Distribution of SNPs based on CG/CG to TG/CA mutation in different regions of human genome

Submitted in partial fulfilment of the requirement of the

Degree of

MASTERS OF SCIENCE IN BIOTECHNOLOGY

Under the guidance of:

Dr. Vikas Handa

Assistant professor

Submitted by:

Arash Kumari

Roll No. 301001005

DEPARTMENT OF BIOTECHNOLOGY AND ENVIRONMENTAL SCIENCES

THAPAR UNIVERSITY

Patiala-147004

JULY -2012

CERTIFICATE

This is to certify that thesis entitled "**Distribution of SNP based on CG/CG to TG/CA mutation in different region of human genome**" submitted by Arash Kumari in partial fulfilment of the requirement for the award of the degree of masters in Biotechnology, Department of Biotechnology and Environmental Sciences, Thapar University, Patiala, is a record of student's own work carried out by her under my supervision and guidance. The report has not been submitted for the award of any degree or certificate in this or any other university or Institute.

18/07/2012.

Dr. Vikas Handa
Supervisor
DBTES, TU
Patiala

Dr. M.S.Reddy
Head
DBTES, TU
Patiala

Dr.S..K.Majumdar
Dean
(Academic Affairs)
Thapar University
Patiala

CANDIDATE'S DECLARATION

I, hereby declare that the work presented in the dissertation entitled "**Distribution of SNP based on CG/CG to TG/CA mutation in different region of human genome**" in partial fulfilment of the requirement for the award of the degree of masters in Biotechnology, Department of Biotechnology and Environmental Sciences, Thapar University, Patiala, is an authentic record of my own work during the period of six months from January 2012 to June 2012, under the supervision of Dr. Vikas Handa, Assistant professor, Department of Biotechnology & Environmental Sciences, Thapar University. The report has not been submitted for the award of any other degree or certificate in this or any other university.

Arash Kumari.

Arash Kumari

Roll No. 301001005

Place: Patiala

Date: 18-07-2012

ACKNOWLEDGEMENT

This dissertation would not have been possible without the help and support of the kind people around me, to only some of whom it is possible to give particular mention here.

I am deeply indebted to my guide, **Dr. Vikas Handa,** Assistant Professor, Department of Biotechnology and Environmental sciences, T.U., Patiala whose help, stimulating suggestions and encouragement helped me in all the time of research and writing of this thesis, not to mention his advice and unsurpassed knowledge. His Guidance in pursuance of this work has been invaluable on both an academic and a personal level, for which I am extremely grateful.

I am sincerely thankful to **Dr. M. S. Reddy**, Head of Department of Biotechnology & Environmental sciences, T.U., Patiala, his immense concern throughout the project work. I wish to acknowledge the kind help cooperation and moral support of all faculty members of DBTES. Their suggestions and constructive criticism were highly useful.

I am obliged to **Ms. Methoxy Rishiraj**, Ph.D. scholar, at Department of Biotechnology and Environmental sciences, T.U., Patiala, for her kind cooperation and support which make my work easier and enjoyable. Above all, I am grateful to almighty God and parents for blessing me to complete this work successfully. For any errors or inadequacies that may remain in this work, of course, the responsibility is entirely my own.

Date: (Arash Kumari)

Place:

DEDICATED

TO

MY PARENTS

CONTENTS

LIST OF ABBREVATIONS

A - Adenine

AdoMet - S-adenosyl-l-methionine

B - C or G or T but not A

C - Cytosine

C-5 - carbon at 5^{th} position

CG-SNP - SNP generated due to CG/CG to TG/CA mutation

D - A or G or T but not C

DNMT - DNA methyltransferase

DNMT1 - DNA methyltransferase1

DNMT3a - DNA methyltransferase 3a

DNMT3b - DNA methyltransferase 3b

ExpCpG - expected frequency of CpG dinucleotide

G - Guanine

H - A or C or T but not G

K – Keto bases (T or G)

N - Any nucleotide (A or C or G or T)

N4 - nitrogen at 4^{th} position

N6 - nitrogen at 6^{th} position

ObsCpG - observed frequency of CpG dinucleotide

ObsCpG/ExpCpG - ratio of observed frequency of CpG dinucleotide and Expected frequency of CpG dinucleotide

R – Purine

SNP - Single Nucleotide Polymorphism

S - Strong bonding nucleotide (G or C)

V- (A or G or C but not T)

T - Thymine

W – Weak bonding nucleotide (A or T)

Y - Pyrimidine

LIST OF TABLES

Page no.

LIST OF FIGURES

Page no.

ABSTRACT

Methylated cytosines undergo spontaneous deamination leading to its conversion into thymine placed opposite to guanine in the complementary strand. If the mismatch is not repaired before replication it generates mutation. CpG mutation is one of the contributory factors that have generated SNPs in human genome [22]. In present study we have investigated the correlation between methylation level at flanks of CG-SNP (based on CG/CG to TG/CA) and the allele frequencies of these SNP across the human genome. Further we have investigated the distribution of CG-SNP in different genomic regions such as exons, introns, 5' UTR and 3' UTR to study the effect of evolutionary pressure on changes in sequences using SNPs as tools. Also, we have determined flanking sequences of up to ±4 base pairs surrounding the central CG-SNP sites that show characteristic high and low allele frequencies in human genome.

CHAPTER1

INTRODUCTION

1. INTRODUCTION

The genome dynamically responds to the environment. Stress, diet, behaviour, toxins and other factors activate chemical switches that regulate gene expression and this interaction studies comes under the epigenetics. Conrad Waddington (1905-1975) is often credited with coining the term epigenetics in 1942 as “the branch of biology which studies the causal interactions between genes and their products, which bring the phenotype into being” [1].

Epigenetics is the study of heritable changes in phenotype or gene expression caused by mechanisms other than changes in the underlying DNA sequence. It refers to functionally relevant modifications to the genome that do not involve a change in the nucleotide sequence [2]. Examples of such changes are DNA methylation and histone modification, both of which serve to regulate gene expression without altering the underlying DNA sequence. These changes may remain through cell divisions for the remainder of the cell's life and may also last for multiple generations. Like genetic changes, epigenetic changes are preserved when a cell divides. A cell's epigenome is the overall epigenetic state of a cell [3].

DNA Methylation

DNA methylation is a stable but not irreversible epigenetic signal that silences gene expression. It has a variety of important functions in mammals, including control of gene expression, cellular differentiation and development, preservation of chromosomal integrity, parental imprinting and X-chromosome inactivation. In addition, it has been implicated in brain function and the development of the immune system [3]. Somatic alterations in genomic methylation patterns contribute to the etiology of human cancers and ageing. It is tightly interwoven with the modification of histone tails and other epigenetic signals. DNA methylation is functionally connected to the modification pattern of histone tails, which can be acetylated, methylated, ubiquitinated or phosphorylated [4].

DNA methylation was discovered in calf thymus DNA by Hotchkiss in 1948. It occurs at the N6 position of adenine residues and the N4 and C5 positions of cytosine residues, only the last type being observed in higher eukaryotes, including mammals [5].

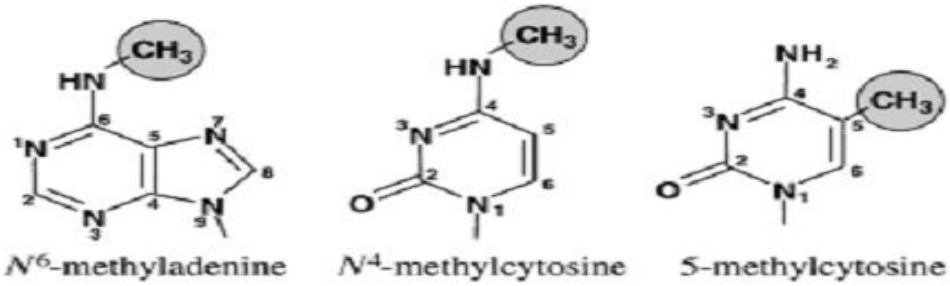

Figure 1.1: Structures of methylated bases occurring in DNA

Source: Jeltsch A, Chem BioChem 2002, 3, 274-293

The DNA methyltransferase family of enzymes catalyze the transfer of a methyl group to the 5-carbon position of cytosine. All DNA methyltransferase use S-adenosyl-L-methionine (AdoMet) as the source of methyl group. Three active DNA methyltransferases have been identified in mammals. They are named DNMT1, DNMT3A, and DNMT3B that methylate cytosine in 5'CpG-3'context. A fourth enzyme also exist named DNMT2[6; 7; 8].

