidinB2 were found to elicit extensive cytotoxic effect in highly invasive triple-negative MDA-MB-231 breast cancer cells. While EGCG reduced 50% cell viability at 200 μM concentration, procyanidin B2 was effective at a comparatively lower concentration of 150 μM for 24 h. Conversely, the same LC_{50} of EGCG and procyanidin B2 treatment exhibited no cytotoxic activity to normal keratinocytes (HaCaT).

In contrast to conventional chemotherapeutic drugs, procyanidin B2 is non-toxic in nature. Additionally, it is natural and a dietary component with substantial anticancer effects on breast cancer cells. In conclusion, we have unraveled the role of procyanidin B2 as an epigenetic modulator which precisely targets DNMTs and reverses the silencing of tumor suppressor genes. The outcome of this investigation holds procyanidin B2 as a promising inhibitor against cancer targeting the enzyme DNMTs. Further, by executing experiments in animal models and clinical settings, it may be recommended for incorporation in the list of compounds in chemoprevention of breast cancer.

Chapter 5

Conclusions

The occurrence of inter/intra-tumor heterogeneity and clonal evolution in breast cancer requires accurate elucidation of geographical and chronological variations in patient samples. Despite the success of GWAS in identifying loci associated with tumor initiation, there is still a substantial proportion of the causality to be explored. Interestingly, EWAS have the potential to capture disease-associated epigenetic variations, primarily differential methylation. In this thesis, we have integrated genotype-epitype dataset to identify haplotype-specific DNA methylation in breast cancer, subsequently excavated locus specific diagnostic and prognostic marker. The biomarker identification was conjointly associated with sequential therapeutic strategies for identifying novel drugs, to achieve low dose, the customized and high-impact treatment we seek.

Considering the conjoint study based on genotype-epitype interactions, chapter 2, details about the comparative and comprehensive study of risk alleles in breast cancer and matched normal tissues. The identification of risk allele supports the potential implementation of meQTLs as a risk factor in cancer, wherein DNA methylation functions as a mediator for the respective risk allele. Likewise, risk associated with a polymorphism display germline variant in the cancer tissue and the matched normal counterparts. Mutated genes are inherited in breast cancer and are a well-defined example of breast cancer susceptibility. The increased predisposition of germline mutations in breast cancer tissues is more susceptible for inheritance. The conjoint study of genotype-epitype association has enlightened a novel approach to elucidate the genome-wide distribution of differentially methylated loci. Based on the integrated study, three significant CpG-SNP pairs have been identified that clearly demonstrates variability in the distribution of polymorphic allele is linked to differential methylation pattern which in turn is associated with the gene expression. The fluctuation in major and minor allele distribution associated with SNPs rs9891975, rs4421026 and rs17235834 regulates the methylation level of CpG sites SNP cg02058408, cg05388880 and cg25198340, respectively in tumor and normal samples. These CpG sites lead to differential gene expression of *ST5*, *CMAH* and *FYN* genes, respectively in breast cancer patients in comparison to normal population. Thus, findings based upon novel mechanism constituting genetic variation, DNA methylation and gene expression may serve as novel biomarkers for early diagnosis.

Owing to the identification of diagnostic marker, the next step was to detect the prognostic potential of the biomarkers in determining the overall risk associated with the survival. In chapter 3, the detailed analysis based upon meQTLs is integrated with clinicopathological factors to detect the risk related to overall survival of breast cancer patients. Unlike GWAS, environmental factors directly confound on EWAS, affecting both epigenotype and phenotype and exaggerating the risk associated with the progression of breast cancer. Indeed, when DNA methylation is integrated with SNP array data, it gives a more appropriate clue in understanding principle coordinates of both genetic and epigenetic states in dissecting the risk associated with overall survival. In our study, we have investigated the genome-wide distribution of meQTLs and their cumulative effect on risk stratification of breast cancer patients. The Cox proportional hazard model based on multiple covariates provides an empirical estimate of overall risk. The comprehensive assessment based on meQTLs depicts that variable genotype associated with particular SNP results in differential methylation distribution. These differentially methylated CpG sites have been identified in delaminate the breast cancer patients into the high and low-risk group. In particular, the quantification in methylation level was observed at CpG sites, cg05370838, cg00956490 and cg11340537. These differentially methylated regions were the consequence of discrepancy in allelic distribution associated rs2230576, rs940453, and rs2640785 SNPs, respectively. Furthermore the differentially methylated CpGs were strongly associated with the expression of *ADAM8*, *CREB5*, and *EXPH5* gene, respectively. These differentially methylated CpGs were identified to have a promising association with tumor progression and overall survival of breast cancer patients. Besides, the exclusive effects of SNPs were also interrogated to assess the risk of cancer progression. In summary, conjoint analysis based upon differential methylated CpG sites and SNPs have resulted in the identification of novel susceptible loci that holds prognostic relevance in breast cancer. Further ahead, the functional studies on the candidate genes are required to explicate their potential relevance to the pathophysiology and treatment efficacy of breast cancer.

Tumors displaying global differential methylation hold the benefit of the restoration of these global patterns. Considering their dynamic and reversible nature, the modifying enzymes need to be targeted by small molecules or inhibitors, adding to the drug arsenal to improve their specificity and reduce their toxicity. In chapter 4, we elucidate combinatorial approach of *in-silico* and *in-vitro* analysis in the development of personalized medicine therapies. DNMT, the key epigenetic manipulator, was targeted for pharmacological inhibition and cancer reprogramming. DNMT inhibitors known till date were excavated and examined in-lieu of identification of novel and potent inhibitor. EGCG, being efficient of all had its own limitations. However, on analysis of 32 EGCG analogues, procyanidin B2-3, 3'-di-O-gallate (procyanidin B2) emerged as potent

inhibitor attenuating DNMT activity at IC_{50} of 6.88±0.647 µM and successfully restoring the expression of *E-cadherin*, Maspin and *BRCA1* tumor suppressor genes. Moreover, the toxic property of procyanidin B2 has the ability to discern the triple negative breast cancer (MDA-MB231) cells to normal cells. In summary, the identified epigenetic modulator will have considerable clinical effect in remodeling the malignant cells, and will hold a prime position in breast cancer therapy.

Finally we would like to conclude by saying that the conjoint analysis based upon genetic and epigenetic marker will enlighten the researcher and clinicians to design new strategy in resolving the complexity associated with diagnosis, prognosis and therapeutic implications of breast cancer treatment.

Scope for further research

Integrative studies of genotype-epitype interactions have provided tantalizing insight into the global distribution of meQTLs; however, the significant details need to be evaluated. For instance, there are many questions that need to be answered regarding the genomic architecture of meQTLs (e.g. exact number of loci having differentially methylated regions, how far are the DMRs extended across the functional loci or have positional biasness). The distal or the *trans* effect of these meQTLs on gene expression needs to be evaluated. Moreover, these DMRs are closely linked to another epigenetic process such as histone modifications and non-coding RNA, needs to be interrogated conjointly. Genetically and stochastically driven DMRs holds functional significance and requires systematic investigation across tissues and cell types. Moreover, beside the significant influence of SNP, other genetic variants such as insertion, deletion, duplication and copy number variations need to be incorporated to resolve the complexity associated with disease etiology.

High-throughput technology is characterized by next generation sequencing aids in genome-wide methylome-profiling. Nowadays, it is feasible to map allelic polymorphism associated with DNA methylation at a single base-pair resolution as it provides detailed information about the extent and location of meQTLs. However, technical limitation related to bisulphite sequencing to determine epialleles and epi-haplotypes in the genome is still constrained. It has been identified that sequencing based upon single molecule resolution is required for high epi-allelic and quantitative information. The establishment of the repository and electronic access to samples and associated data will enable to dissect the specific sequence polymorphism, specifically in cancer patients. Finally, appropriate designing of EWAS database needs to developed and conducted, to enable the tools for analysis and interpretation of EWAS data. To achieve clear insight into the exact mechanism of initiation and propagation of tumors condition, cooperation between scientist, clinicians and resource provider is required to pioneer the

conjoint study of GWAS and EWAS. The recognition of the epigenome-wide study of differential methylation has opened new avenues for drug discovery and therapeutics. Therapeutic implications could be combined with conventional therapies to develop personalized treatment and render unresponsive tumors susceptible to treatment at reduced dose. Such advancement may restrict the side effects of treatment and will improve the compliance associated with dose regimens and overall quality of life

Bibliography

1. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
2. Sharma, S., T.K. Kelly, and P.A. Jones, *Epigenetics in cancer.* Carcinogenesis, 2010. **31**(1): p. 27-36.
3. Jones, P.A. and D. Takai, *The role of DNA methylation in mammalian epigenetics.* Science, 2001. **293**(5532): p. 1068-70.
4. Doerfler, W., *DNA methylation and gene activity.* Annu Rev Biochem, 1983. **52**: p. 93-124.
5. Santi, D.V., C.E. Garrett, and P.J. Barr, *On the mechanism of inhibition of DNA-cytosine methyltransferases by cytosine analogs.* Cell, 1983. **33**(1): p. 9-10.
6. Bird, A., et al., *A fraction of the mouse genome that is derived from islands of nonmethylated, CpG-rich DNA.* Cell, 1985. **40**(1): p. 91-9.
7. Antequera, F. and A. Bird, *Number of CpG islands and genes in human and mouse.* Proc Natl Acad Sci U S A, 1993. **90**(24): p. 11995-9.
8. Yoder, J.A., C.P. Walsh, and T.H. Bestor, *Cytosine methylation and the ecology of intragenomic parasites.* Trends Genet, 1997. **13**(8): p. 335-40.
9. O'Neill, R.J., M.J. O'Neill, and J.A. Graves, *Undermethylation associated with retroelement activation and chromosome remodelling in an interspecific mammalian hybrid.* Nature, 1998. **393**(6680): p. 68-72.
10. Bird, A.P., *Gene number, noise reduction and biological complexity.* Trends Genet, 1995. **11**(3): p. 94-100.
11. Lei, H., et al., *De novo DNA cytosine methyltransferase activities in mouse embryonic stem cells.* Development, 1996. **122**(10): p. 3195-205.
12. Patra, S.K., M. Deb, and A. Patra, *Molecular marks for epigenetic identification of developmental and cancer stem cells.* Clin Epigenetics, 2011. **2**(1): p. 27-53.
13. Jaenisch, R., *DNA methylation and imprinting: why bother?* Trends Genet, 1997. **13**(8): p. 323-9.
14. Bartolomei, M.S. and S.M. Tilghman, *Genomic imprinting in mammals.* Annu Rev Genet, 1997. **31**: p. 493-525.
15. Moore, T. and D. Haig, *Genomic imprinting in mammalian development: a parental tug-of-war.* Trends Genet, 1991. **7**(2): p. 45-9.
16. Goto, T. and M. Monk, *Regulation of X-chromosome inactivation in development in mice and humans.* Microbiol Mol Biol Rev, 1998. **62**(2): p. 362-78.
17. Beard, C., E. Li, and R. Jaenisch, *Loss of methylation activates Xist in somatic but not in embryonic cells.* Genes Dev, 1995. **9**(19): p. 2325-34.
18. Reither, S., et al., *Catalytic mechanism of DNA-(cytosine-C5)-methyltransferases revisited: covalent intermediate formation is not essential for methyl group transfer by the murine Dnmt3a enzyme.* J Mol Biol, 2003. **329**(4): p. 675-84.
19. Jeltsch, A., W. Nellen, and F. Lyko, *Two substrates are better than one: dual specificities for Dnmt2 methyltransferases.* Trends Biochem Sci, 2006. **31**(6): p. 306-8.
20. Patra, S.K. and S. Bettuzzi, *Epigenetic DNA-(cytosine-5-carbon) modifications: 5-aza-2'-deoxycytidine and DNA-demethylation.* Biochemistry (Mosc), 2009. **74**(6): p. 613-9.