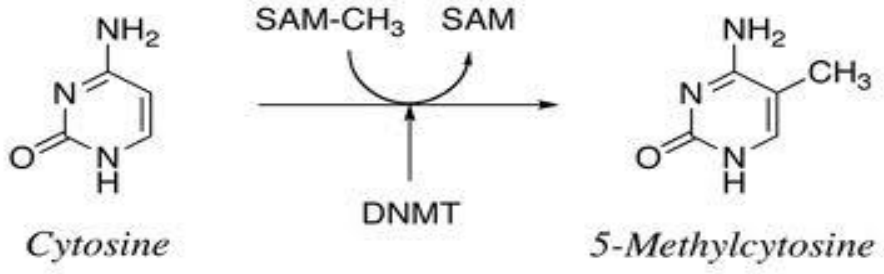

Figure 1.2: Conversion of cytosine to 5-methylctosine

Source: Dr Adele Murrell (CRUK CRI)

DNA methyltransferases

In mammalian cells, DNA methylation occurs mainly at the C5 position of CpG dinucleotides and is carried out by two general classes of enzymatic activities – maintenance methylation and *de novo* methylation. Maintenance methylation activity is necessary to preserve DNA methylation after every cellular DNA replication cycle. The *de novo* methylation set up DNA methylation patterns early in development. Maintenance methyltransferases add methylation to DNA when one strand is already methylated. These work throughout the life of the organism to maintain the methylation pattern that had been established by the de novo methyltransferases [10] .

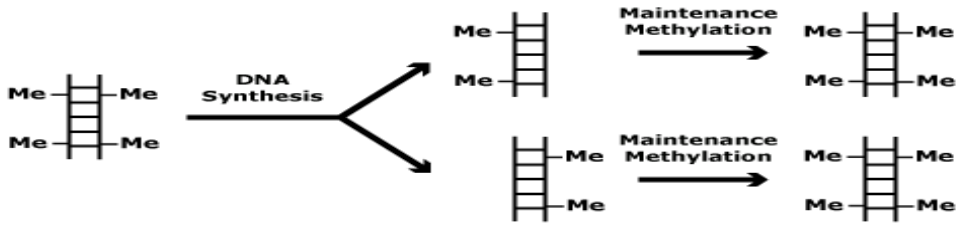

Figure 1.3: Maintenance methylation - DNMT1

Source: UCSF School of medicine

De novo methyltransferases recognize some unmethylated position in the DNA and methylate cytosines of those CpGs. These are expressed mainly in early embryo development and they set up the pattern of methylation [10].

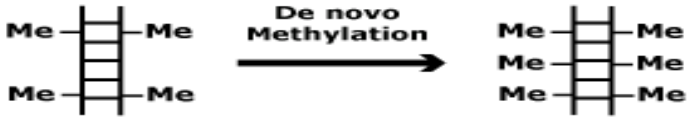

Figure 1.4: *De novo* methylation - DNMT3a, DNMT3b
Source: UCSF School of medicine

CpG dinucleotide depletion

Most eukaryotes methylate only a small percentage of CpG sites, but 70-80% of CpG cytosines are methylated in vertebrates. Methylated C residues spontaneously deaminate to form T residues over evolutionary time; hence CpG dinucleotides steadily mutate to TpG dinucleotides, which is evidenced by the under-representation of CpG dinucleotides in the human genome (they occur at only ~20-25% of the expected frequency) [12; 13]. On the other hand, spontaneous deamination of unmethylated C residues gives rise to U residues, a mutation that is quickly recognized and repaired by the cell. When the equivalent deamination reaction occurs on 5-methylcytosine, however, the product, thymine, is not repaired by DNA repair enzymes (and 5-methylcytosine is an order of magnitude less susceptible to deamination than cytosine)[13].

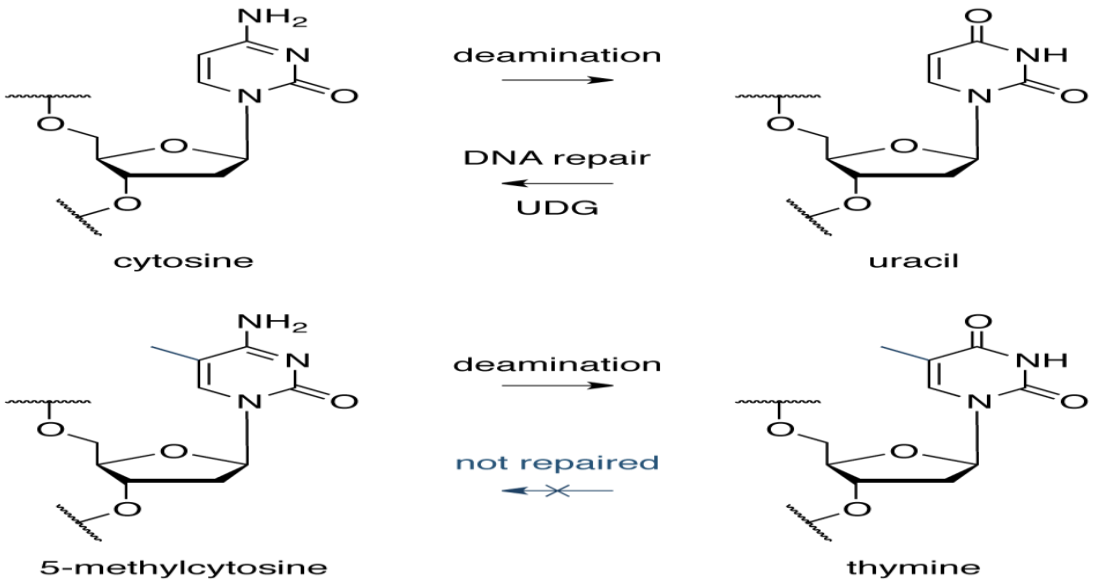

Figure 1.5: CpG dinucleotide depletion

Source: Oxidative DNA damage: mechanisms, mutation, and disease - FASEB J., 2003

Unmethylated CpGs are often grouped in clusters called CpG islands, which are present in the 5' regulatory regions of many genes. CpG islands are associated with nearly half of the genes in mammalian genomes and largely remain unmethylated. CpG islands are rich in GC content and have relatively higher frequency of CpG dinucleotides when compared to rest of the genome [16].

Single nucleotide polymorphism

Single nucleotide polymorphisms, or SNPs, are DNA sequence variations that occur when a single nucleotide (A, T, C, or G) in the genome sequence is altered. For example a SNP might change the DNA sequence A**A**GGCTAA to A**T**GGCTAA. For a variation to be considered a SNP, it must occur in at least 1% of the population. SNPs, which make up about 90% of all human genetic variation, occur every 100 to 300 bases along the 3-billion-base human genome. Two of every three SNPs involve the replacement of cytosine (C) with thymine (T). SNPs can occur in coding (gene) and noncoding regions of the genome. Many SNPs have no effect on cell function, but scientists believe others SNP could predispose people to disease or influence their response to a drug [27].

Although more than 99% of human DNA sequences are the same, variations in DNA sequence can have a major impact on how humans respond to disease; environmental factors such as bacteria, viruses, toxins, and chemicals; and drugs and other therapies. This makes SNPs valuable for biomedical research and for developing pharmaceutical products or medical diagnostics. SNPs are also evolutionarily stable—not changing much from generation to generation-making them easier to follow in population studies [18].

SNP maps can help in identifying the multiple genes associated with complex ailments such as cancer, diabetes, vascular disease, and some forms of mental illness. Several groups worked to find SNPs and ultimately create SNP maps of the human genome. Among these were the U.S. Human Genome Project (HGP) and a large group of pharmaceutical companies called the SNP Consortium or TSC project[26]. Besides the TSC website, SNP data are also available from the following resources:

- dbSNP database - From the National Center for Biotechnology Information (NCBI).
- HGVbase (Human Genome Variation Database) - A human gene-based polymorphism database.

Since SNPs are arising out of mutations, it is evident many of such mutations (transitions) in context of CpG dinucleotide sequences are related to cytosine methylation at fifth position [21]. The present work is based on analysis of such CG-SNPs.

CHAPTER2

REVIEW OF LITERATURE

2. REVIEW OF LITERATURE

2.1. Epigenetics

The term 'epigenetics', which literally means 'outside conventional genetics', is used to describe the study of stable alterations in gene expression potential that arise during development and cell proliferation[1]. Epigenetic processes are essential for development and differentiation, but they can also arise in mature humans and mice, either by random change or under the influence of the environment [2].

Research over the past few years has focused on two molecular mechanisms that mediate epigenetic phenomena: DNA methylation and histone modifications Epigenetic effects by means of DNA methylation are essential in the development and function of healthy cells. It was proposed in 1975 that DNA methylation might be responsible for the stable maintenance of a particular gene expression pattern through mitotic cell division [3]. It has a variety of important functions in mammals, including control of gene expression, cellular differentiation and development, preservation of chromosomal integrity, parental imprinting and X-chromosome inactivation [5].

2.2. DNA Methylation

DNA methylation was discovered in calf thymus DNA by Hotchkiss in 1948. It occurs at the N6 position of adenine residues and the N4 and C5 positions of cytosine residues, only the last type being observed in higher eukaryotes, including mammals. In any case the methyl groups are positioned in the major groove of the DNA, where they do not interfere with the Watson/Crick base pairing capacities of the nucleotides [5].

DNA methylation is carried out by enzymes called DNA methyltransferases on cytosine and adenine bases. All DNA methyltransferases use S-adenosyl-L-methionine (AdoMet) as the sources of methyl group being transferred to the DNA bases. Prokaryotic cytosine and adenine methylation can influence gene transcription, affect cell viability, play important role in mismatch repair of DNA and also serve the restriction-modification systems that protect the bacterial host DNA from cleavage by specific endonucleases [4].

The methylation pattern is inherited by daughter cell genomes during DNA replication by the action of DNA methyltransferase 1 (Dnmt1), which exhibits high preference for a hemimethylated DNA substrate [6; 7; 8]. The genomic methylation pattern is set by *de novo* DNA methylation during gametogenesis in a sex specific fashion and later, after extensive demethylation of the genome, during embryogenesis [9]. The *de novo* methylation is carried out by two *de novo* DNA methyltransferases (MTases), Dnmt3a and Dnmt3b, which methylate unmethylated and hemimethylated DNA [10]. The influence of flanking sequence on the catalytic activity of the Dnmt3a and Dnmt3b *de novo* DNA methyltransferases using a set of synthetic oligonucleotide substrates that covers all possible ±1 flanks in quantitative terms was investigated and methylation kinetics experiments revealed a >13-fold difference between the preferred (RCGY) and disfavored ±1 flanking base-pairs (YCGR). In addition, AT-rich flanks are preferred over GC-rich ones.[11]

2.3. CpG dinucleotides

In mammals, cytosine methylation takes place predominantly at palindromic CpG dinucleotides in both strands of the DNA. The mammalian genomes contain about 60 million CpG dinucleotides and 70–80% of those are modified in a non-random pattern. The transition rate of methylated CpG (5mCpG) to TpG is 10 to50 fold higher than other transitional changes [12];[13].

Because cytosines in GpC dinucleotides are not methylated in mammalian genomes, the difference between the CpG transition rate, measured by the number of CpG to TpG/CpA per CpG dinucleotide in a sequence, and GpC transition rate, measured by the number of GpC to GpT/ApC per GpC dinucleotide, represents the rate of methylation-dependent transition or 5mC deamination rate. By applying this approach to human single nucleotide polymorphism (SNP) data it was indicated that the 5mC deamination rate was exponentially dependent on local GC content [14].

Since one 5mC change would cause loss of two CpGs and the gain of one TpG and one CpA, this should be discernable as a correlation- between CpG deficiency and TpG plus CpA excess. This result provides strong evidence for the conversion of mCpG to TpG plus CpA during evolution. In addition, it shows that the excess of TpG:CpA in vertebrates is as dramatic as the deficiency of CpG:CpG, but has escaped notice because CpG is a self-complementary [15].

2.4. CpG Islands

Although vertebrate DNA is generally depleted in the dinucleotide CpG, it has been shown that some vertebrate genes contain CpG islands, regions of DNA with a high G+C content and a high frequency of CpG dinucleotides relative to the bulk genome. CpG islands are associated with the 5’ ends of all housekeeping genes and many tissue-specific genes, and with the 3’ ends of some tissue-specific genes. A few genes contained both 5’ and 3’ CpG islands, separated by several thousand base-pairs of CpG-depleted DNA. The 5’ CpG islands extended through 5’flanking DNA, exons and introns, whereas most of the 3’ CpG islands appeared to be associated with exons. CpG islands were generally found in the same position relative to the transcription unit of equivalent genes in different species, with some notable exceptions [14].

The first large-scale computational analysis of CpG islands using vertebrate sequences in GenBank was performed by Gardiner-Garden and Frommer, who defined a CpG island as being a 200-bp region of DNA with a high G+C content (greater than 50%) and observed CpG/expected CpG ratio (ObsCpG /ExpCpG) of greater or equal to 0.6. For example, the human *Alu*s, which are highly repetitive short interspersed elements, have an approximately 280-bp consensus sequence, and some of these have relative high %GC and ObsCpG/ExpCpG. This composition makes it difficult to distinguish bona fide CpG islands from the nearly 1,000,000 *Alu* copies per haploid genome [16].

Later study based on genomic analysis of Human, Arabidopsis and some non vertebrate genomes redefined CpG islands with more rigorous values of the above mentioned parameters i.e., regions of DNA of greater than 500 bp with a G+C equal to or greater than 55% and observed CpG expected CpG of 0.65 were more likely to be associated with the 5' regions of genes and this definition excluded most *Alu*-repetitive elements [17].

2.5. Single Nucleotide Polymorphism

Single nucleotide polymorphisms (SNPs) are the most abundant genetic variation in vertebrate genomes. They have been important tools in many biological fields, including mutation mechanisms, genome evolution, disease studies, pharmacogenomics, and fine mapping [15].

Mutation at the nucleotide level does not occur randomly. Recent studies of mutational mechanisms revealed that the influence of neighbouring nucleotides on SNPs was strong in the human and mouse genomes [16]. Specifically, strong biases relative to the genome average were observed at the two adjacent sites of the SNPs and small biases could extend farther, i.e., as far as 200 nucleotides at each flanking side. Further, the bias patterns varied among the SNP types, e.g., the extent of the biases for transition SNPs (A/G and C/T) was much stronger than those for transversion SNPs (A/C, G/T, A/T, and C/G) [17].

Because the SNPs identified in the today's genomes reflect the combinatory evolutionary processes such as methylated CpG mutation hotspots, high transition rate, selection on functional elements, and error-prone DNA replication and repair, a small effective SNP size suggests the strong influence of one or several genetic factors, especially the CpG effects in vertebrate genomes [18].