21. Santi, D.V., A. Norment, and C.E. Garrett, *Covalent bond formation between a DNA-cytosine methyltransferase and DNA containing 5-azacytosine.* Proc Natl Acad Sci U S A, 1984. **81**(22): p. 6993-7.
22. Svedruzic, Z.M. and N.O. Reich, *The mechanism of target base attack in DNA cytosine carbon 5 methylation.* Biochemistry, 2004. **43**(36): p. 11460-73.
23. O'Gara, M., et al., *Enzymatic C5-cytosine methylation of DNA: mechanistic implications of new crystal structures for HhaL methyltransferase-DNA-AdoHcy complexes.* J Mol Biol, 1996. **261**(5): p. 634-45.
24. Zhang, X. and T.C. Bruice, *The mechanism of M.HhaI DNA C5 cytosine methyltransferase enzyme: a quantum mechanics/molecular mechanics approach.* Proc Natl Acad Sci U S A, 2006. **103**(16): p. 6148-53.
25. Vilkaitis, G., et al., *The mechanism of DNA cytosine-5 methylation. Kinetic and mutational dissection of Hhai methyltransferase.* J Biol Chem, 2001. **276**(24): p. 20924-34.
26. Cheng, X., *Structure and function of DNA methyltransferases.* Annu Rev Biophys Biomol Struct, 1995. **24**: p. 293-318.
27. Goll, M.G. and T.H. Bestor, *Eukaryotic cytosine methyltransferases.* Annu Rev Biochem, 2005. **74**: p. 481-514.
28. Goyal, R., R. Reinhardt, and A. Jeltsch, *Accuracy of DNA methylation pattern preservation by the Dnmt1 methyltransferase.* Nucleic Acids Res, 2006. **34**(4): p. 1182-8.
29. Okano, M., et al., *DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development.* Cell, 1999. **99**(3): p. 247-57.
30. Pradhan, S., et al., *Recombinant human DNA (cytosine-5) methyltransferase. I. Expression, purification, and comparison of de novo and maintenance methylation.* J Biol Chem, 1999. **274**(46): p. 33002-10.
31. Lopatina, N., et al., *Differential maintenance and de novo methylating activity by three DNA methyltransferases in aging and immortalized fibroblasts.* J Cell Biochem, 2002. **84**(2): p. 324-34.
32. Bachman, K.E., M.R. Rountree, and S.B. Baylin, *Dnmt3a and Dnmt3b are transcriptional repressors that exhibit unique localization properties to heterochromatin.* J Biol Chem, 2001. **276**(34): p. 32282-7.
33. Bestor, T.H. and G.L. Verdine, *DNA methyltransferases.* Curr Opin Cell Biol, 1994. **6**(3): p. 380-9.
34. Adams, R.L., *Eukaryotic DNA methyltransferases--structure and function.* Bioessays, 1995. **17**(2): p. 139-45.
35. Malone, T., R.M. Blumenthal, and X. Cheng, *Structure-guided analysis reveals nine sequence motifs conserved among DNA amino-methyltransferases, and suggests a catalytic mechanism for these enzymes.* J Mol Biol, 1995. **253**(4): p. 618-32.
36. Kar, S., et al., *An insight into the various regulatory mechanisms modulating human DNA methyltransferase 1 stability and function.* Epigenetics, 2012. **7**(9): p. 994-1007.
37. Cheng, X. and R.M. Blumenthal, *Mammalian DNA methyltransferases: a structural perspective.* Structure, 2008. **16**(3): p. 341-50.
38. Goll, M.G., et al., *Methylation of tRNAAsp by the DNA methyltransferase homolog Dnmt2.* Science, 2006. **311**(5759): p. 395-8.

39. Cheng, X., et al., *Crystal structure of the HhaI DNA methyltransferase complexed with S-adenosyl-L-methionine.* Cell, 1993. **74**(2): p. 299-307.
40. Bestor, T.H., *The DNA methyltransferases of mammals.* Hum Mol Genet, 2000. **9**(16): p. 2395-402.
41. Pradhan, M., et al., *CXXC domain of human DNMT1 is essential for enzymatic activity.* Biochemistry, 2008. **47**(38): p. 10000-9.
42. Aapola, U., et al., *Epigenetic modifications affect Dnmt3L expression.* Biochem J, 2004. **380**(Pt 3): p. 705-13.
43. Kunert, N., et al., *A Dnmt2-like protein mediates DNA methylation in Drosophila.* Development, 2003. **130**(21): p. 5083-90.
44. Widschwendter, M. and P.A. Jones, *DNA methylation and breast carcinogenesis.* Oncogene, 2002. **21**(35): p. 5462-82.
45. Parbin, S., et al., *Histone deacetylases: a saga of perturbed acetylation homeostasis in cancer.* J Histochem Cytochem, 2014. **62**(1): p. 11-33.
46. Smith, L.T., G.A. Otterson, and C. Plass, *Unraveling the epigenetic code of cancer for therapy.* Trends Genet, 2007. **23**(9): p. 449-56.
47. Kanai, Y., *Genome-wide DNA methylation profiles in precancerous conditions and cancers.* Cancer Sci, 2010. **101**(1): p. 36-45.
48. Esteller, M., *Epigenetic gene silencing in cancer: the DNA hypermethylome.* Hum Mol Genet, 2007. **16 Spec No 1**: p. R50-9.
49. Esteller, M., et al., *A gene hypermethylation profile of human cancer.* Cancer Res, 2001. **61**(8): p. 3225-9.
50. Ohm, J.E., et al., *A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing.* Nat Genet, 2007. **39**(2): p. 237-42.
51. Widschwendter, M., et al., *Epigenetic stem cell signature in cancer.* Nat Genet, 2007. **39**(2): p. 157-8.
52. Kar, S., et al., *Epigenetic choreography of stem cells: the DNA demethylation episode of development.* Cell Mol Life Sci, 2014. **71**(6): p. 1017-32.
53. Beck, S. and V.K. Rakyan, *The methylome: approaches for global DNA methylation profiling.* Trends Genet, 2008. **24**(5): p. 231-7.
54. Ushijima, T., *Detection and interpretation of altered methylation patterns in cancer cells.* Nat Rev Cancer, 2005. **5**(3): p. 223-31.
55. Hatada, I., *Emerging technologies for genome-wide DNA methylation profiling in cancer.* Crit Rev Oncog, 2006. **12**(3-4): p. 205-23.
56. Bock, C., *Analysing and interpreting DNA methylation data.* Nat Rev Genet, 2012. **13**(10): p. 705-19.
57. Laird, P.W., *The power and the promise of DNA methylation markers.* Nat Rev Cancer, 2003. **3**(4): p. 253-66.
58. Schones, D.E. and K. Zhao, *Genome-wide approaches to studying chromatin modifications.* Nat Rev Genet, 2008. **9**(3): p. 179-91.
59. Fraga, M.F. and M. Esteller, *DNA methylation: a profile of methods and applications.* Biotechniques, 2002. **33**(3): p. 632, 634, 636-49.
60. Hashimoto, K., et al., *Improved quantification of DNA methylation using methylation-sensitive restriction enzymes and real-time PCR.* Epigenetics, 2007. **2**(2): p. 86-91.

61. Bird, A.P. and E.M. Southern, *Use of restriction enzymes to study eukaryotic DNA methylation: I. The methylation pattern in ribosomal DNA from Xenopus laevis.* J Mol Biol, 1978. **118**(1): p. 27-47.
62. Waalwijk, C. and R.A. Flavell, *DNA methylation at a CCGG sequence in the large intron of the rabbit beta-globin gene: tissue-specific variations.* Nucleic Acids Res, 1978. **5**(12): p. 4631-4.
63. Liang, G., et al., *Identification of DNA methylation differences during tumorigenesis by methylation-sensitive arbitrarily primed polymerase chain reaction.* Methods, 2002. **27**(2): p. 150-5.
64. Frigola, J., et al., *Methylome profiling of cancer cells by amplification of inter-methylated sites (AIMS).* Nucleic Acids Res, 2002. **30**(7): p. e28.
65. Keshet, I., et al., *Evidence for an instructive mechanism of de novo methylation in cancer cells.* Nat Genet, 2006. **38**(2): p. 149-53.
66. Weber, M., et al., *Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome.* Nat Genet, 2007. **39**(4): p. 457-66.
67. Rakyan, V.K., et al., *An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs).* Genome Res, 2008. **18**(9): p. 1518-29.
68. Robinson, M.D., et al., *Evaluation of affinity-based genome-wide DNA methylation data: effects of CpG density, amplification bias, and copy number variation.* Genome Res, 2010. **20**(12): p. 1719-29.
69. Zuo, T., et al., *Methods in DNA methylation profiling.* Epigenomics, 2009. **1**(2): p. 331-45.
70. Hayatsu, H., *Discovery of bisulfite-mediated cytosine conversion to uracil, the key reaction for DNA methylation analysis--a personal account.* Proc Jpn Acad Ser B Phys Biol Sci, 2008. **84**(8): p. 321-30.
71. Wang, R.Y., C.W. Gehrke, and M. Ehrlich, *Comparison of bisulfite modification of 5-methyldeoxycytidine and deoxycytidine residues.* Nucleic Acids Res, 1980. **8**(20): p. 4777-90.
72. Herman, J.G., et al., *Methylation-specific PCR: a novel PCR assay for methylation status of CpG islands.* Proc Natl Acad Sci U S A, 1996. **93**(18): p. 9821-6.
73. Xiong, Z. and P.W. Laird, *COBRA: a sensitive and quantitative DNA methylation assay.* Nucleic Acids Res, 1997. **25**(12): p. 2532-4.
74. Clark, S.J., et al., *High sensitivity mapping of methylated cytosines.* Nucleic Acids Res, 1994. **22**(15): p. 2990-7.
75. Xi, Y. and W. Li, *BSMAP: whole genome bisulfite sequence MAPping program.* BMC Bioinformatics, 2009. **10**: p. 232.
76. Smith, A.D., et al., *Updates to the RMAP short-read mapping software.* Bioinformatics, 2009. **25**(21): p. 2841-2.
77. Xi, Y., et al., *RRBSMAP: a fast, accurate and user-friendly alignment tool for reduced representation bisulfite sequencing.* Bioinformatics, 2012. **28**(3): p. 430-2.
78. Jiang, P., et al., *Methy-Pipe: an integrated bioinformatics pipeline for whole genome bisulfite sequencing data analysis.* PLoS One, 2014. **9**(6): p. e100360.
79. Krueger, F. and S.R. Andrews, *Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications.* Bioinformatics, 2011. **27**(11): p. 1571-2.

80. Pedersen, B., et al., *MethylCoder: software pipeline for bisulfite-treated sequences.* Bioinformatics, 2011. **27**(17): p. 2435-6.
81. Bock, C., et al., *Quantitative comparison of genome-wide DNA methylation mapping technologies.* Nat Biotechnol, 2010. **28**(10): p. 1106-14.
82. Kent, W.J., et al., *BigWig and BigBed: enabling browsing of large distributed datasets.* Bioinformatics, 2010. **26**(17): p. 2204-7.
83. Karolchik, D., et al., *The UCSC Genome Browser Database: 2008 update.* Nucleic Acids Res, 2008. **36**(Database issue): p. D773-9.
84. Flicek, P., et al., *Ensembl 2008.* Nucleic Acids Res, 2008. **36**(Database issue): p. D707-14.
85. Zhou, X., et al., *The Human Epigenome Browser at Washington University.* Nat Methods, 2011. **8**(12): p. 989-90.
86. Anders, S., *Visualization of genomic data with the Hilbert curve.* Bioinformatics, 2009. **25**(10): p. 1231-5.
87. Toedling, J., et al., *Ringo--an R/Bioconductor package for analyzing ChIP-chip readouts.* BMC Bioinformatics, 2007. **8**: p. 221.
88. Barfield, R.T., et al., *CpGassoc: an R function for analysis of DNA methylation microarray data.* Bioinformatics, 2012. **28**(9): p. 1280-1.
89. Withers, B.E. and J.C. Dunbar, *The endonuclease isoschizomers, SmaI and XmaI, bend DNA in opposite orientations.* Nucleic Acids Res, 1993. **21**(11): p. 2571-7.
90. Yan, P.S., S.H. Wei, and T.H. Huang, *Differential methylation hybridization using CpG island arrays.* Methods Mol Biol, 2002. **200**: p. 87-100.
91. Yan, P.S., et al., *Differential methylation hybridization: profiling DNA methylation with a high-density CpG island microarray.* Methods Mol Biol, 2009. **507**: p. 89-106.
92. Balog, R.P., et al., *Parallel assessment of CpG methylation by two-color hybridization with oligonucleotide arrays.* Anal Biochem, 2002. **309**(2): p. 301-10.
93. Fassbender, A., et al., *Quantitative DNA methylation profiling on a high-density oligonucleotide microarray.* Methods Mol Biol, 2010. **576**: p. 155-70.
94. Gagan, J. and E.M. Van Allen, *Next-generation sequencing to guide cancer therapy.* Genome Med, 2015. **7**(1): p. 80.
95. Reuter, J.A., D.V. Spacek, and M.P. Snyder, *High-throughput sequencing technologies.* Mol Cell, 2015. **58**(4): p. 586-97.
96. Mardis, E.R., *Next-generation sequencing platforms.* Annu Rev Anal Chem (Palo Alto Calif), 2013. **6**: p. 287-303.
97. Quail, M.A., et al., *A large genome center's improvements to the Illumina sequencing system.* Nat Methods, 2008. **5**(12): p. 1005-10.
98. Voelkerding, K.V., S.A. Dames, and J.D. Durtschi, *Next-generation sequencing: from basic research to diagnostics.* Clin Chem, 2009. **55**(4): p. 641-58.
99. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.
100. Ding, L., et al., *Analysis of next-generation genomic data in cancer: accomplishments and challenges.* Hum Mol Genet, 2010. **19**(R2): p. R188-96.
101. Cancer Genome Atlas Research, N., et al., *The Cancer Genome Atlas Pan-Cancer analysis project.* Nat Genet, 2013. **45**(10): p. 1113-20.

102. Huang, Y.W., T.H. Huang, and L.S. Wang, *Profiling DNA methylomes from microarray to genome-scale sequencing.* Technol Cancer Res Treat, 2010. **9**(2): p. 139-47.
103. Li, Y. and T.O. Tollefsbol, *DNA methylation detection: bisulfite genomic sequencing analysis.* Methods Mol Biol, 2011. **791**: p. 11-21.
104. Baylin, S.B. and J.G. Herman, *DNA hypermethylation in tumorigenesis: epigenetics joins genetics.* Trends Genet, 2000. **16**(4): p. 168-74.
105. Brown, R. and G. Strathdee, *Epigenomics and epigenetic therapy of cancer.* Trends Mol Med, 2002. **8**(4 Suppl): p. S43-8.
106. Arrowsmith, C.H., et al., *Epigenetic protein families: a new frontier for drug discovery.* Nat Rev Drug Discov, 2012. **11**(5): p. 384-400.
107. Mai, A. and L. Altucci, *Epi-drugs to fight cancer: from chemistry to cancer treatment, the road ahead.* Int J Biochem Cell Biol, 2009. **41**(1): p. 199-213.
108. Feinberg, A.P., R. Ohlsson, and S. Henikoff, *The epigenetic progenitor origin of human cancer.* Nat Rev Genet, 2006. **7**(1): p. 21-33.
109. Kalari, S. and G.P. Pfeifer, *Identification of driver and passenger DNA methylation in cancer by epigenomic analysis.* Adv Genet, 2010. **70**: p. 277-308.
110. Robertson, K.D., et al., *The human DNA methyltransferases (DNMTs) 1, 3a and 3b: coordinate mRNA expression in normal tissues and overexpression in tumors.* Nucleic Acids Res, 1999. **27**(11): p. 2291-8.
111. Sorm, F., et al., *5-Azacytidine, a new, highly effective cancerostatic.* Experientia, 1964. **20**(4): p. 202-3.
112. Vesely, J. and A. Cihak, *5-Aza-2'-deoxycytidine: preclinical studies in mice.* Neoplasma, 1980. **27**(2): p. 113-9.
113. Matei, D., et al., *Epigenetic resensitization to platinum in ovarian cancer.* Cancer Res, 2012. **72**(9): p. 2197-205.
114. Zou, H.Z., et al., *Detection of aberrant p16 methylation in the serum of colorectal cancer patients.* Clin Cancer Res, 2002. **8**(1): p. 188-91.
115. Stresemann, C. and F. Lyko, *Modes of action of the DNA methyltransferase inhibitors azacytidine and decitabine.* Int J Cancer, 2008. **123**(1): p. 8-13.
116. Li, L.H., et al., *Cytotoxicity and mode of action of 5-azacytidine on L1210 leukemia.* Cancer Res, 1970. **30**(11): p. 2760-9.
117. Yang, X., et al., *Targeting DNA methylation for epigenetic therapy.* Trends Pharmacol Sci, 2010. **31**(11): p. 536-46.
118. Godfrey, K.M., et al., *Epigenetic mechanisms and the mismatch concept of the developmental origins of health and disease.* Pediatr Res, 2007. **61**(5 Pt 2): p. 5R-10R.
119. Gluckman, P.D., M.A. Hanson, and T. Buklijas, *A conceptual framework for the developmental origins of health and disease.* J Dev Orig Health Dis, 2010. **1**(1): p. 6-18.
120. Heng, H.H., et al., *Genetic and epigenetic heterogeneity in cancer: the ultimate challenge for drug therapy.* Curr Drug Targets, 2010. **11**(10): p. 1304-16.
121. Ehrlich, M., *DNA methylation and cancer-associated genetic instability.* Adv Exp Med Biol, 2005. **570**: p. 363-92.
122. Tsutsumi, Y., *Hypomethylation of the retrotransposon LINE-1 in malignancy.* Jpn J Clin Oncol, 2000. **30**(7): p. 289-90.

123. Roman-Gomez, J., et al., *Promoter hypomethylation of the LINE-1 retrotransposable elements activates sense/antisense transcription and marks the progression of chronic myeloid leukemia.* Oncogene, 2005. **24**(48): p. 7213-23.
124. Cui, H., et al., *Loss of imprinting in normal tissue of colorectal cancer patients with microsatellite instability.* Nat Med, 1998. **4**(11): p. 1276-80.
125. Cruz-Correa, M., et al., *Loss of imprinting of insulin growth factor II gene: a potential heritable biomarker for colon neoplasia predisposition.* Gastroenterology, 2004. **126**(4): p. 964-70.
126. Ji, W., et al., *DNA demethylation and pericentromeric rearrangements of chromosome 1.* Mutat Res, 1997. **379**(1): p. 33-41.
127. Suter, C.M., D.I. Martin, and R.L. Ward, *Hypomethylation of L1 retrotransposons in colorectal cancer and adjacent normal tissue.* Int J Colorectal Dis, 2004. **19**(2): p. 95-101.
128. Herman, J.G. and S.B. Baylin, *Gene silencing in cancer in association with promoter hypermethylation.* N Engl J Med, 2003. **349**(21): p. 2042-54.
129. Jones, P.A. and S.B. Baylin, *The epigenomics of cancer.* Cell, 2007. **128**(4): p. 683-92.
130. Tao, Z., et al., *Breast Cancer: Epidemiology and Etiology.* Cell Biochem Biophys, 2014.
131. Bocker, W., *[WHO classification of breast tumors and tumors of the female genital organs: pathology and genetics].* Verh Dtsch Ges Pathol, 2002. **86**: p. 116-9.
132. Weigelt, B., J.L. Peterse, and L.J. van 't Veer, *Breast cancer metastasis: markers and models.* Nat Rev Cancer, 2005. **5**(8): p. 591-602.
133. Skibinski, A. and C. Kuperwasser, *The origin of breast tumor heterogeneity.* Oncogene, 2015.
134. Perou, C.M., et al., *Molecular portraits of human breast tumours.* Nature, 2000. **406**(6797): p. 747-52.
135. Sorlie, T., et al., *Repeated observation of breast tumor subtypes in independent gene expression data sets.* Proc Natl Acad Sci U S A, 2003. **100**(14): p. 8418-23.
136. Sorlie, T., et al., *Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms.* BMC Genomics, 2006. **7**: p. 127.
137. Schagdarsurengin, U., G.P. Pfeifer, and R. Dammann, *Frequent epigenetic inactivation of cystatin M in breast carcinoma.* Oncogene, 2007. **26**(21): p. 3089-94.
138. Heyn, H. and M. Esteller, *DNA methylation profiling in the clinic: applications and challenges.* Nat Rev Genet, 2012. **13**(10): p. 679-92.
139. Dedeurwaerder, S., D. Fumagalli, and F. Fuks, *Unravelling the epigenomic dimension of breast cancers.* Curr Opin Oncol, 2011. **23**(6): p. 559-65.
140. Seidman, H., *Cancer of the breast. Statistical and epidemiological data.* Cancer, 1969. **24**(6): p. 1355-78.
141. Ciatto, S., *Detection of breast cancer local recurrences.* Ann Oncol, 1995. **6 Suppl 2**: p. 23-6.
142. Lilienfeld, A.M., *THE EPIDEMIOLOGY OF BREAST CANCER.* Cancer Res, 1963. **23**: p. 1503-13.

143. Friedrich, M., *[X-ray examination of the breast (author's transl)]*. Rontgenblatter, 1981. **34**(4): p. 151-60.
144. Barnes, A.B., *Diagnosis and treatment of abnormal breast secretions.* N Engl J Med, 1966. **275**(21): p. 1184-7.
145. Schlitter, H.E. and H. Burger, *[Mammography. Examination method for tracing and early recognition of mammary carcinoma]*. Med Klin, 1966. **61**(44): p. 1739-43.
146. Moore, F.D., et al., *Carcinoma of the breast. A decade of new results with old concepts.* N Engl J Med, 1967. **277**(7): p. 343-50.
147. Naylor, S., *Biomarkers: current perspectives and future prospects.* Expert Rev Mol Diagn, 2003. **3**(5): p. 525-9.
148. Mayeux, R., *Biomarkers: potential uses and limitations.* NeuroRx, 2004. **1**(2): p. 182-8.
149. Strimbu, K. and J.A. Tavel, *What are biomarkers?* Curr Opin HIV AIDS, 2010. **5**(6): p. 463-6.
150. Strasser-Weippl, K. and P.E. Goss, *Suitable trial designs and cohorts for preventive breast cancer agents.* Nat Rev Clin Oncol, 2013. **10**(12): p. 677-87.
151. Rebbeck, T.R., et al., *Genetic heterogeneity in hereditary breast cancer: role of BRCA1 and BRCA2.* Am J Hum Genet, 1996. **59**(3): p. 547-53.
152. Miki, Y., et al., *A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1.* Science, 1994. **266**(5182): p. 66-71.
153. Wooster, R., et al., *Identification of the breast cancer susceptibility gene BRCA2.* Nature, 1995. **378**(6559): p. 789-92.
154. Zheng, L., et al., *Lessons learned from BRCA1 and BRCA2.* Oncogene, 2000. **19**(53): p. 6159-75.
155. Chen, S. and G. Parmigiani, *Meta-analysis of BRCA1 and BRCA2 penetrance.* J Clin Oncol, 2007. **25**(11): p. 1329-33.
156. Shih, H.A., et al., *BRCA1 and BRCA2 mutation frequency in women evaluated in a breast cancer risk evaluation clinic.* J Clin Oncol, 2002. **20**(4): p. 994-9.
157. Shiovitz, S. and L.A. Korde, *Genetics of breast cancer: a topic in evolution.* Ann Oncol, 2015. **26**(7): p. 1291-9.
158. Nagy, R., K. Sweet, and C. Eng, *Highly penetrant hereditary cancer syndromes.* Oncogene, 2004. **23**(38): p. 6445-70.
159. Masciari, S., et al., *Breast cancer phenotype in women with TP53 germline mutations: a Li-Fraumeni syndrome consortium effort.* Breast Cancer Res Treat, 2012. **133**(3): p. 1125-30.
160. Wilson, J.R., et al., *A novel HER2-positive breast cancer phenotype arising from germline TP53 mutations.* J Med Genet, 2010. **47**(11): p. 771-4.
161. Zhou, X.P., et al., *Germline PTEN promoter mutations and deletions in Cowden/Bannayan-Riley-Ruvalcaba syndrome result in aberrant PTEN protein and dysregulation of the phosphoinositol-3-kinase/Akt pathway.* Am J Hum Genet, 2003. **73**(2): p. 404-11.
162. Brownstein, M.H., M. Wolf, and J.B. Bikowski, *Cowden's disease: a cutaneous marker of breast cancer.* Cancer, 1978. **41**(6): p. 2393-8.
163. Nelen, M.R., et al., *Novel PTEN mutations in patients with Cowden disease: absence of clear genotype-phenotype correlations.* Eur J Hum Genet, 1999. **7**(3): p. 267-73.

164. Stracker, T.H., T. Usui, and J.H. Petrini, *Taking the time to make important decisions: the checkpoint effector kinases Chk1 and Chk2 and the DNA damage response.* DNA Repair (Amst), 2009. **8**(9): p. 1047-54.
165. Adank, M.A., et al., *CHEK2*1100delC homozygosity is associated with a high breast cancer risk in women.* J Med Genet, 2011. **48**(12): p. 860-3.
166. De Brakeleer, S., et al., *Cancer predisposing missense and protein truncating BARD1 mutations in non-BRCA1 or BRCA2 breast cancer families.* Hum Mutat, 2010. **31**(3): p. E1175-85.
167. Xia, B., et al., *Control of BRCA2 cellular and clinical functions by a nuclear partner, PALB2.* Mol Cell, 2006. **22**(6): p. 719-29.
168. Ahmed, M. and N. Rahman, *ATM and breast cancer susceptibility.* Oncogene, 2006. **25**(43): p. 5906-11.
169. Blanco, A., et al., *RAD51C germline mutations found in Spanish site-specific breast cancer and breast-ovarian cancer families.* Breast Cancer Res Treat, 2014. **147**(1): p. 133-43.
170. Rupnik, A., M. Grenon, and N. Lowndes, *The MRN complex.* Curr Biol, 2008. **18**(11): p. R455-7.
171. Barnes, D.R., et al., *Estimating single nucleotide polymorphism associations using pedigree data: applications to breast cancer.* Br J Cancer, 2013. **108**(12): p. 2610-22.
172. Johnson, N., et al., *Counting potentially functional variants in BRCA1, BRCA2 and ATM predicts breast cancer susceptibility.* Hum Mol Genet, 2007. **16**(9): p. 1051-7.
173. Cox, A., et al., *A common coding variant in CASP8 is associated with breast cancer risk.* Nat Genet, 2007. **39**(3): p. 352-8.
174. Stacey, S.N., et al., *Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer.* Nat Genet, 2007. **39**(7): p. 865-9.
175. Zhang, L., et al., *Association of genetic polymorphisms of ER-alpha and the estradiol-synthesizing enzyme genes CYP17 and CYP19 with breast cancer risk in Chinese women.* Breast Cancer Res Treat, 2009. **114**(2): p. 327-38.
176. Abbasi, S., M. Nouri, and C. Azimi, *Estrogen receptor genes variations and breast cancer risk in Iran.* Int J Clin Exp Med, 2012. **5**(4): p. 332-41.
177. Virmani, A.K., et al., *Aberrant methylation of the adenomatous polyposis coli (APC) gene promoter 1A in breast and lung carcinomas.* Clin Cancer Res, 2001. **7**(7): p. 1998-2004.
178. Herman, J.G., et al., *Inactivation of the CDKN2/p16/MTS1 gene is frequently associated with aberrant DNA methylation in all common human cancers.* Cancer Res, 1995. **55**(20): p. 4525-30.
179. Dammann, R., G. Yang, and G.P. Pfeifer, *Hypermethylation of the cpG island of Ras association domain family 1A (RASSF1A), a putative tumor suppressor gene from the 3p21.3 locus, occurs in a large percentage of human breast cancers.* Cancer Res, 2001. **61**(7): p. 3105-9.
180. Evron, E., et al., *Loss of cyclin D2 expression in the majority of breast cancers is associated with promoter hypermethylation.* Cancer Res, 2001. **61**(6): p. 2782-7.
181. Li, B., et al., *CpG methylation as a basis for breast tumor-specific loss of NES1/kallikrein 10 expression.* Cancer Res, 2001. **61**(21): p. 8014-21.