2.6. Cytosine Methylation and SNP

Human polymorphisms originate as mutations, and the influence of context on mutagenesis should be reflected in the distribution of sequences surrounding single nucleotide polymorphisms (SNPs). Tomso and Bell have performed a computational survey of nearly two million human SNPs to determine if sequence-dependent hotspots for polymorphism exist in the human genome. They show that sequences containing CpG dinucleotides, which occur at low frequencies in the human genome, are 6.7-fold more abundant at polymorphic sites than expected. In contrast, polymorphisms in CpG sequences located within CpG islands, important regulatory regions that modulate gene expression, are 6.8-fold less prevalent than expected. The distribution of polymorphic alleles at CpGs in CpG islands is also significantly different from that in non-island regions. These data strongly support a role for 5-methylcytosine deamination in the generation of human variation, and suggest that variation at CpGs in islands is suppressed.[19]

Substitution patterns at polymorphic sites and bias patterns in nucleotides neighbouring polymorphic sites are important for understanding molecular mechanisms of mutation and genome evolution. Single nucleotide polymorphism (SNP) data and information about surrounding sequence motifs are suitable for studying mutational processes in human and other genomes [20].

There is considerable recent interest in SNPs within every gene in the genome or regularly spaced across the genome as tools for association-mapping of disease-susceptibility genes [21] or identifying polymorphic sites within a known gene that are associated with a trait of interest and may be functional [22].

There are more than two and one-half million SNPs available in the public domain. At present, most of the SNPs are deposited by The SNP Consortium (TSC), the Sanger Genome Centre, and Washington University [23]. This large data set provides an opportunity to investigate substitution patterns as well as neighbouring-nucleotide effects representative of the whole genome, including genic and intergenic regions.

Zhao and Boerwinkle investigated substitution patterns and neighbouring-nucleotide effects for 2,576,903 single nucleotide polymorphisms (SNPs) publicly available through the National Centre for Biotechnology Information (NCBI).

Biased Proportions of substitutions in percentile across the human genome					
Transition SNP		**Transversion SNP**			
A/G	**C/T**	A/C	G/T	A/T	C/G
32.77	**32.81**	8.98	9.06	7.46	8.92

Figure 2.1 Data showing Biased Proportions of substitutions

Source: Zhongming Zhao and Eric Boerwinkle, 2002

The two nucleotides immediately neighbouring the variable site showed major deviation from genome-wide and chromosome-specific expectations, although lesser biases extended as far as 200 bp.

Biased Proportions in percentile at adjacent position relative to the SNP site							
On the 5'site				On the 3'site			
A	**C**	G	T	A	C	**G**	T
1.43	**4.91**	-1.70	-4.62	-4.44	-1.59	**5.05**	0.99

Figure 2.2 Data showing Biased Proportions at adjacent position relative to the SNP site

Source: Zhongming Zhao and Eric Boerwinkle, 2002

This data shows that neighbouring-nucleotide patterns for transitions were dominated by the hypermutability effects of CpG dinucleotides. Transitions were more common than transversions, and the probability of a transversion increased with increasing A + T content at the two adjacent sites. These data provide genome-wide information about the effects of neighbouring nucleotides on mutational and evolutionary processes giving rise to contemporary patterns of nucleotide occurrence surrounding SNPs [16].

CHAPTER3

OBJECTIVE

3. OBJECTIVE

1. To find the correlation between methylation level at flanks of CG-SNP (based on CG/CG to TG/CA mutation) and the allele frequencies of these SNP across the genome.

2. To analyze CG-SNPs (based on CG/CG to TG/CA mutation) in different regions of genome such as exons, introns, 5'UTR and 3'UTR to study the effect of evolution pressure.

3. To establish the tetra nucleotide consensus flanking sequence for CG-SNPs (based on CG/CG to TG/CA mutation) based upon their allele frequencies in human genome.

CHAPTER4

MATERIAL AND METHDOLOGY

4. MATERIAL AND METHDOLOGY

4.1 Criteria for selection of sequences

SNP (Single Nucleotide polymorphism) concerned in this work are those which are generated as a result of CG/CG to TG/CA mutation in different region of Human genome. One of the reasons for this mutation is methylation at C-5 of cytosine in CpG dinucleotide. So first task was to find data containing CpG sites where methylation at C-5 of cytosine had been determined experimentally. The degree of methylation could be determined by various method like methylation-sensitive polymerase chain reaction and additionally that the methylation profile was generated with bisulphite genomic sequencing.

We are using here two sources to obtain methylation data.

4.1.1. Data already published

4.1.2. Methylation database (MethDB)

DNA methylation data are heterogeneous and can range from the global estimation of the total 5-mC content to the exact methylation pattern of a single DNA sequence region. In MethDB; therefore, two types of methylation data exist: *1*) sequence-specific methylation profiles and methylation patterns; and *2*) total methylation content. A methylation pattern is the sequence of the five nucleotide bases (including 5-mC) in a single DNA molecule. Methylation profiles are representations of the average methylation along a sequence.

DATA RECUSION FROM METHDB

MethDB can be accessed from any Web browser (www.methdb.net or www.methdb.de). On the front page, a navigation bar can be found on the left-hand side, and some general information is displayed in the central frame. Clicking on “Search” brings the user to a page with different search options.

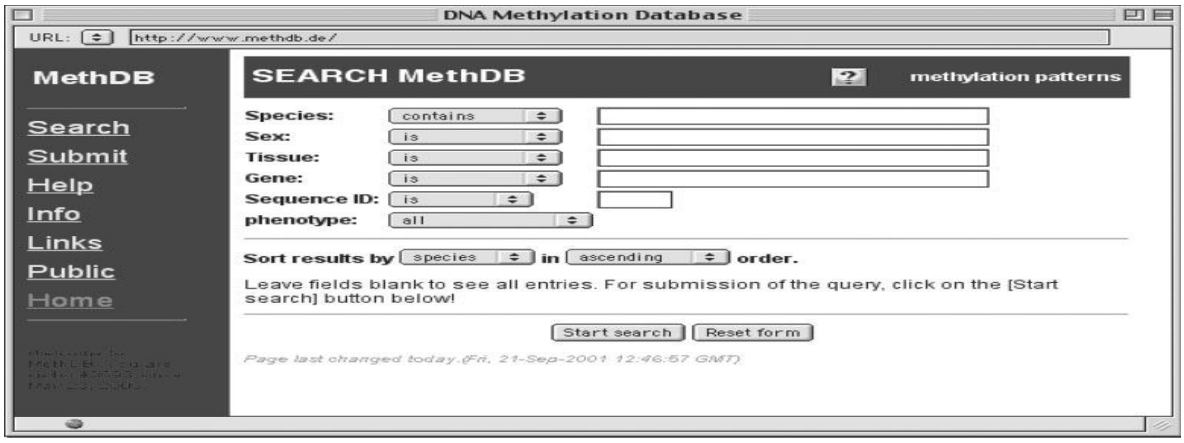

Figure 3.6: Query mask for the experiments that produced sequence-specific methylation patterns and profiles.

Source : DNA Methylation Database "MethDB": a User Guide

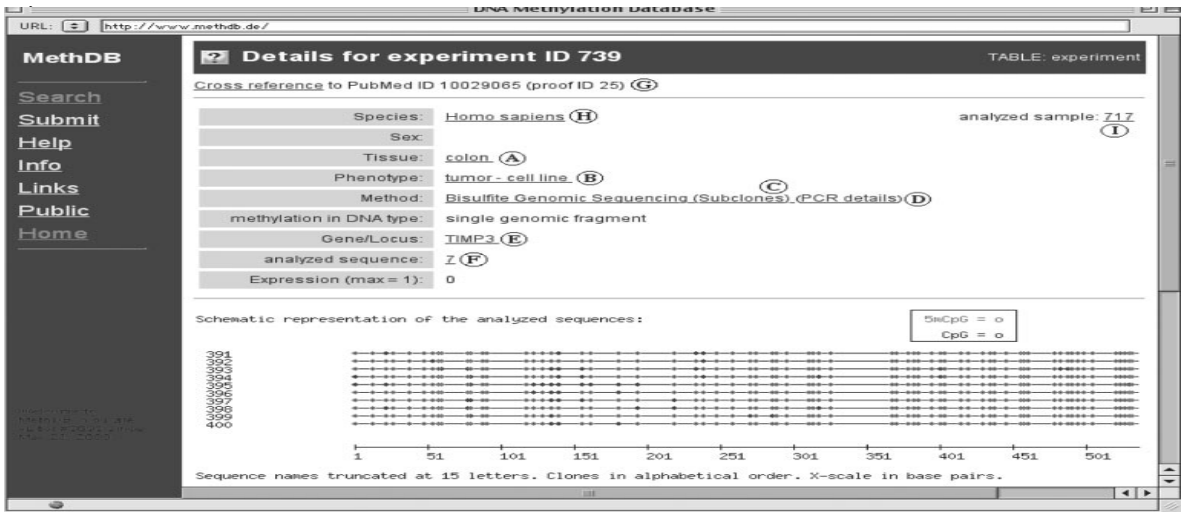

Figure 3.7: Experimental details page. Hypertext links lead to descriptions of the tissue (*A*), the phenotype (*B*) and the experimental procedure (*C*), the polymerase chain reaction (PCR) (*1*) details *D*) and the investigated gene (*E*) and give details about the studied sequence (*F*). Links to PubMed (*G*) and the National Center for Biotechnology taxonomy browser (*H*) provide access to background information in external databases. The sample data can be viewed via the sample ID (*I*) and the guanine- adenine- thymine-cytosine-5-methylcytosine (GATC5-MC) sequence via the pattern ID.

Source: DNA Methylation Database "MethDB": a User Guide

4.2 Sequence Selection

1.	**Source for Table 01 gene:** Human Molecular Genetics, 2011, Vol. 20, No. 3 401–412 doi:10.1093/hmg/ddq476 **Figure2.** (B–D) DNA methylation and histone modifications in the promoter regions of OCT4 (B), SOX2 (C) and NANOG (D) during mhESC differentiation. **Figure3.** (B–E) DNA methylation and histone modifications in the promoter regions of SOX17 (B), GATA4 (C), FOXA2 (D) and CXCR4 (E) during differentiation of hESCs to hepatocytes. **Figure4.** (B and C) DNA methylation and histone modifications in the promoter regions of ALB (B) and HNF4A (C) during differentiation of hESCs into hepatocytes. **Supplementary Table 3**. Primers used for detection of genomic DNA in bisulfite-specific PCR
2.	**Source for Table 02 gene :** PLoS Genet 7(5): e1002085. doi:10.1371/journal.pgen.1002085. **Figure2. (F)** Bisulfite sequencing analysis of the 8 genes in endometrial cells (UtE1104), UtE-iPS-11 and HUES-8 cells. **Supporting Information:** Table S8 Primer list
3.	**Source for Table 03 gene** : Molecular Biology of the Cell Vol. 21, 2066–2077, June 15, 2010 **Figure1.** (C) Bisulfite sequencing analysis of CpG methylation of promoters shown in B. **Supplemental Table S1.** MeDIP-PCR and bisulfite sequencing primers used in this study.
4.	**Source for Table 04- Table 09 gene:** Methylation database (MethDB) available at web browser (www.methdb.de).

Table 4.1 Source of methylation for different genes used in study. And detailed information of sequence analysed regarding chromosome no., position in chromosome (4.1.1-4.1.9).

Sr. No.	Gene Name	Chr. No.	Accession No.	From	To
1	ALB	4	NT_022778.16	74269819	74270218
2	CXCR4	2	NT_022135.16	136875465	136876095
3	FOXA2	20	NT_011387.8	22566703	22567294
4	GATA4	8	NT_077531.4	11560911	11561432
5	HNF4	20	NT_011362.10	42984118	42984560
6	NANOG	12	NT_009714.17	7941622	7941943
7	OCT4	6	NT_007592.15	31138208	31138705
8	SOX2	3	NT_005612.16	181429514	181429730
9	SOX17	19	NT_008183.19	55369985	55370471

Table 4.1.1

Sr. No.	Gene Name	Chr. No	Accession No.	From	To
10	EPHA1	7	NT_007914.15	143105570	143106195
11	GBP3	1	NT_032977.9	89488101	89488492
12	LYST	1	NT_167186.11	236046897	23047245
13	PTPN6	12	NT_009759.16	7055730	7056140
14	RAB25	1	NT_004487.19	156030738	156031231
15	SALL4	20	NT_011362.10	50418802	50419114
16	SP100	2	NT_005403.17	231280652	231280990
17	UBE1L	3	NT__022517.18	49791029	49791055

Table 4.1.2

Sr. No.	Gene Name	Chr. No	Accession No.	From	To
18	CDH1	12	NT_009759.16	6717543	6717820
19	FOXD3	1	NT_032977.9	63786817	63786929
20	GLIS 1	1	NT_032977.9	54200326	54200490
21	HOXA9	7	NT_007819.17	27196875	27197149
22	TERT	5	NT_006576.16	1295561	1295850
23	OXT	20	NT_011387.8	3051835	3052009
24	TNNI2	11	NT_009237.18	17859480	17859676

Table 4.1.3

Sr. No.	Gene Name	Chr. No.	Accession No.	From	To
25	HSFXDNA, FMR1, FRAXA	X	NT_011681.16	146993175	146993315
26	MAGE-A1	X	NT_167198.1	152485899	152486180
27	LAGE-1	X	NT_167198.1	153813319	153813516
28	CDM	X	NT_167198.1	152988222	152988696
29	MSSK1	X	NT_167198.1	153046432	153046810
30	SLC6A8	X	NT_167198.1	152954606	152955002
32	Xist	X	NT_011669.17	73072347	73072666
33	SYBL1	X	NT_167198.1	155111230	155110783
34	GLA	X	NT_011651.17	100662924	100662806

Table 4.1.4

Sr. No.	Gene Name	Chr. No.	Accession No.	From	To
35	ADAMTS5, ADMP-2, ADAMTS1, FLJ36738	21	NT_011512.11	28337911	28338202
36	HLCS, HCS	21	NT_011512.11	38353232	38352956
37	DSCR6, RIPPLY3	21	NT_011512.11	38378113	38378311
38	DSCR3, DCRA, DSCRA, MGC	21	NT_011512.11	38630682	38630810
39	RIPK4, DIK, PKK, RIP4, ANKK2, ANKRD3, MGC, MGC	21	NT_011515.12	43186564	43186748
40	H2BFS	21	NT_011515.12	44985270	44985500
41	C21orf29, TSPEAR, MGC11251 21188	21	NT_011515.12	46126132	46126724
42	C21orf70, PRED56	21	NT_011515.12	46368032	46368380
43	NRIP1, RIP140	21	NT_011512.11	16437718	16438103
44	STCH	21	NT_011512.11	15755449	15755892
45	ABCC13, PRED6	21	NT_011512.11	15646300	15646652
46	RBM11	21	NT_011512.11	15588393	15588854
47	COL6A,PP3610,DKFZp58 E1322	21	NT_011515.12	47517581	47518080
48	TFF3, ITF,TFT,HITF,Hp1b	21	NT_011515.12	27543089	27543316
49	TFF1, Ps2	21	NT_011515.12	4373773	43736773