182. Farias, E.F., et al., *Retinoic acid receptor alpha2 is a growth suppressor epigenetically silenced in MCF-7 human breast cancer cells.* Cell Growth Differ, 2002. **13**(8): p. 335-41.
183. Rodriguez, B.A., et al., *Epigenetic repression of the estrogen-regulated Homeobox B13 gene in breast cancer.* Carcinogenesis, 2008. **29**(7): p. 1459-65.
184. Versmold, B., et al., *Epigenetic silencing of the candidate tumor suppressor gene PROX1 in sporadic breast cancer.* Int J Cancer, 2007. **121**(3): p. 547-54.
185. Hesson, L., et al., *NORE1A, a homologue of RASSF1A tumour suppressor gene is inactivated in human cancers.* Oncogene, 2003. **22**(6): p. 947-54.
186. Ferguson, A.T., et al., *High frequency of hypermethylation at the 14-3-3 sigma locus leads to gene silencing in breast cancer.* Proc Natl Acad Sci U S A, 2000. **97**(11): p. 6049-54.
187. Lo, P.K., et al., *Epigenetic suppression of secreted frizzled related protein 1 (SFRP1) expression in human breast cancer.* Cancer Biol Ther, 2006. **5**(3): p. 281-6.
188. Mongan, N.P. and L.J. Gudas, *Valproic acid, in combination with all-trans retinoic acid and 5-aza-2'-deoxycytidine, restores expression of silenced RARbeta2 in breast cancer cells.* Mol Cancer Ther, 2005. **4**(3): p. 477-86.
189. Evron, E., et al., *Detection of breast cancer cells in ductal lavage fluid by methylation-specific PCR.* Lancet, 2001. **357**(9265): p. 1335-6.
190. Krop, I.E., et al., *HIN-1, a putative cytokine highly expressed in normal but not cancerous mammary epithelial cells.* Proc Natl Acad Sci U S A, 2001. **98**(17): p. 9796-801.
191. Silva, J., et al., *Concomitant expression of p16INK4a and p14ARF in primary breast cancer and analysis of inactivation mechanisms.* J Pathol, 2003. **199**(3): p. 289-97.
192. Ai, L., et al., *Inactivation of Wnt inhibitory factor-1 (WIF1) expression by epigenetic silencing is a common event in breast cancer.* Carcinogenesis, 2006. **27**(7): p. 1341-8.
193. Early Breast Cancer Trialists' Collaborative, G., *Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials.* Lancet, 2005. **365**(9472): p. 1687-717.
194. Koscielny, S., et al., *Breast cancer: relationship between the size of the primary tumour and the probability of metastatic dissemination.* Br J Cancer, 1984. **49**(6): p. 709-15.
195. Borst, M.J. and J.A. Ingold, *Metastatic patterns of invasive lobular versus invasive ductal carcinoma of the breast.* Surgery, 1993. **114**(4): p. 637-41; discussion 641-2.
196. Arpino, G., et al., *Infiltrating lobular carcinoma of the breast: tumor characteristics and clinical outcome.* Breast Cancer Res, 2004. **6**(3): p. R149-56.
197. Slamon, D.J., et al., *Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2.* N Engl J Med, 2001. **344**(11): p. 783-92.
198. Ross, J.S., et al., *The Her-2/neu gene and protein in breast cancer 2003: biomarker and target of therapy.* Oncologist, 2003. **8**(4): p. 307-25.

199. Elston, C.W. and I.O. Ellis, *Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up.* Histopathology, 1991. **19**(5): p. 403-10.
200. Page, D.L., *Prognosis and breast cancer. Recognition of lethal and favorable prognostic types.* Am J Surg Pathol, 1991. **15**(4): p. 334-49.
201. Carter, C.L., C. Allen, and D.E. Henson, *Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases.* Cancer, 1989. **63**(1): p. 181-7.
202. Pinder, S.E., et al., *Pathological prognostic factors in breast cancer. III. Vascular invasion: relationship with recurrence and survival in a large study with long-term follow-up.* Histopathology, 1994. **24**(1): p. 41-7.
203. Harbeck, N., et al., *Enhanced benefit from adjuvant chemotherapy in breast cancer patients classified high-risk according to urokinase-type plasminogen activator (uPA) and plasminogen activator inhibitor type 1 (n = 3424).* Cancer Res, 2002. **62**(16): p. 4617-22.
204. Slamon, D.J., et al., *Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene.* Science, 1987. **235**(4785): p. 177-82.
205. Sorlie, T., et al., *Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.* Proc Natl Acad Sci U S A, 2001. **98**(19): p. 10869-74.
206. van 't Veer, L.J., et al., *Gene expression profiling predicts clinical outcome of breast cancer.* Nature, 2002. **415**(6871): p. 530-6.
207. Egeblad, M. and Z. Werb, *New functions for the matrix metalloproteinases in cancer progression.* Nat Rev Cancer, 2002. **2**(3): p. 161-74.
208. Ma, X.J., et al., *A two-gene expression ratio predicts clinical outcome in breast cancer patients treated with tamoxifen.* Cancer Cell, 2004. **5**(6): p. 607-16.
209. Campa, D., et al., *Genetic risk variants associated with in situ breast cancer.* Breast Cancer Res, 2015. **17**: p. 82.
210. Lindstrom, S., et al., *Genome-wide association study identifies multiple loci associated with both mammographic density and breast cancer risk.* Nat Commun, 2014. **5**: p. 5303.
211. Gayther, S.A. and B.A. Ponder, *Mutations of the BRCA1 and BRCA2 genes and the possibilities for predictive testing.* Mol Med Today, 1997. **3**(4): p. 168-74.
212. Dodova, R.I., et al., *Spectrum and frequencies of BRCA1/2 mutations in Bulgarian high risk breast cancer patients.* BMC Cancer, 2015. **15**(1): p. 523.
213. Eccles, D.M., et al., *BRCA1 and BRCA2 genetic testing-pitfalls and recommendations for managing variants of uncertain clinical significance.* Ann Oncol, 2015.
214. Malkin, D., et al., *Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms.* Science, 1990. **250**(4985): p. 1233-8.
215. Liaw, D., et al., *Germline mutations of the PTEN gene in Cowden disease, an inherited breast and thyroid cancer syndrome.* Nat Genet, 1997. **16**(1): p. 64-7.
216. Easton, D.F., et al., *Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium.* Am J Hum Genet, 1993. **52**(4): p. 678-701.

217. Sapkota, Y., *Germline DNA variations in breast cancer predisposition and prognosis: a systematic review of the literature.* Cytogenet Genome Res, 2014. **144**(2): p. 77-91.
218. Easton, D.F., et al., *A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes.* Am J Hum Genet, 2007. **81**(5): p. 873-83.
219. Ghoussaini, M., et al., *Genome-wide association analysis identifies three new breast cancer susceptibility loci.* Nat Genet, 2012. **44**(3): p. 312-8.
220. Michailidou, K., et al., *Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer.* Nat Genet, 2015. **47**(4): p. 373-80.
221. Shivakumar, L., et al., *The RASSF1A tumor suppressor blocks cell cycle progression and inhibits cyclin D1 accumulation.* Mol Cell Biol, 2002. **22**(12): p. 4309-18.
222. Widschwendter, M., et al., *Methylation and silencing of the retinoic acid receptor-beta2 gene in breast cancer.* J Natl Cancer Inst, 2000. **92**(10): p. 826-32.
223. Mehrotra, J., et al., *Very high frequency of hypermethylated genes in breast cancer metastasis to the bone, brain, and lung.* Clin Cancer Res, 2004. **10**(9): p. 3104-9.
224. Sadr-Nabavi, A., et al., *Decreased expression of angiogenesis antagonist EFEMP1 in sporadic breast cancer is caused by aberrant promoter methylation and points to an impact of EFEMP1 as molecular biomarker.* Int J Cancer, 2009. **124**(7): p. 1727-35.
225. Dworkin, A.M., T.H. Huang, and A.E. Toland, *Epigenetic alterations in the breast: Implications for breast cancer detection, prognosis and treatment.* Semin Cancer Biol, 2009. **19**(3): p. 165-71.
226. Fang, F., et al., *Breast cancer methylomes establish an epigenomic foundation for metastasis.* Sci Transl Med, 2011. **3**(75): p. 75ra25.
227. Dedeurwaerder, S., et al., *DNA methylation profiling reveals a predominant immune component in breast cancers.* EMBO Mol Med, 2011. **3**(12): p. 726-41.
228. Conway, K., et al., *DNA methylation profiling in the Carolina Breast Cancer Study defines cancer subclasses differing in clinicopathologic characteristics and survival.* Breast Cancer Res, 2014. **16**(5): p. 450.
229. Stirzaker, C., et al., *Methylome sequencing in triple-negative breast cancer reveals distinct methylation clusters with prognostic value.* Nat Commun, 2015. **6**: p. 5899.
230. Hudis, C.A., *Trastuzumab--mechanism of action and use in clinical practice.* N Engl J Med, 2007. **357**(1): p. 39-51.
231. Baselga, J., et al., *Everolimus in postmenopausal hormone-receptor-positive advanced breast cancer.* N Engl J Med, 2012. **366**(6): p. 520-9.
232. Daroqui, M.C., et al., *TGF-beta autocrine pathway and MAPK signaling promote cell invasiveness and in vivo mammary adenocarcinoma tumor progression.* Oncol Rep, 2012. **28**(2): p. 567-75.
233. Baselga, J., et al., *Phase II randomized study of neoadjuvant everolimus plus letrozole compared with placebo plus letrozole in patients with estrogen receptor-positive breast cancer.* J Clin Oncol, 2009. **27**(16): p. 2630-7.
234. Tate, C.R., et al., *Targeting triple-negative breast cancer cells with the histone deacetylase inhibitor panobinostat.* Breast Cancer Res, 2012. **14**(3): p. R79.

235. Yakes, F.M., et al., *Cabozantinib (XL184), a novel MET and VEGFR2 inhibitor, simultaneously suppresses metastasis, angiogenesis, and tumor growth.* Mol Cancer Ther, 2011. **10**(12): p. 2298-308.
236. Yap, T.A., et al., *Intratumor heterogeneity: seeing the wood for the trees.* Sci Transl Med, 2012. **4**(127): p. 127ps10.
237. Huang, X., et al., *Heterotrimerization of the growth factor receptors erbB2, erbB3, and insulin-like growth factor-i receptor in breast cancer cells resistant to herceptin.* Cancer Res, 2010. **70**(3): p. 1204-14.
238. Rodon, J., et al., *Phase I dose-escalation and -expansion study of buparlisib (BKM120), an oral pan-Class I PI3K inhibitor, in patients with advanced solid tumors.* Invest New Drugs, 2014. **32**(4): p. 670-81.
239. Hassounah, N.B., T.A. Bunch, and K.M. McDermott, *Molecular pathways: the role of primary cilia in cancer progression and therapeutics with a focus on Hedgehog signaling.* Clin Cancer Res, 2012. **18**(9): p. 2429-35.
240. Yu, F., et al., *let-7 regulates self renewal and tumorigenicity of breast cancer cells.* Cell, 2007. **131**(6): p. 1109-23.
241. Tanos, T. and C. Brisken, *What signals operate in the mammary niche?* Breast Dis, 2008. **29**: p. 69-82.
242. Easton, D.F., et al., *Genome-wide association study identifies novel breast cancer susceptibility loci.* Nature, 2007. **447**(7148): p. 1087-93.
243. Mayer, E.L. and I.E. Krop, *Advances in targeting SRC in the treatment of breast cancer and other solid malignancies.* Clin Cancer Res, 2010. **16**(14): p. 3526-32.
244. Shattuck, D.L., et al., *Met receptor contributes to trastuzumab resistance of Her2-overexpressing breast cancer cells.* Cancer Res, 2008. **68**(5): p. 1471-7.
245. Pollak, M., *The insulin and insulin-like growth factor receptor family in neoplasia: an update.* Nat Rev Cancer, 2012. **12**(3): p. 159-69.
246. Andre, F., et al., *Targeting FGFR with dovitinib (TKI258): preclinical and clinical data in breast cancer.* Clin Cancer Res, 2013. **19**(13): p. 3693-702.
247. Barker, H.E., et al., *LOXL2-mediated matrix remodeling in metastasis and mammary gland involution.* Cancer Res, 2011. **71**(5): p. 1561-72.
248. Yamnik, R.L., et al., *S6 kinase 1 regulates estrogen receptor alpha in control of breast cancer cell proliferation.* J Biol Chem, 2009. **284**(10): p. 6361-9.
249. Turner, N., et al., *FGFR1 amplification drives endocrine therapy resistance and is a therapeutic target in breast cancer.* Cancer Res, 2010. **70**(5): p. 2085-94.
250. Hamelers, I.H., et al., *Synergistic proliferative action of insulin-like growth factor I and 17 beta-estradiol in MCF-7S breast tumor cells.* Exp Cell Res, 2002. **273**(1): p. 107-17.
251. Keyomarsi, K., et al., *Cyclin E and survival in patients with breast cancer.* N Engl J Med, 2002. **347**(20): p. 1566-75.
252. Tsai, H.C., et al., *Transient low doses of DNA-demethylating agents exert durable antitumor effects on hematological and epithelial tumor cells.* Cancer Cell, 2012. **21**(3): p. 430-46.
253. Ginestier, C., et al., *CXCR1 blockade selectively targets human breast cancer stem cells in vitro and in xenografts.* J Clin Invest, 2010. **120**(2): p. 485-97.
254. Gastaldi, S., et al., *Met signaling regulates growth, repopulating potential and basal cell-fate commitment of mammary luminal progenitors: implications for basal-like breast cancer.* Oncogene, 2013. **32**(11): p. 1428-40.