Table 4.1.5

Sr. No.	Gene Name	Chr. No.	Accession No.	From	To
50	BRCA1	17	NT_010783.15	6551515	6552127
51	HIC-1	17	NT_010718.16	1960827	1961108
52	SLC6A8 pseudo	16	NT_010393.16	32897033	32896565
53	D15S63	15	NT_026446.14	25101536	25101751
54	SNRPN	15	NT_026446.14	25199998	25200271
55	CSPG4,HMW MAA	15	NT_010194.17	76005057	76005259
56	Nanog	12	NT_009714.17	7941605	7941940

Table 4.1.6

Sr. No.	Gene Name	Chr. No.	Accession No.	From	To
57	KAI1	11	NT_009237.18	44587105	44587322
58	GSTP1	11	NT_167190.1	67350574	67351027
59	G6PD	11	NT_167190.1	67350574	67351027
60	CCND1, BCL1	11	NT_167190.1	69455690	69455898
61	GSTP1	11	NT_167190.1	67350574	67351027
62	SRBC	11	NT_009237.18	6341968	6342137
63	CGRP	11	NT_009237.18	14933562	14933842
64	RET	10	NT_033985.7	43571705	43572125

Table 4.1.7

Sr. No.	Gene Name	Chr. No.	Accession No.	From	To
65	BCR-ABL	9	NT_035014.4	133710847	133710550
66	CpG40	9	NT_008413.18	19934648	19934852
67	CpG28	9	NT_008413.18	21395857	21396341
68	CpG32	9	NT_008413.18	21968084	21968616
69	CpG25	9	NT_008413.18	21989427	21989940
70	CpG176	9	NT_008413.18	21995595	21996007
71	CpG54	9	NT_008413.18	23850645	23851278
72	DMR 2-48	9	NT_008413.18	32956299	32955928
73	HUMCIP,CDKN2B,P15	9	NT_008413.18	21998878	21999304
74	CDKN2A,P16-INK4	9	NT_008413.18	21974678	21974897

Table 4.1.8

Sr. No.	Gene Name	Chr. No.	Accession No.	From	To
75	SOX17	8	NT_008183.19	55369986	55370472
76	GATA4	8	NT_077531.4	11561353	11561853
77	MDR1	7	NT_007933.15	87229448	87230634
78	ER	6	NT_025741.15	56298565	56298781
79	CDX1	5	NT_029289.11	149545957	1495461480
80	APC	5	NT_034772.6	113073184	113073603
81	RASSF1	3	NT_022517.18	32010124	32070124
82	SOX2	3	NT_005612.16	181429446	181429662
83	TPEF	2	NT_005403.17	193059238	193059027

Table 4.1.9

4.3 SNP detection in the analysed sequence

Using the above information regarding chromosome start and end position, SNP were find in the analysed sequence by using one of the either approach:

S.No.	Approach for SNP detection	URL
1.	dbSNP database of NCBI	http://www.ncbi.nlm.nih.gov/snp/
2.	Chip Bioinformatics tool	http://snpper.chip.org/
3.	Ensembl genome browser	http://www.ensembl.org

Table 4.2

Allele Frequencies of these SNP were recorded. Here we selecting only those SNP whose immediate neighbouring flanks are cytosine on the 5' side for A/G transition and Guanine on the 3'side for C/T transition.

To study the distribution of SNP in different regions of genome such as exons, introns, 5'UTR and 3'UTR, seven unrelated genes Pou5f1, NFYC, XPG, XPB, POLE3, C-MYC, NANOG, were selected and analyzed.

To establish the flanking sequences of up to ±4 base pairs surrounding the central CG-SNP sites, data has been collected from dbSNP of NCBI which is publically available.

CHAPTER5

RESULTS

5. RESULTS:

Investigation of the relation between DNA methylation and the allele frequencies of CG-SNP

SNPs arise from substitution point mutations and have >/= 1% allele frequency for the least frequent allele. Out of the 6(12) possible substitutions, it may already be expected that those based on transition would be over represented and consequently transversions would be under represented. It is reflected in the reports on SNP data analysis also. In vertebrates, since CpGs are mutable owing to methylation of the cytosine, they contribute significantly to SNPs. There are published reports that have shown that CpG mutation is one of the contributory factors that have generated SNPs in vertebrate genome [22]. Zhao and Boerwinkle have shown that neighbouring-nucleotide patterns for transitions were dominated by the hypermutability effects of CpG dinucleotides and transitions were more common than transversions. It is interesting to investigate the effect of extent of methylation at CpG sites i.e., at just adjacent flanks of CG-SNP (based on CG/CG to TG/CA) and the allele frequencies of these SNP across the genome. We are considering 2 type of substitutions as CG-SNP, one is 5'-C**R**-3'and another is 5'-**Y**G-3'. A significant correlation was expected between mean methylation percentage and allele frequencies in CG-SNPs.

It is hypothesized that higher levels of methylation at CpGs site are expected to result in substitution with higher allele frequencies for 'T's and 'A's. In order to test the hypothesis, genomic sequences from 83 genes (for which methylation data was available) were analysed and from which CG-SNP were searched out. The extent of methylation was compared against allele frequencies for these CG-SNPs. Out of total analysed genes, 31 genes do not have any available data of CG-SNPs. And rest 52 genes contain the 108 CG-SNPs for which methylation data was available, however only 39 CG-SNPs have known allele frequencies.

We determined Pearson coefficient of correlation between DNA methylation and allele frequency of the CG-SNPs and found that there was no significant correlation. It may be inferred that hardly any methylation propensity kind of information be harnessed from the allele frequencies of CG-SNPs.

Distribution analysis of CG-SNP in different region of genome

Further, we investigated the distribution of CG-SNP in different genomic region like intron, exon, 5'UTR and 3'UTR for various genes. For this purpose seven genes were selected at random and number of total SNP, number of CG-SNP, and length for the all the selected genomic regions were determined. The data shown in the following table:

GENE	Exon			5' UTR			3'UTR			Intron		
	X	Y	Z	X	Y	Z	X	Y	Z	X	Y	Z
NFYC	29	5	5187	12	3	306	9	0	775	1227	225	74846
OCT-4	172	25	2749	34	3	54	56	12	264	1412	68	3588
NANOG	28	3	2094	4	0	216	20	5	964	147	41	4566
C-MYC	47	9	2363	13	4	73	11	3	2180	45	8	3002
POLE3	16	3	2801	12	1	248	24	6	1573	30	5	715
XPB	81	18	11172	5	0	95	8	2	306	630	135	25714
XPG	194	25	5460	13	2	122	6	2	98	535	97	24700

Table 5.1 Distribution of CG-SNP in different genomic region

(Whereas, X- Total SNPs, Y- CG-SNPs, Z - Length of region)

When the number of total SNPs per unit length and the number of CG-SNPs per unit length of the gene regions were plotted a significant variation in distribution of SNPs as well as CG-SNPs was found amongst the four regions of the gene, namely Exons, Introns, 5' UTRs and 3' UTRs, it was observed that there are significantly higher number of SNPs (as well as CG-SNPs) in 5' UTR regions when compared to the other three. Though the number of SNPs is least for Exons as it expected because of their direct role in contributing to the protein sequence, there is only a marginal difference from 3'UTR and Intron regions.

On the other hand not much difference was found between the mutual comparative representation of CG-SNPs and total SNPs indicating no effect of CG methylation based mutations in different gene regions. However 3'UTR regions appear to have marginally higher ratio of CG-SNPs against total SNPs when compared to the other three regions.