255. Chandarlapaty, S., et al., *AKT inhibition relieves feedback suppression of receptor tyrosine kinase expression and activity.* Cancer Cell, 2011. **19**(1): p. 58-71.
256. Lehmann, B.D., et al., *Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies.* J Clin Invest, 2011. **121**(7): p. 2750-67.
257. Fioretti, F.M., et al., *Revising the role of the androgen receptor in breast cancer.* J Mol Endocrinol, 2014. **52**(3): p. R257-65.
258. Garcia, S., et al., *Overexpression of c-Met and of the transducers PI3K, FAK and JAK in breast carcinomas correlates with shorter survival and neoangiogenesis.* Int J Oncol, 2007. **31**(1): p. 49-58.
259. Diaz-Padilla, I., et al., *A phase II study of single-agent RO4929097, a gamma-secretase inhibitor of Notch signaling, in patients with recurrent platinum-resistant epithelial ovarian cancer: A study of the Princess Margaret, Chicago and California phase II consortia.* Gynecol Oncol, 2015.
260. Chang, C.C., et al., *Leptin-STAT3-G9a signaling promotes obesity-mediated breast cancer progression.* Cancer Res, 2015.
261. Presta, M., et al., *Fibroblast growth factor/fibroblast growth factor receptor system in angiogenesis.* Cytokine Growth Factor Rev, 2005. **16**(2): p. 159-78.
262. Ibrahim, Y.H., et al., *PI3K inhibition impairs BRCA1/2 expression and sensitizes BRCA-proficient triple-negative breast cancer to PARP inhibition.* Cancer Discov, 2012. **2**(11): p. 1036-47.
263. Meiss, F. and R. Zeiser, *Vismodegib.* Recent Results Cancer Res, 2014. **201**: p. 405-17.
264. Hiscox, S., et al., *Chronic exposure to fulvestrant promotes overexpression of the c-Met receptor in breast cancer cells: implications for tumour-stroma interactions.* Endocr Relat Cancer, 2006. **13**(4): p. 1085-99.
265. Rao, R., et al., *Combination of pan-histone deacetylase inhibitor and autophagy inhibitor exerts superior efficacy against triple-negative human breast cancer cells.* Mol Cancer Ther, 2012. **11**(4): p. 973-83.
266. Ishii, T., et al., *Inhibition mechanism exploration of investigational drug TAK-441 as inhibitor against Vismodegib-resistant Smoothened mutant.* Eur J Pharmacol, 2014. **723**: p. 305-13.
267. Tyan, S.W., et al., *Breast cancer cells induce cancer-associated fibroblasts to secrete hepatocyte growth factor to enhance breast tumorigenesis.* PLoS One, 2011. **6**(1): p. e15313.
268. Gurney, A., et al., *Wnt pathway inhibition via the targeting of Frizzled receptors results in decreased growth and tumorigenicity of human tumors.* Proc Natl Acad Sci U S A, 2012. **109**(29): p. 11717-22.
269. Garcia, S., et al., *Poor prognosis in breast carcinomas correlates with increased expression of targetable CD146 and c-Met and with proteomic basal-like phenotype.* Hum Pathol, 2007. **38**(6): p. 830-41.
270. Liu, J., et al., *Targeting Wnt-driven cancer through the inhibition of Porcupine by LGK974.* Proc Natl Acad Sci U S A, 2013. **110**(50): p. 20224-9.
271. Finak, G., et al., *Stromal gene expression predicts clinical outcome in breast cancer.* Nat Med, 2008. **14**(5): p. 518-27.

272. Hall, R.D., J.E. Gray, and A.A. Chiappori, *Beyond the standard of care: a review of novel immunotherapy trials for the treatment of lung cancer.* Cancer Control, 2013. **20**(1): p. 22-31.
273. Van Bergen, T., et al., *The role of LOX and LOXL2 in scar formation after glaucoma surgery.* Invest Ophthalmol Vis Sci, 2013. **54**(8): p. 5788-96.
274. Cheng, C., et al., *Evaluation of treatment response of cilengitide in an experimental model of breast cancer bone metastasis using dynamic PET with 18F-FDG.* Hell J Nucl Med, 2011. **14**(1): p. 15-20.
275. Mendes-Pereira, A.M., et al., *Synthetic lethal targeting of PTEN mutant cells with PARP inhibitors.* EMBO Mol Med, 2009. **1**(6-7): p. 315-22.
276. Bernard-Pierrot, I., et al., *Characterization of the recurrent 8p11-12 amplicon identifies PPAPDC1B, a phosphatase protein, as a new therapeutic target in breast cancer.* Cancer Res, 2008. **68**(17): p. 7165-75.
277. Fabbri, M., et al., *Epigenetic regulation of miRNAs in cancer.* Adv Exp Med Biol, 2013. **754**: p. 137-48.
278. Deb, M., et al., *Chromatin dynamics: H3K4 methylation and H3 variant replacement during development and in cancer.* Cell Mol Life Sci, 2014. **71**(18): p. 3439-63.
279. Raha, P., et al., *Combined histone deacetylase inhibition and tamoxifen induces apoptosis in tamoxifen-resistant breast cancer models, by reversing Bcl-2 overexpression.* Breast Cancer Res, 2015. **17**: p. 26.
280. Beagle, B.R., et al., *mTOR kinase inhibitors synergize with histone deacetylase inhibitors to kill B-cell acute lymphoblastic leukemia cells.* Oncotarget, 2015. **6**(4): p. 2088-100.
281. Parbin, S., et al., *Insights into the molecular interactions of thymoquinone with histone deacetylase: evaluation of the therapeutic intervention potential against breast cancer.* Mol Biosyst, 2016. **12**(1): p. 48-58.
282. Huang, X., et al., *HDAC inhibitor SNDX-275 enhances efficacy of trastuzumab in erbB2-overexpressing breast cancer cells and exhibits potential to overcome trastuzumab resistance.* Cancer Lett, 2011. **307**(1): p. 72-9.
283. Yardley, D.A., et al., *Randomized phase II, double-blind, placebo-controlled study of exemestane with or without entinostat in postmenopausal women with locally recurrent or metastatic estrogen receptor-positive breast cancer progressing on treatment with a nonsteroidal aromatase inhibitor.* J Clin Oncol, 2013. **31**(17): p. 2128-35.
284. Kaufmann, M., et al., *Exemestane is superior to megestrol acetate after tamoxifen failure in postmenopausal women with advanced breast cancer: results of a phase III randomized double-blind trial. The Exemestane Study Group.* J Clin Oncol, 2000. **18**(7): p. 1399-411.
285. Thakur, S., et al., *ING1 and 5-azacytidine act synergistically to block breast cancer cell growth.* PLoS One, 2012. **7**(8): p. e43671.
286. Mirza, S., et al., *Demethylating agent 5-aza-2-deoxycytidine enhances susceptibility of breast cancer cells to anticancer agents.* Mol Cell Biochem, 2010. **342**(1-2): p. 101-9.
287. Kar, S., et al., *Expression profiling of DNA methylation-mediated epigenetic gene-silencing factors in breast cancer.* Clin Epigenetics, 2014. **6**(1): p. 20.

288. Connolly, R. and V. Stearns, *Epigenetics as a therapeutic target in breast cancer.* J Mammary Gland Biol Neoplasia, 2012. **17**(3-4): p. 191-204.
289. Xu, J., et al., *Evidence that tumor necrosis factor-related apoptosis-inducing ligand induction by 5-Aza-2'-deoxycytidine sensitizes human breast cancer cells to adriamycin.* Cancer Res, 2007. **67**(3): p. 1203-11.
290. Peng, L., et al., *SIRT1 deacetylates the DNA methyltransferase 1 (DNMT1) protein and alters its activities.* Mol Cell Biol, 2011. **31**(23): p. 4720-34.
291. Bellarosa, D., et al., *SAHA/Vorinostat induces the expression of the CD137 receptor/ligand system and enhances apoptosis mediated by soluble CD137 receptor in a human breast cancer cell line.* Int J Oncol, 2012. **41**(4): p. 1486-94.
292. Munster, P.N., et al., *A phase II study of the histone deacetylase inhibitor vorinostat combined with tamoxifen for the treatment of patients with hormone therapy-resistant breast cancer.* Br J Cancer, 2011. **104**(12): p. 1828-35.
293. *Entinostat plus exemestane has activity in ER+ advanced breast cancer.* Cancer Discov, 2013. **3**(7): p. OF17.
294. Kelly, W.K., et al., *Phase I study of an oral histone deacetylase inhibitor, suberoylanilide hydroxamic acid, in patients with advanced cancer.* J Clin Oncol, 2005. **23**(17): p. 3923-31.
295. Ramalingam, S.S., et al., *Phase I and pharmacokinetic study of vorinostat, a histone deacetylase inhibitor, in combination with carboplatin and paclitaxel for advanced solid malignancies.* Clin Cancer Res, 2007. **13**(12): p. 3605-10.
296. Lee, F., M.N. Jure-Kunkel, and M.E. Salvati, *Synergistic activity of ixabepilone plus other anticancer agents: preclinical and clinical evidence.* Ther Adv Med Oncol, 2011. **3**(1): p. 11-25.
297. Tu, Y., et al., *A phase I-II study of the histone deacetylase inhibitor vorinostat plus sequential weekly paclitaxel and doxorubicin-cyclophosphamide in locally advanced breast cancer.* Breast Cancer Res Treat, 2014. **146**(1): p. 145-52.
298. Connolly, R.M. and V. Stearns, *Current approaches for neoadjuvant chemotherapy in breast cancer.* Eur J Pharmacol, 2013. **717**(1-3): p. 58-66.
299. Lee, J., et al., *A class I histone deacetylase inhibitor, entinostat, enhances lapatinib efficacy in HER2-overexpressing breast cancer cells through FOXO3-mediated Bim1 expression.* Breast Cancer Res Treat, 2014. **146**(2): p. 259-72.
300. Weiss, A.J., et al., *Phase II study of 5-azacytidine in solid tumors.* Cancer Treat Rep, 1977. **61**(1): p. 55-8.
301. Tomillero, A. and M.A. Moral, *Gateways to clinical trials.* Methods Find Exp Clin Pharmacol, 2008. **30**(7): p. 543-88.
302. Shao, H., et al., *Improved response to nab-paclitaxel compared with cremophor-solubilized paclitaxel is independent of secreted protein acidic and rich in cysteine expression in non-small cell lung cancer.* J Thorac Oncol, 2011. **6**(6): p. 998-1005.
303. Nautiyal, J., et al., *Src inhibitor dasatinib inhibits growth of breast cancer cells by modulating EGFR signaling.* Cancer Lett, 2009. **283**(2): p. 143-51.
304. Campone, M., et al., *Phase II study of single-agent bosutinib, a Src/Abl tyrosine kinase inhibitor, in patients with locally advanced or metastatic breast cancer pretreated with chemotherapy.* Ann Oncol, 2012. **23**(3): p. 610-7.

305. Gucalp, A., et al., *Phase II trial of saracatinib (AZD0530), an oral SRC-inhibitor for the treatment of patients with hormone receptor-negative metastatic breast cancer.* Clin Breast Cancer, 2011. **11**(5): p. 306-11.
306. Britton, D.J., et al., *Bidirectional cross talk between ERalpha and EGFR signalling pathways regulates tamoxifen-resistant growth.* Breast Cancer Res Treat, 2006. **96**(2): p. 131-46.
307. Rizzo, P., et al., *Cross-talk between notch and the estrogen receptor in breast cancer suggests novel therapeutic approaches.* Cancer Res, 2008. **68**(13): p. 5226-35.
308. Phillips, G.D., et al., *Dual targeting of HER2-positive cancer with trastuzumab emtansine and pertuzumab: critical role for neuregulin blockade in antitumor response to combination therapy.* Clin Cancer Res, 2014. **20**(2): p. 456-68.
309. Barretina, J., et al., *The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity.* Nature, 2012. **483**(7391): p. 603-7.
310. Amos, C.I., *Successful design and conduct of genome-wide association studies.* Hum Mol Genet, 2007. **16 Spec No. 2**: p. R220-5.
311. Nica, A.C., et al., *Candidate causal regulatory effects by integration of expression QTLs with complex trait genetic associations.* PLoS Genet, 2010. **6**(4): p. e1000895.
312. Nicolae, D.L., et al., *Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS.* PLoS Genet, 2010. **6**(4): p. e1000888.
313. Tycko, B., *Allele-specific DNA methylation: beyond imprinting.* Hum Mol Genet, 2010. **19**(R2): p. R210-20.
314. Mills, A.A., *Throwing the cancer switch: reciprocal roles of polycomb and trithorax proteins.* Nat Rev Cancer, 2010. **10**(10): p. 669-82.
315. You, J.S. and P.A. Jones, *Cancer genetics and epigenetics: two sides of the same coin?* Cancer Cell, 2012. **22**(1): p. 9-20.
316. Kelly, T.K., D.D. De Carvalho, and P.A. Jones, *Epigenetic modifications as therapeutic targets.* Nat Biotechnol, 2010. **28**(10): p. 1069-1078.
317. Genomes Project, C., et al., *A map of human genome variation from population-scale sequencing.* Nature, 2010. **467**(7319): p. 1061-73.
318. Bae, J.B., *Perspectives of international human epigenome consortium.* Genomics Inform, 2013. **11**(1): p. 7-14.
319. Chadwick, L.H., *The NIH Roadmap Epigenomics Program data resource.* Epigenomics, 2012. **4**(3): p. 317-24.
320. Cipollini, G., et al., *Genetic alterations in hereditary breast cancer.* Ann Oncol, 2004. **15 Suppl 1**: p. I7-I13.
321. *Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease.* Lancet, 2001. **358**(9291): p. 1389-99.
322. DePinho, R.A., *The age of cancer.* Nature, 2000. **408**(6809): p. 248-254.
323. Elenbaas, B. and R.A. Weinberg, *Heterotypic signaling between epithelial tumor cells and fibroblasts in carcinoma formation.* Exp Cell Res, 2001. **264**(1): p. 169-84.

324. Hindorff, L.A., et al., *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.* Proc Natl Acad Sci U S A, 2009. **106**(23): p. 9362-7.
325. Gibbs, J.R., et al., *Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain.* PLoS Genet, 2010. **6**(5): p. e1000952.
326. Bell, J.T., et al., *DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines.* Genome Biol, 2011. **12**(1): p. R10.
327. Robinson, M.D., et al., *Statistical methods for detecting differentially methylated loci and regions.* Front Genet, 2014. **5**: p. 324.
328. Bibikova, M., et al., *High-throughput DNA methylation profiling using universal bead arrays.* Genome Res, 2006. **16**(3): p. 383-93.
329. Di, X., et al., *Dynamic model based algorithms for screening and genotyping over 100 K SNPs on oligonucleotide microarrays.* Bioinformatics, 2005. **21**(9): p. 1958-63.
330. Liu, W.M., et al., *Algorithms for large-scale genotyping microarrays.* Bioinformatics, 2003. **19**(18): p. 2397-403.
331. Zhao, Q., et al., *Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA.* Brief Bioinform, 2014.
332. Tomczak, K., P. Czerwinska, and M. Wiznerowicz, *The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge.* Contemp Oncol (Pozn), 2015. **19**(1A): p. A68-77.
333. Carvalho, B., et al., *Exploration, normalization, and genotype calls of high-density oligonucleotide SNP array data.* Biostatistics, 2007. **8**(2): p. 485-99.
334. Robertson, G., et al., *De novo assembly and analysis of RNA-seq data.* Nat Methods, 2010. **7**(11): p. 909-12.
335. Wang, K., et al., *MapSplice: accurate mapping of RNA-seq reads for splice junction discovery.* Nucleic Acids Res, 2010. **38**(18): p. e178.
336. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC Bioinformatics, 2011. **12**: p. 323.
337. Kauffmann, A., et al., *Importing ArrayExpress datasets into R/Bioconductor.* Bioinformatics, 2009. **25**(16): p. 2092-4.
338. Pinneau, S.R., et al., *Analysis of factor variance: one-way classification.* Percept Mot Skills, 1966. **23**(3): p. 1209-10.
339. Norton, H.W. and J.V. Neel, *Hardy-Weinberg Equilibrium and Primitive Populations.* Am J Hum Genet, 1965. **17**: p. 91-2.
340. Chan, Y. and R.P. Walmsley, *Learning and understanding the Kruskal-Wallis one-way analysis-of-variance-by-ranks test for differences among three or more independent groups.* Phys Ther, 1997. **77**(12): p. 1755-62.
341. Daca-Roszak, P., et al., *Impact of SNPs on methylation readouts by Illumina Infinium HumanMethylation450 BeadChip Array: implications for comparative population studies.* BMC Genomics, 2015. **16**: p. 1003.
342. Heyn, H., et al., *Linkage of DNA methylation quantitative trait loci to human cancer risk.* Cell Rep, 2014. **7**(2): p. 331-8.
343. Bishara, A.J. and J.B. Hittner, *Testing the significance of a correlation with nonnormal data: comparison of Pearson, Spearman, transformation, and resampling approaches.* Psychol Methods, 2012. **17**(3): p. 399-417.

344. Russnes, H.G., et al., *Insight into the heterogeneity of breast cancer through next-generation sequencing.* J Clin Invest, 2011. **121**(10): p. 3810-8.
345. Stingl, J. and C. Caldas, *Molecular heterogeneity of breast carcinomas and the cancer stem cell hypothesis.* Nat Rev Cancer, 2007. **7**(10): p. 791-9.
346. Polyak, K., *Breast cancer: origins and evolution.* J Clin Invest, 2007. **117**(11): p. 3155-63.
347. Greaves, M. and C.C. Maley, *Clonal evolution in cancer.* Nature, 2012. **481**(7381): p. 306-13.
348. Gerlinger, M. and C. Swanton, *How Darwinian models inform therapeutic failure initiated by clonal heterogeneity in cancer medicine.* Br J Cancer, 2010. **103**(8): p. 1139-43.
349. Polyak, K., *Breast cancer stem cells: a case of mistaken identity?* Stem Cell Rev, 2007. **3**(2): p. 107-9.
350. Fletcher, O. and F. Dudbridge, *Candidate gene-environment interactions in breast cancer.* BMC Med, 2014. **12**: p. 195.
351. Margaritte, P., et al., *Linkage of familial breast cancer to chromosome 17q21 may not be restricted to early-onset disease.* Am J Hum Genet, 1992. **50**(6): p. 1231-4.
352. Al-Hajj, M., et al., *Prospective identification of tumorigenic breast cancer cells.* Proc Natl Acad Sci U S A, 2003. **100**(7): p. 3983-8.
353. Chen, J.J., et al., *BRCA1, BRCA2, and Rad51 operate in a common DNA damage response pathway.* Cancer Res, 1999. **59**(7 Suppl): p. 1752s-1756s.
354. Chuang, H.Y., et al., *Network-based classification of breast cancer metastasis.* Mol Syst Biol, 2007. **3**: p. 140.
355. Kar, S.P., et al., *Network-based integration of GWAS and gene expression identifies a HOX-centric network associated with serous ovarian cancer risk.* Cancer Epidemiol Biomarkers Prev, 2015.
356. Darabi, H., et al., *Polymorphisms in a Putative Enhancer at the 10q21.2 Breast Cancer Risk Locus Regulate NRBF2 Expression.* Am J Hum Genet, 2015. **97**(1): p. 22-34.
357. Sirchia, S.M., et al., *Evidence of epigenetic changes affecting the chromatin state of the retinoic acid receptor beta2 promoter in breast cancer cells.* Oncogene, 2000. **19**(12): p. 1556-63.
358. Landberg, G., et al., *Downregulation of the potential suppressor gene IGFBP-rP1 in human breast cancer is associated with inactivation of the retinoblastoma protein, cyclin E overexpression and increased proliferation in estrogen receptor negative tumors.* Oncogene, 2001. **20**(27): p. 3497-505.
359. Teschendorff, A.E. and M. Widschwendter, *Differential variability improves the identification of cancer risk markers in DNA methylation studies profiling precursor cancer lesions.* Bioinformatics, 2012. **28**(11): p. 1487-94.
360. Lichy, J.H., et al., *Identification of a human chromosome 11 gene which is differentially regulated in tumorigenic and nontumorigenic somatic cell hybrids of HeLa cells.* Cell Growth Differ, 1992. **3**(8): p. 541-8.
361. Majidi, M., A.E. Hubbs, and J.H. Lichy, *Activation of extracellular signal-regulated kinase 2 by a novel Abl-binding protein, ST5.* J Biol Chem, 1998. **273**(26): p. 16608-14.
362. Amid, C., et al., *Comparative genomic sequencing reveals a strikingly similar architecture of a conserved syntenic region on human chromosome 11p15.3*

(including gene ST5) and mouse chromosome 7. Cytogenet Cell Genet, 2001. **93**(3-4): p. 284-90.
363. Irie, A., et al., *The molecular basis for the absence of N-glycolylneuraminic acid in humans.* J Biol Chem, 1998. **273**(25): p. 15866-71.
364. Kawano, T., et al., *Molecular cloning of cytidine monophospho-N-acetylneuraminic acid hydroxylase. Regulation of species- and tissue-specific expression of N-glycolylneuraminic acid.* J Biol Chem, 1995. **270**(27): p. 16458-63.
365. Inoue, S., C. Sato, and K. Kitajima, *Extensive enrichment of N-glycolylneuraminic acid in extracellular sialoglycoproteins abundantly synthesized and secreted by human cancer cells.* Glycobiology, 2010. **20**(6): p. 752-62.
366. Gabri, M.R., et al., *Exogenous incorporation of neugc-rich mucin augments n-glycolyl sialic acid content and promotes malignant phenotype in mouse tumor cell lines.* J Exp Clin Cancer Res, 2009. **28**: p. 146.
367. Samraj, A.N., et al., *Involvement of a non-human sialic Acid in human cancer.* Front Oncol, 2014. **4**: p. 33.
368. Zhong, Y., et al., *N-Glycolyl GM3 ganglioside immunoexpression in oral mucosal melanomas of Chinese.* Oral Dis, 2012. **18**(8): p. 741-7.
369. Sorensen, K.D., et al., *Chromosomal deletion, promoter hypermethylation and downregulation of FYN in prostate cancer.* Int J Cancer, 2008. **122**(3): p. 509-19.
370. Yamashita, S., et al., *Chemical genomic screening for methylation-silenced genes in gastric cancer cell lines using 5-aza-2'-deoxycytidine treatment and oligonucleotide microarray.* Cancer Sci, 2006. **97**(1): p. 64-71.
371. van Oosterwijk, J.G., et al., *Src kinases in chondrosarcoma chemoresistance and migration: dasatinib sensitises to doxorubicin in TP53 mutant cells.* Br J Cancer, 2013. **109**(5): p. 1214-22.
372. Kulis, M., et al., *Epigenomic analysis detects widespread gene-body DNA hypomethylation in chronic lymphocytic leukemia.* Nat Genet, 2012. **44**(11): p. 1236-42.
373. Kulis, M., et al., *Intragenic DNA methylation in transcriptional regulation, normal differentiation and cancer.* Biochim Biophys Acta, 2013. **1829**(11): p. 1161-74.
374. Hahn, M.A., et al., *Relationship between gene body DNA methylation and intragenic H3K9me3 and H3K36me3 chromatin marks.* PLoS One, 2011. **6**(4): p. e18844.
375. Lister, R., et al., *Human DNA methylomes at base resolution show widespread epigenomic differences.* Nature, 2009. **462**(7271): p. 315-22.
376. Ball, M.P., et al., *Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells.* Nat Biotechnol, 2009. **27**(4): p. 361-8.
377. Hon, G.C., et al., *Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer.* Genome Res, 2012. **22**(2): p. 246-58.
378. Rauch, T.A., et al., *A human B cell methylome at 100-base pair resolution.* Proc Natl Acad Sci U S A, 2009. **106**(3): p. 671-8.
379. Ali, H.R., et al., *Genome-driven integrated classification of breast cancer validated in over 7,500 samples.* Genome Biol, 2014. **15**(8): p. 431.