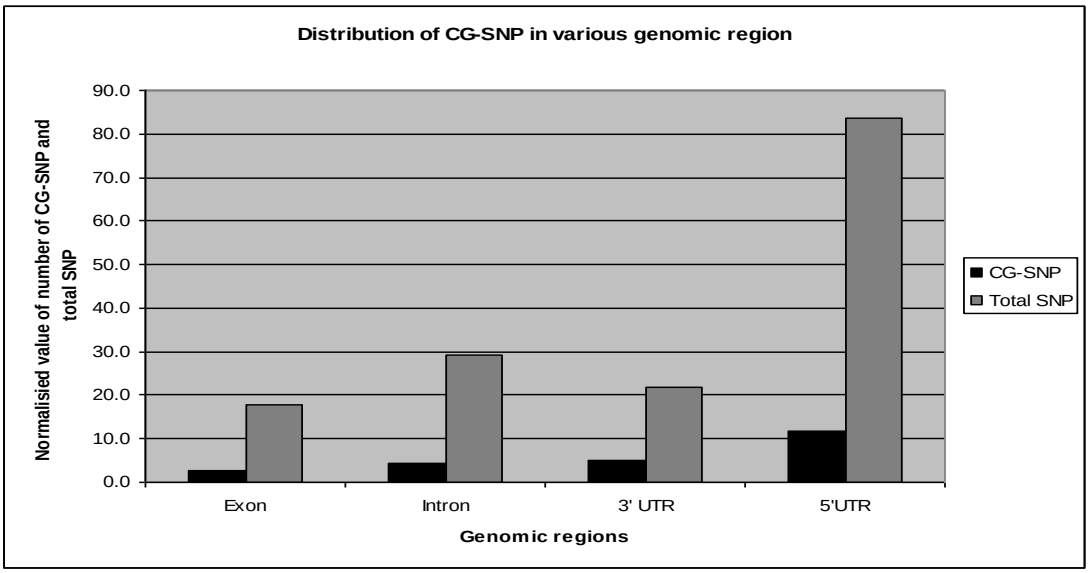

Figure 5.1 Distribution of CG-SNP in different region of genome.

Determining the consensus sequence for CG-SNP flanking bases for high and low allele frequency classes

A data set of 981 CG-SNP sites selected randomly from human genome with their respective allele frequencies was generated. The data set consisted of a heterogeneous population of CG-SNPs with minimum allele frequencies greater than 1%. The data set was arranged according to the order of allele frequencies (A or T). The data set was divided into several classes based on very high and very low allele frequencies ie greater than or equal to 68, 84, 90, 92 and 96 percentile and upto 4, 8, 10, 16 and 32 percentile. When the ordered allele frequencies were plotted, a sigmoidal graph was observed. Based on the sigmoidal profile of allele frequencies, additiona classes were selected that represented the break points in the graph.

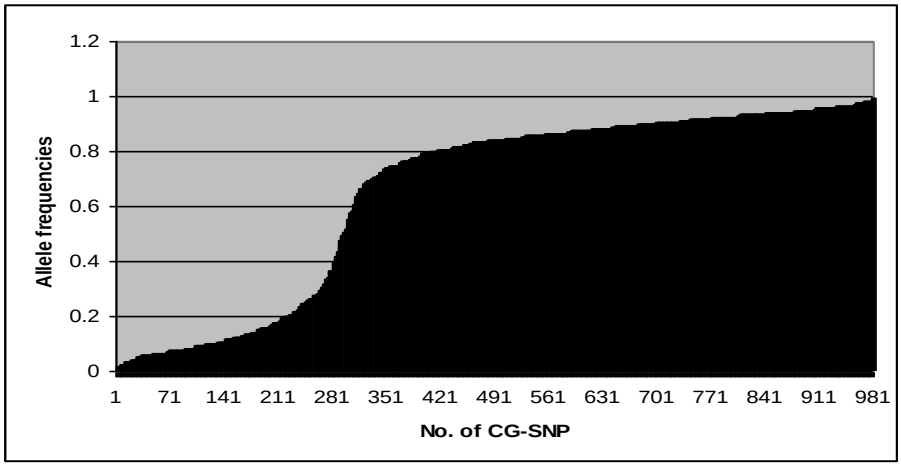

Figure 5.2 Allele frequencies (C or G) of various CG-SNP arranged in increasing order

Using the flanking bases data of the selected classes of high or low allele frequencies, number of each of the four bases were determined for all the 8 flanking sites (-4, -3, -2, -1, +1, +2, +3 & +4). From this matrix of information, over represented bases were identified for each of the 8 positions to generate consensus sequences for all the classes (shown in the table). Using consensus sequence of all the classes one master consensus sequence each for high allele frequency and low allele freuquency were determines as (5'- KVCACGBGCS -3') and (5'- MHBVCGGRRH -3') respectively.

Frequency range (in percentile)	Low frequency consensus sequence	Frequency range (in percentile)	High frequency consensus sequence
4	BVCACGBGCV	96	BVCACGBGCV
8	MVBVCGMRBV	92	TVCACGBSYV
10	MVCVCGKDRD	90	KNCACGBGCV
16	AVCVCGGRRH	84	KGCACGBSBS
26.40	VHBVCGKRRH	68	GACCCGBGSS
27.62	AWBVCGGRRH	66.97	SMCCCGBGCB
27.93	MHBVCGGRRH	58.71	SMCCCGBGCB
32	MHCVCGMGRA		
Master consensus	**MHBVCGGRRH**	Master consensus	**KVCACGBGCS**

Table 5.2 Consensus sequences of bases flanking CG sites at ±1, ±2, ±3, ±4 for high and low frequencies of CG-SNPs.

CHAPTER6

DISSCUSSION

6. DISSCUSSION:

DNA methylation is an important epigenetic modification that plays important role in several processes including gene regulation and chromatin remodelling. It is an enzymatic process that involves transfer of a methyl group to the 5 position of cytosine in CpG context. The methylated CpG is mutagenic, undergoes spontaneous deamination and thus gets converted into TpG/CpA. This phenomenon has lead to loss of CpGs in the course of evolution and at present vertebrate genomes exhibit underrepresentation of CpGs. Thus mutability of methylated cytosines in CpG context has vastly influenced the vertebrate genome structure with prominent examples of evolution of GC rich CpG islands. Some of the less conspicuous effects are point mutations that may effect structure or function of proteins as well as regulation of their genes. Some these mutations may be exhibited as Single Nucleotide Polymorphism (SNP) sites. Thus we find DNA methylation not only regulating the gene expression but indirectly it is also affecting the structure and function of proteins as well as their genetic regulation of expression.

As there are published reports establishing link between CpG methylation and their effect on generation of SNPs, in the current work it has been attempted to exploit the relation between methyled CpG mutation to TpG/CpA and their conversion into SNPs. The first experiment was to determine the correlation studies between allele ('T's or 'A's) frequencies and methylation levels. For this CG-SNPs were selected in which either C or G position was exhibiting SNP and it also had methylation information available in the methylation databases or literature. The hypothesis was based on the fact that if a particular CpG is methylated with higher preference, its mean methylation levels would be high which should also be reflected upon higher chances of mutation. This in turn would mean mutations in the course of evolution resulting in greater propagation and higher allele frequency. No significant correlation was observed in this study. Differential evolutionary advantage of mutations influencing their propagation among population might be the reason diluting the hypothesized effect there by explaining lack of any correlation. Our result is also confirmed by a very recent report[28].

Next it was aimed at investigating the distribution of CG-SNPs in different type of regions in genome and compares it with the distribution of total SNPs in the same regions. It was observed that both CG-SNPs as well as total SNPs are over represented in 5' UTR when compared to Exons, Introns and 3' UTRs. The other three regions had marginal difference including expected minimum values for Exons that are containing the most sensitive coding regions. It is an interesting observation that 5' UTRs are the least sensitive to mutations as far as this analysis is based on SNPs. Further work is highly desirable to prove further into this effect. On the other hand there was not much difference in the distributions between total SNPs and CG-SNPs except for a marginally higher ratio of CG-SNP to total SNP frequencies for 3' UTRs. It may implies that 3' UTRs have evolved with higher preference for As and Ts in the regions. It is wondering to think if it has any evolutionary relationship with poly A tail linked to the 3' UTR of the mRNAs in terms of structural stability of RNA or some functional roles.