380. Haibe-Kains, B., et al., *A three-gene model to robustly identify breast cancer molecular subtypes.* J Natl Cancer Inst, 2012. **104**(4): p. 311-25.
381. Cancer Genome Atlas, N., *Comprehensive molecular portraits of human breast tumours.* Nature, 2012. **490**(7418): p. 61-70.
382. Jovanovic, J., et al., *The epigenetics of breast cancer.* Mol Oncol, 2010. **4**(3): p. 242-54.
383. Ronneberg, J.A., et al., *Methylation profiling with a panel of cancer related genes: association with estrogen receptor, TP53 mutation status and expression subtypes in sporadic breast cancer.* Mol Oncol, 2011. **5**(1): p. 61-76.
384. Fleischer, T., et al., *Genome-wide DNA methylation profiles in progression to in situ and invasive carcinoma of the breast with impact on gene transcription and prognosis.* Genome Biol, 2014. **15**(8): p. 435.
385. Fackler, M.J., et al., *Genome-wide methylation analysis identifies genes specific to breast cancer hormone receptor status and risk of recurrence.* Cancer Res, 2011. **71**(19): p. 6195-207.
386. Gyorffy, B., et al., *Aberrant DNA methylation impacts gene expression and prognosis in breast cancer subtypes.* Int J Cancer, 2015.
387. Schnitt, S.J., *Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy.* Mod Pathol, 2010. **23 Suppl 2**: p. S60-4.
388. Liu, Z., X.S. Zhang, and S. Zhang, *Breast tumor subgroups reveal diverse clinical prognostic power.* Sci Rep, 2014. **4**: p. 4002.
389. Dejeux, E., et al., *DNA methylation profiling in doxorubicin treated primary locally advanced breast tumours identifies novel genes associated with survival and treatment response.* Mol Cancer, 2010. **9**: p. 68.
390. Fleischer, T., et al., *Integrated analysis of high-resolution DNA methylation profiles, gene expression, germline genotypes and clinical end points in breast cancer patients.* Int J Cancer, 2014. **134**(11): p. 2615-25.
391. Kamalakaran, S., et al., *DNA methylation patterns in luminal breast cancers differ from non-luminal subtypes and can identify relapse risk independent of other clinical variables.* Mol Oncol, 2011. **5**(1): p. 77-92.
392. Palmer, J.R., et al., *Genetic susceptibility loci for subtypes of breast cancer in an African American population.* Cancer Epidemiol Biomarkers Prev, 2013. **22**(1): p. 127-34.
393. Pirie, A., et al., *Common germline polymorphisms associated with breast cancer-specific survival.* Breast Cancer Res, 2015. **17**(1): p. 58.
394. Andersson, R., et al., *An atlas of active enhancers across human cell types and tissues.* Nature, 2014. **507**(7493): p. 455-61.
395. Whyte, W.A., et al., *Master transcription factors and mediator establish super-enhancers at key cell identity genes.* Cell, 2013. **153**(2): p. 307-19.
396. Ma, C.X. and M.J. Ellis, *The Cancer Genome Atlas: clinical applications for breast cancer.* Oncology (Williston Park), 2013. **27**(12): p. 1263-9, 1274-9.
397. Costa, V., et al., *RNA-Seq and human complex diseases: recent accomplishments and future perspectives.* Eur J Hum Genet, 2013. **21**(2): p. 134-42.
398. Zhi, D., et al., *SNPs located at CpG sites modulate genome-epigenome interaction.* Epigenetics, 2013. **8**(8): p. 802-6.
399. Gentile, J.R., A.H. Roden, and R.D. Klein, *An analysis-of-variance model for the intrasubject replication design.* J Appl Behav Anal, 1972. **5**(2): p. 193-8.

400. Prentice, R.L. and J.D. Kalbfleisch, *Hazard rate models with covariates.* Biometrics, 1979. **35**(1): p. 25-39.
401. Harrell, F.E., Jr., et al., *Regression modelling strategies for improved prognostic prediction.* Stat Med, 1984. **3**(2): p. 143-52.
402. Bradburn, M.J., et al., *Survival analysis part II: multivariate data analysis--an introduction to concepts and methods.* Br J Cancer, 2003. **89**(3): p. 431-6.
403. Clark, T.G., et al., *Survival analysis part I: basic concepts and first analyses.* Br J Cancer, 2003. **89**(2): p. 232-8.
404. Stablein, D.M., W.H. Carter, Jr., and J.W. Novak, *Analysis of survival data with nonproportional hazard functions.* Control Clin Trials, 1981. **2**(2): p. 149-59.
405. Wang, X.S., et al., *Molecular cloning and characterization of a novel protein kinase with a catalytic domain homologous to mitogen-activated protein kinase kinase kinase.* J Biol Chem, 1996. **271**(49): p. 31607-11.
406. Zardavas, D., T.M. Fouad, and M. Piccart, *Optimal adjuvant treatment for patients with HER2-positive breast cancer in 2015.* Breast, 2015.
407. Murray, J.L., et al., *Prognostic value of single nucleotide polymorphisms of candidate genes associated with inflammation in early stage breast cancer.* Breast Cancer Res Treat, 2013. **138**(3): p. 917-24.
408. Cheng, Q., et al., *Amplification and high-level expression of heat shock protein 90 marks aggressive phenotypes of human epidermal growth factor receptor 2 negative breast cancer.* Breast Cancer Res, 2012. **14**(2): p. R62.
409. Li, H., et al., *Functional annotation of HOT regions in the human genome: implications for human disease and cancer.* Sci Rep, 2015. **5**: p. 11633.
410. Fletcher, M.N., et al., *Master regulators of FGFR2 signalling and breast cancer risk.* Nat Commun, 2013. **4**: p. 2464.
411. Kim, H.S., J.D. Minna, and M.A. White, *GWAS meets TCGA to illuminate mechanisms of cancer predisposition.* Cell, 2013. **152**(3): p. 387-9.
412. Taqi, M.M., et al., *Prodynorphin CpG-SNPs associated with alcohol dependence: elevated methylation in the brain of human alcoholics.* Addict Biol, 2011. **16**(3): p. 499-509.
413. Oberdoerffer, S., *A conserved role for intragenic DNA methylation in alternative pre-mRNA splicing.* Transcription, 2012. **3**(3): p. 106-9.
414. Osmark, P., et al., *Unique splicing pattern of the TCF7L2 gene in human pancreatic islets.* Diabetologia, 2009. **52**(5): p. 850-4.
415. Shukla, S., et al., *CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing.* Nature, 2011. **479**(7371): p. 74-9.
416. Prickett, T.D., et al., *Somatic mutations in MAP3K5 attenuate its proapoptotic function in melanoma through increased binding to thioredoxin.* J Invest Dermatol, 2014. **134**(2): p. 452-60.
417. Yu, J.S. and A.K. Kim, *Platycodin D induces reactive oxygen species-mediated apoptosis signal-regulating kinase 1 activation and endoplasmic reticulum stress response in human breast cancer cells.* J Med Food, 2012. **15**(8): p. 691-9.
418. McGrath, J.A., et al., *Germline Mutation in EXPH5 Implicates the Rab27B Effector Protein Slac2-b in Inherited Skin Fragility.* Am J Hum Genet, 2012. **91**(6): p. 1115-21.
419. Pigors, M., et al., *Molecular heterogeneity of epidermolysis bullosa simplex: contribution of EXPH5 mutations.* J Invest Dermatol, 2014. **134**(3): p. 842-5.

420. Liu, F., et al., *Dissecting the mechanism of colorectal tumorigenesis based on RNA-sequencing data.* Exp Mol Pathol, 2015. **98**(2): p. 246-53.
421. Nomura, N., et al., *Isolation and characterization of a novel member of the gene family encoding the cAMP response element-binding protein CRE-BP1.* J Biol Chem, 1993. **268**(6): p. 4259-66.
422. Dong, B., et al., *An infectious retrovirus susceptible to an IFN antiviral pathway from human prostate tumors.* Proc Natl Acad Sci U S A, 2007. **104**(5): p. 1655-60.
423. Su, W.H., et al., *How genome-wide SNP-SNP interactions relate to nasopharyngeal carcinoma susceptibility.* PLoS One, 2013. **8**(12): p. e83034.
424. Qi, L. and Y. Ding, *Involvement of the CREB5 regulatory network in colorectal cancer metastasis.* Yi Chuan, 2014. **36**(7): p. 679-84.
425. Wang, K.S., et al., *Parent-of-origin effects of FAS and PDLIM1 in attention-deficit/hyperactivity disorder.* J Psychiatry Neurosci, 2012. **37**(1): p. 46-52.
426. Yoshiyama, K., et al., *CD156 (human ADAM8): expression, primary amino acid sequence, and gene location.* Genomics, 1997. **41**(1): p. 56-62.
427. Romagnoli, M., et al., *ADAM8 expression in invasive breast cancer promotes tumor dissemination and metastasis.* EMBO Mol Med, 2014. **6**(2): p. 278-94.
428. Shen, Z., et al., *Both macrophages and hypoxia play critical role in regulating invasion of gastric cancer in vitro.* Acta Oncol, 2013. **52**(4): p. 852-60.
429. Li, S.Q., et al., *Neutralization of ADAM8 ameliorates liver injury and accelerates liver repair in carbon tetrachloride-induced acute liver injury.* J Toxicol Sci, 2014. **39**(2): p. 339-51.
430. Yang, Z., et al., *Expression of A disintegrin and metalloprotease 8 is associated with cell growth and poor survival in colorectal cancer.* BMC Cancer, 2014. **14**: p. 568.
431. Errico, A., *Gastrointestinal cancer: ADAM8 provides new hope in pancreatic cancer.* Nat Rev Clin Oncol, 2015. **12**(3): p. 126.
432. Gole, L., et al., *Characterization of breakpoints in the GABRG3 and TSPY genes in a family with a t(Y;15)(p11.2;q12).* Am J Med Genet A, 2004. **125A**(2): p. 177-80.
433. Freichel, M., V. Tsvilovskyy, and J.E. Camacho-Londono, *TRPC4- and TRPC4-containing channels.* Handb Exp Pharmacol, 2014. **222**: p. 85-128.
434. Vanden Abeele, F., et al., *Two types of store-operated Ca2+ channels with different activation modes and molecular origin in LNCaP human prostate cancer epithelial cells.* J Biol Chem, 2004. **279**(29): p. 30326-37.
435. Veliceasa, D., et al., *Transient potential receptor channel 4 controls thrombospondin-1 secretion and angiogenesis in renal cell carcinoma.* FEBS J, 2007. **274**(24): p. 6365-77.
436. Zeng, B., et al., *TRPC channels and their splice variants are essential for promoting human ovarian cancer cell proliferation and tumorigenesis.* Curr Cancer Drug Targets, 2013. **13**(1): p. 103-16.
437. Jiang, H.N., et al., *Involvement of TRPC channels in lung cancer cell differentiation and the correlation analysis in human non-small cell lung cancer.* PLoS One, 2013. **8**(6): p. e67637.
438. Martelotto, L.G., et al., *Breast cancer intra-tumor heterogeneity.* Breast Cancer Res, 2014. **16**(3): p. 210.

439. de Bono, J.S. and A. Ashworth, *Translating cancer research into targeted therapeutics.* Nature, 2010. **467**(7315): p. 543-9.
440. Yuan, E., et al., *A single nucleotide polymorphism chip-based method for combined genetic and epigenetic profiling: validation in decitabine therapy and tumor/normal comparisons.* Cancer Res, 2006. **66**(7): p. 3443-51.
441. Lyko, F. and R. Brown, *DNA methyltransferase inhibitors and the development of epigenetic cancer therapies.* J Natl Cancer Inst, 2005. **97**(20): p. 1498-506.
442. Issa, J.P., et al., *Increased cytosine DNA-methyltransferase activity during colon cancer progression.* J Natl Cancer Inst, 1993. **85**(15): p. 1235-40.
443. Suzuki, T., et al., *Design, synthesis, inhibitory activity, and binding mode study of novel DNA methyltransferase 1 inhibitors.* Bioorg Med Chem Lett, 2010. **20**(3): p. 1124-7.
444. Shilpi, A., et al., *Mechanisms of DNA methyltransferase-inhibitor interactions: Procyanidin B2 shows new promise for therapeutic intervention of cancer.* Chem Biol Interact, 2015. **233**: p. 122-38.
445. Cihak, A., *Biological effects of 5-azacytidine in eukaryotes.* Oncology, 1974. **30**(5): p. 405-22.
446. Kaminskas, E., et al., *FDA drug approval summary: azacitidine (5-azacytidine, Vidaza) for injectable suspension.* Oncologist, 2005. **10**(3): p. 176-82.
447. Kantarjian, H., et al., *Decitabine improves patient outcomes in myelodysplastic syndromes: results of a phase III randomized study.* Cancer, 2006. **106**(8): p. 1794-803.
448. Patra, A., et al., *5-Aza-2'-deoxycytidine stress response and apoptosis in prostate cancer.* Clin Epigenetics, 2011. **2**(2): p. 339-48.
449. Momparler, R.L., L.F. Momparler, and J. Samson, *Comparison of the antileukemic activity of 5-AZA-2'-deoxycytidine, 1-beta-D-arabinofuranosylcytosine and 5-azacytidine against L1210 leukemia.* Leuk Res, 1984. **8**(6): p. 1043-9.
450. Fang, M.Z., et al., *Tea polyphenol (-)-epigallocatechin-3-gallate inhibits DNA methyltransferase and reactivates methylation-silenced genes in cancer cell lines.* Cancer Res, 2003. **63**(22): p. 7563-70.
451. Jagadeesh, S., et al., *Mahanine reverses an epigenetically silenced tumor suppressor gene RASSF1A in human prostate cancer cells.* Biochem Biophys Res Commun, 2007. **362**(1): p. 212-7.
452. Liu, Z., et al., *Curcumin is a potent DNA hypomethylation agent.* Bioorg Med Chem Lett, 2009. **19**(3): p. 706-9.
453. Arce, C., et al., *Hydralazine target: from blood vessels to the epigenome.* J Transl Med, 2006. **4**: p. 10.
454. Villar-Garea, A., et al., *Procaine is a DNA-demethylating agent with growth-inhibitory effects in human cancer cells.* Cancer Res, 2003. **63**(16): p. 4984-9.
455. Lee, B.H., et al., *Procainamide is a specific inhibitor of DNA methyltransferase 1.* J Biol Chem, 2005. **280**(49): p. 40749-56.
456. Isakovic, L., et al., *Constrained (l-)-S-adenosyl-l-homocysteine (SAH) analogues as DNA methyltransferase inhibitors.* Bioorg Med Chem Lett, 2009. **19**(10): p. 2742-6.

457. Saavedra, O.M., et al., *SAR around (l)-S-adenosyl-l-homocysteine, an inhibitor of human DNA methyltransferase (DNMT) enzymes.* Bioorg Med Chem Lett, 2009. **19**(10): p. 2747-51.
458. Song, J., et al., *Structure of DNMT1-DNA complex reveals a role for autoinhibition in maintenance DNA methylation.* Science, 2011. **331**(6020): p. 1036-40.
459. Jia, D., et al., *Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation.* Nature, 2007. **449**(7159): p. 248-51.
460. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.
461. Takeshita, K., et al., *Structural insight into maintenance methylation by mouse DNA methyltransferase 1 (Dnmt1).* Proc Natl Acad Sci U S A, 2011. **108**(22): p. 9055-9.
462. Brooksbank, C., G. Cameron, and J. Thornton, *The European Bioinformatics Institute's data resources: towards systems biology.* Nucleic Acids Res, 2005. **33**(Database issue): p. D46-53.
463. Apweiler, R., et al., *UniProt: the Universal Protein knowledgebase.* Nucleic Acids Res, 2004. **32**(Database issue): p. D115-9.
464. Thompson, J.D., D.G. Higgins, and T.J. Gibson, *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.* Nucleic Acids Res, 1994. **22**(22): p. 4673-80.
465. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.
466. Morris, G.M., et al., *Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function.* Journal of Computational Chemistry, 1998. **19**(14): p. 1639-1662.
467. Venkatachalam, C.M., et al., *LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites.* J Mol Graph Model, 2003. **21**(4): p. 289-307.
468. Friesner, R.A., et al., *Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy.* J Med Chem, 2004. **47**(7): p. 1739-49.
469. Wu, G., et al., *Detailed analysis of grid-based molecular docking: A case study of CDOCKER-A CHARMm-based MD docking algorithm.* J Comput Chem, 2003. **24**(13): p. 1549-62.
470. Pronk, S., et al., *GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit.* Bioinformatics, 2013. **29**(7): p. 845-854.
471. Schuttelkopf, A.W. and D.M. van Aalten, *PRODRG: a tool for high-throughput crystallography of protein-ligand complexes.* Acta Crystallogr D Biol Crystallogr, 2004. **60**(Pt 8): p. 1355-63.
472. Hou, T., et al., *Assessing the performance of the MM/PBSA and MM/GBSA methods. 1. The accuracy of binding free energy calculations based on molecular dynamics simulations.* J Chem Inf Model, 2011. **51**(1): p. 69-82.
473. Kollman, P.A., et al., *Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models.* Acc Chem Res, 2000. **33**(12): p. 889-97.

474. Tsui, V. and D.A. Case, *Theory and applications of the generalized Born solvation model in macromolecular simulations.* Biopolymers, 2000. **56**(4): p. 275-91.
475. Baldwin, A.S., Jr., *The NF-kappa B and I kappa B proteins: new discoveries and insights.* Annu Rev Immunol, 1996. **14**: p. 649-83.
476. Dinicola, S., et al., *Apoptosis-inducing factor and caspase-dependent apoptotic pathways triggered by different grape seed extracts on human colon cancer cell line Caco-2.* Br J Nutr, 2010. **104**(6): p. 824-32.
477. Kaur, M., C. Agarwal, and R. Agarwal, *Anticancer and cancer chemopreventive potential of grape seed extract and other grape-based products.* J Nutr, 2009. **139**(9): p. 1806S-12S.
478. Patra, S.K., *Ras regulation of DNA-methylation and cancer.* Exp Cell Res, 2008. **314**(6): p. 1193-201.
479. Nass, S.J., et al., *Aberrant methylation of the estrogen receptor and E-cadherin 5' CpG islands increases with malignant progression in human breast cancer.* Cancer Res, 2000. **60**(16): p. 4346-8.
480. Oshiro, M.M., et al., *Mutant p53 and aberrant cytosine methylation cooperate to silence gene expression.* Oncogene, 2003. **22**(23): p. 3624-34.
481. Ben Gacem, R., et al., *Clinicopathologic significance of DNA methyltransferase 1, 3a, and 3b overexpression in Tunisian breast cancers.* Hum Pathol, 2012. **43**(10): p. 1731-8.
482. Hurtubise, A. and R.L. Momparler, *Evaluation of antineoplastic action of 5-aza-2'-deoxycytidine (Dacogen) and docetaxel (Taxotere) on human breast, lung and prostate carcinoma cell lines.* Anticancer Drugs, 2004. **15**(2): p. 161-7.

Curriculum Vitae

Arunima Shilpi

Research Scholar,
Epigenetics and Cancer Research Laboratory,
Department of Life Science,
National Institute of Technology,
Rourkela, Odisha, India.
Pin code: 769008
Email ID: 511ls102@nitrkl.ac.in,
arunima.bioinformatics@yahoo.com

OBJECTIVE

To be dedicated towards my work and utilize logical and analytical skills to resolve a given scientific problem

SCHOLARSHIP AND ACHIEVEMENTS

- Qualified CSIR-UGC-JRF JUNE, 2011 Rank-62
- Qualified GATE, 2011 IN BIOTECHNOLOGY: 96 percentile
- Qualified CSIR NET-LS DECEMBER, 2010: Rank-73

EDUCATION

- Visiting Research Scholar at Northwestern University, Chicago, IL, USA. (June 2014-Decmber 2014).
- Pursing **Ph.D.** at NIT Rourkela in Epigenetics and Cancer Biology- 2011.
- **Post-Graduation:** M.Sc. in Bioinformatics from Banasthali University, Rajasthan, 2009.
- **Graduation:** B.Sc. in Biotechnology, Genetics and Biochemistry from Garden City College (Bangalore University), Karnataka, 2007.
- **Intermediate:** ISCE board from St. Joseph Convent, Patna (Bihar).
- **Matriculation:** ICSE board from St. Paul's School, Barauni Refinery (Bihar).

ACADEMIC DISTINCTIONS

- **Post-Graduation** : 76%
- **Graduation** : 84%
- **Intermediate** : 78%
- **Matriculation** : 76%
- **Thesis title**: "Profiling of DNA methylation and single nucleotide polymorphisms for diagnosis, prognosis and targeting DNA methyltransferase of therapeutic intervention of breast cancer.

PUBLICATIONS

1. **Shilpi A**, Bi Y , Jung S, Patra S.K., Davuluri R.V. "Identification of genetic and epigenetic variants associated with breast cancer prognosis by integrative bioinformatics analysis of multi-omics data". Accepted in Cancer Informatics.
2. **Shilpi A**, Parbin S, Pradhan N, Kar S, Sengupta D, Deb M, Patra S.K. "Genetic and epigenetic biomarkers and their synergistic predisposition in breast cancer". Under review. Cellular and Molecular Life Sciences.
3. **Shilpi A**, Parbin S, Sengupta D, Kar S, Deb M, Rath SK, Pradhan N, Rakshit M, Patra SK. Mechanisms of DNA methyltransferase-inhibitor interactions: Procyanidin B2 shows new promise for therapeutic intervention of cancer. Chem Biol Interact. 2015 Mar 31; 233: 122-138.
4. **Shilpi A**, Parbin S, Sengupta D, Kar S, Deb M, Rath SK, Pradhan N, Rakshit M, Patra SK; Molecular Dynamic Simulation and Free Energy of Binding Analysis of Novel Compounds from ZINC Database against Wild Type and Engineered Mutant (C1191I, C1191K)DNA Methyltransferase 1. J J Enzymol Enzy Eng. 2015, 1(1): 002.
5. Parbin S, **Shilpi A**, Kar S, Pradhan N, Sengupta D, Deb M, Rath SK, Patra SK. "Insights into the molecular interactions of thymoquinone with histone deacetylase: evaluation of the therapeutic intervention potential against breast cancer". Mol Biosyst. 2015 Dec 15; 12(1):48-58.
6. Parbin S, Kar S, **Shilpi A**, Sengupta D, Deb M, Rath SK, Patra SK. "Histonedeacetylases: a saga of perturbed acetylation homeostasis in cancer. J Histochem Cytochem." 2014 Jan; 62(1):11-33.
7. Kar S, Sengupta D, Deb M, **Shilpi A**, Parbin S, Rath SK, Pradhan N, Rakshit M, Patra SK. Expression profiling of DNA methylation-mediated epigenetic gene-silencing factors in breast cancer. Clin Epigenetics. 2014 Oct 13;6 (1):20.
8. Deb M, Kar S, Sengupta D, **Shilpi A**, Parbin S, Rath SK, Londhe VA, Patra SK. Chromatin dynamics: H3K4 methylation and H3 variant replacement during development and in cancer. Cell Mol Life Sci. 2014 Sep;71(18):3439-63.
9. Kar S, Parbin S, Deb M, **Shilpi A**, Sengupta D, Rath SK, Rakshit M, Patra A, Patra SK. Epigenetic choreography of stem cells: the DNA demethylation episode of development. Cell Mol Life Sci. 2014 Mar; 71(6):1017-32
10. Kar S, Deb M, Sengupta D, **Shilpi A**, Parbin S, Torrisani J, Pradhan S, Patra S. An insight into the various regulatory mechanisms modulating human DNA methyltransferase 1 stability and function. Epigenetics. 2012 Sep;7(9):994-1007
11. Kar S, Deb M, Sengupta D, **Shilpi A**, Bhutia SK, Patra SK. Intricacies of hedgehog signaling pathways: a perspective in tumorigenesis. Exp Cell Res. 2012 Oct 1;318(16):1959-72.

12. Deb M, Sengupta D, Rath SK, Kar S, Parbin S, **Shilpi A**, Pradhan N, Bhutia SK,Roy S, Patra SK. Clusterin gene is predominantly regulated by histone modifications in human colon cancer and ectopic expression of the nuclear isoform induces cell death. Biochim Biophys Acta. 2015 Aug; 1852(8):1630-45.
13. Deb M, Sengupta D, Kar S, Rath SK, Parbin S, **Shilpi A**, Roy S, Das G, Patra SK. Elucidation of caveolin 1 both as a tumor suppressor and metastasis promoter in light of epigenetic modulators. Tumour Biol. 2014 Dec; 35(12):12031-47.

TOTAL EXPERIENCE (Training / Projects) in years / months

1. **Bhabha Atomic Research Center, Mumbai**, (External Project, **2009**). Worked for six months on "Design of probes to be used in microarray for identification of enteric bacteria."
2. **Banasthali University, Rajasthan** (Internal project **2008)**.Worked for six months on "Drug Target Identification using Metabolic Pathway Information."
3. **IBI Biosolution, Chandigarh**, 2008.Worked for a month on "*In silico* drug designing for the target protein 1UV7 identified in Vibrio cholerae responsible for cholera."
4. **NIIT, Bangalore**,2006.Worked for six month which included certification course in information technology learning about C++ and SQL database.
5. **Syngene Institute of Bioscience, Bangalore**, 2005. Worked for a month on "bacterial and fungal in-vitro cultures, ranking personal hygiene, antibiotic sensitivity test."

ADDITIONAL SKILLS

COMPUTER PROFICIENCY:

Languages :	PERL, C and DATA STRUCTURES IN C, Basics of JAVA, CGI, R
Web Technologies:	HTML
Operating Systems:	LINUX, DOS, Windows
Others	SQL, RDBMS, My SQL, Oracle

BIOINFORMATICS TOOLS:

SWISS-PDBV, RASMOL, Modeller9V4, Quantum3.3, Hex4.5, BLAST (Web based and Standalone), FASTA, MUSCLE, CLUSTAL, Autodock, Glide_XP, CDOCKER, LigandFit, Gromacs, Amber

PERSONAL STRENGTHS

- I have had laboratory exposure on many occasions and I know what it is to have patience and a sincere attitude towards one work.

- Over the years I feel I am on my way to develop a scientific and logical approach to solve queries.
- I can work unflaggingly with resolve for hours at stretch.
- I am an optimist and rarely if ever, do I feel nervous by challenges.

REFERNCES

1. **Dr. Samir Kumar Patra,**
   Associate Professor,
   Department of Life Science,
   National Institute of Technology Rourkela,
   Rourkela, Orissa-769008, India.
   Phone: 91-6612-462683 (Office); 9438168145 (Mobile)
   E-mail: skpatra_99@yahoo.com and samirp@nitrkl.ac.in
2. **Dr. Bibekanand Mallick**,
   Assistant Professor
   Department of Life Science
   National Institute of Technology Rourkela
   Rourkela - 769 008
   Odisha, India
   Phone: (0661) 246-2685
   Email: mallickb@nitrkl.ac.in , vivek.iitian@gmail.com
3. **Dr. Ramana V Davuluri**,
   Professor of Preventive Medicine & Neurological Surgery
   Director of Cancer Informatics Core
   Department of Preventive Medicine
   Northwestern University - Feinberg School of Medicine
   750 N Lake Shore Drive, 11-168
   Chicago, IL, USA 60611
   Tel: 1-312-503-2320
   Fax: 1-312-503-5388
   Email: ramana.davuluri@northwestern.edu

DECLARATION

I hereby declare that all the information mentioned above is true to the best of my knowledge.

Place-Rourkela, India ***Arunima Shilpi***