The final experiment was aimed at finding the consensus sequences flanking the CpGs in CG-SNPs with either very high or very low A or T allele frequencies. The master consensus sequences obtained are (5'- KVCACGBGCS -3') and (5'-MHBVCGGRRH -3') that are significantly distinct from each other. This was yet another attempt to related DNA methylation levels with allele frequencies. The idea was encouraged when the allele frequency distribution plot was found to be bimodal in nature. This has characteristic similarity with the CpG methylation distribution in a report relating flanking sequences with DNA methylation propensity (Fig 6.1)[11]

Since the consensus sequences obtained are not very sharply defined, it is difficult to compare them with the reported consensus sequences. Further elaborating the magnitude of data and finer statistical tools may help to arrive at a better conclusion.

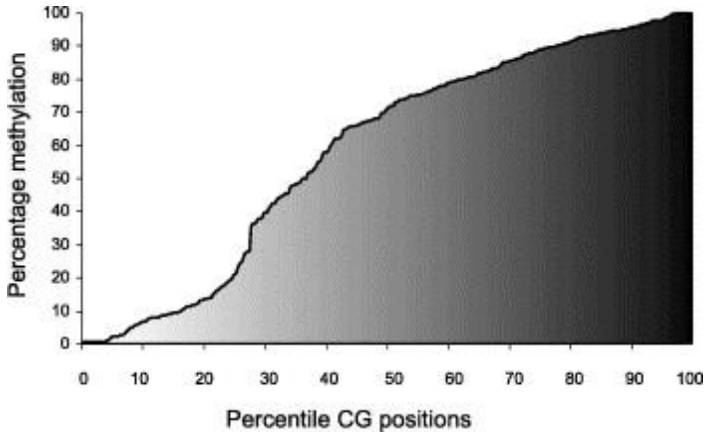

Figure 6.1 Methylation levels at various CG sites in the epigenomic data. The CG sites were arranged in the order of increasing methylation levels.

Source: Handa & Jeltsch, 2005

CHAPTER7

CONCLUSION

7. CONCLUSION

Epigenetics has been playing important role in explaining several biological phenomena in higher eukaryotes while SNPs have emerged as powerful tools in genomics in recent times. Interesting the two worlds are also share an interface. In the present study this interface has been attempted to be studied using three different approaches. It may be concluded from this study that significantly higher representation of SNPs in 5' UTRs is an interesting observation and deserves to be explored further with larger datasets and improved approaches. On the other hand a clear bimodal distribution has been observed in the allele frequencies of CG-SNPs which have resulted also in generation of distinct consensus sequences for extreme allele frequencies.

CHAPTER8

REFERENCES

8. REFERENCES:

1. Waddington, C. (1940). The genetic control of wing development in Drosophila. . *J. Genet.* **41**, 75-80
2. Issa, J. P. (2000). CpG-island methylation in aging and cancer. *Curr. Top. Microbiol. Immunol.* **249**, 101-118.
3. Riggs, A. D. (1975). X inactivation, differentiation, and DNA methylation. *Cytogenet. Cell Genet.* **14**, 9-25
4. Herman, A., Gowher, H. & Jeltsch, A. (2004). Biochemistry and biology of mammalian DNA methyltransferases *CMLS Cell.Mol.Life Sci.* **61**, 2571-2587.
5. Kahang, L. & Shapiro, L. (2001). The CcrM DNA methyltransferases of Agrobacterium tumefaciens is essential, and its activity is cell cycle regulated *T– J Bacteriol.* **183**, 3065-3075.
6. Zucker, K. E., Riggs, A. D. & Smith, S. S. (1985). Purification of human DNA (cytosine-5-)-methyltransferase. *J. Cell. Biochem.*, 337–349.
7. Flynn, J., Azzam, R. & Reich, N. (1998). DNA binding discrimination of the murine DNA cytosine-C5 methyltransferase. *J. Mol. Biol.* **279** 101–116.
8. Pradhan, S., Bacolla, A., Wells, R. D. & Roberts, R. J. (1999). Recombinant human DNA (cytosine-5) methyltransferase. I. Expression, purification, and comparison of de novo and maintenance methylation. *J. Biol. Chem.* **274**, 33002–33010.
9. E., L. (2002). Chromatin modification and epigenetic reprogramming in mammalian development. *Nat Rev Genet.* **3**, 662-673.
10. Okano, M., Xie, S. & Li, E. (1998). Cloning and characterization of a family of novel mammalian DNA (cytosine-5) methyltransferases. *Nature Genet.* **19**, 219–220.

11. Handa, V. & Jeltsch, A. (2005). Profound Flanking Sequence Preference of Dnmt3a and Dnmt3b Mammalian DNA Methyltransferases Shape the Human Epigenome. *J. Mol. Biol.* **348** 1103–1112.
12. Sved, J. & Bird, A. P. (1990). The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc Natl Acad Sci USA* **87**, 4692–4696.
13. Moon, K. J. F. a. W.-J. (2005). CpG Mutation Rates in the Human Genome Are Highly Dependent on Local GC Content *Mol. Biol. Evol.* **22**, 650–658.
14. Razin , A. & Riggs, A. D. (1980). DNA methylation and gene function. *Science* **210**, 604–610.
15. P.Bird, A. (1980). DNA methylation and the frequency of CpG in animal. . *Nucleic Acids Res.* **8**, 1499-1504.
16. Gardiner-Garden, M. & Frommer, M. (1987). CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282.
17. Jones, P. A. & Takai, D. (2001). The role of DNA methylation in mammalian epigenetics. *Science* **293**, 1068–1070.
18. Daekwan, S., Cizhong , J. & Zhongming, Z. (2006). A novel statistical method to estimate the effective SNP size in vertebrate genomes and categorized genomic regions. *BMC Genomics* **7**, 329.
19. Zhao, Z. & Boerwinkle, E. (2002). Neighboring nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome Res.* **12**, 1679–1686.
20. Zhang, F. & Zhao, Z. (2005). SNPNB: analyzing neighboring-nucleotide biases on single nucleotide polymorphisms (SNPs). *Bioinformatics* **21**, 2517-2519.
21. Hughes, A., Packer, B., Welch, R., Bergen, A., Chanock, S. & Yeager, M. (2003). Widespread purifying selection at polymorphic sites in human protein-coding loci. *Proc Natl Acad Sci USA* **M 100**, 15754-15757.

22. Tomso, D. J. & Bell, D. A. (2003). Sequence Context at Human Single Nucleotide Polymorphisms: Overrepresentation of CpG Dinucleotide at Polymorphic Sites and Suppression of Variation in CpG Islands. *J. Mol. Biol.* **327**, 303–308.

23. Zavolan, M. & Kepler, T. B. (2001). Statistical inference of sequence-dependent mutation rates. *Curr. Opin. Genet. Dev.* **11**, 612–615.

24. Risch, N. & Merikangas, K. (2001). The future of genetic studies ofcomplex human diseases. . *Science* **273**, 1516–1517.

25. Huang, Q., Morrison, A. C. & Boerwinkle, E. (2001). Linkage disequilibrium structure and its impact on the localization of a candidate functional mutation. *Genet. Epidemiol* **21**, S620–S625.

26. Marth, G., Yeh, R., Minton, M., Donaldson, R., Li, Q., Duan, S., Davenport, R., Miller, R. D. & Kwok, P.-Y. (2001). Single-nucleotide polymorphisms in the public domain: How useful are they? *Nat. Genet.* **27**, 371–372.

27. SNP fact sheet from Human Genome Project Information available at link: http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml

28. Clemens, W., Finja, C., Georg, H., Johannes, E. & Florian, M. (2012). Linking the Epigenome to the Genome: Correlation of different features to DNA Methylation of CpG Islands *PLoS ONE*, ***7,*** e35327