Comparison of CpG islands in Genomes with different evolutionary lineage

A Thesis

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

Master of Science

In Biotechnology

By

Mehakpreet Kaur

(Reg no: 302001018)

Under the Supervision of

Dr. Vikas Handa

Assistant Professor

Department of Biotechnology

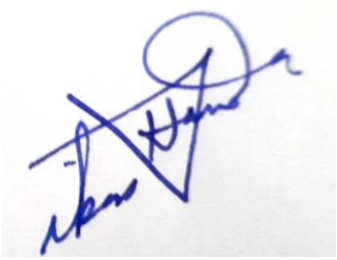
Thapar Institute of Engineering and Technology,

Patiala, Punjab India

June 2022

Certificate

This is to certify that the thesis entitled, **Comparison of CpG islands in Genomes with different evolutionary lineage** being submitted by Mehakpreet Kaur (Reg. No.302001018), in partial fulfilment of the requirements for the award of the degree of Master of Science in Biotechnology, Thapar Institute of Engineering and Technology, Patiala, Punjab is a bonafide work carried out under the guidance and conception of Dr. Vikas Handa and that no part of this thesis has been submitted for the award of any other degree.



Date: 27/07/2022

Dr. Vikas Handa

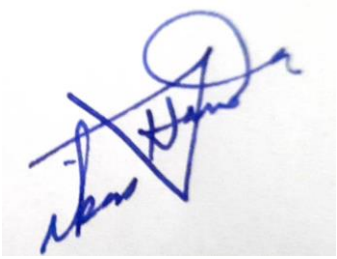
Supervisor

Candidate Declaration

I hereby certify that the project work entitled, **Comparison of CpG islands in Genomes with different evolutionary lineage** in a partial fulfilment of the requirements for the award of the degree of Master of Science in Biotechnology and submitted is an authentic record of my work carried out during the period January 2022 to June 2022 under the guidance of Dr. Vikas Handa, Assistant professor, Department of Biotechnology, Thapar Institute of Engineering and Technology, Patiala, Punjab, India.

Date: 27/07/2022 Mehakpreet Kaur

This is to certify that the above statement made by the student is correct to the best of our knowledge and belief.



Date: 27/07/2022 Dr. Vikas Handa

Supervisor

ACKNOWLEDGEMENT

I am grateful to each one who has helped me throughout the entire project for its successful completion. First, thanks to Almighty for giving me strength and support so that the project could be completed peacefully.

With great reverence, I express my warmest feeling with a deep sense of gratitude to **Dr. Vikas Handa, Assistant professor,** Department of Biotechnology, Thapar Institute of Engineering and Technology who agreed to take upon and guided for this dissertation. I have no words to express my heartfelt thanks to him for his illuminating guidance, unfailing encouragement, supervision, and keen interest during this dissertation.

I would like to express my heartfelt respect to **Dr. M S Reddy,** Head of Department of Biotechnology, Thapar Institute of Engineering and Technology, Patiala for his kind suggestions and foresightedness.

I have immense gratitude towards my parents, Mr. Surinder Pal Singh and Mrs. Maninder Kaur for their affection and faith. Also I would like to thank my siblings Harman and Komal for giving me strength and loving support.

I would like to acknowledge my close friend Ravneet who stood by my side during the tough times. Also I would like to acknowledge my lab friend Tamnayee Basu who helped me in learning small things with great perfection. The whole credit goes to all the people who had their unshakeable faith in me which has always motivated me.

Mehakpreet Kaur

Table of Content

List of figures

List of tables

List of Abbreviations

DNA	Deoxyribonucleic acid
U	Uracil
T	Thymine
G+C	Guanine + Cytosine
A+T	Adenine + Thymine
CpG	C-phosphate-G
5mC	5-Methyl-Cytosine
4mC and m6A	4-Methyl-Cytosine and 6-Methyl-Adenine
SNPs	Single nucleotide polymorphisms
OBS/EXP	Observed/Expected
TFBS	transcription factor binding sites
G4	G quadruplex
oCGIs	Orphan CGIs
MLL	mixed-lineage leukaemia protein
KDM	lysine demethylase
TET	Ten-eleven translocation
CFP1	CxxC finger protein
PRC1 OR PRC2	Polycomb repressive complex
CBP/P300	CREB-binding protein
SETD1A	SET domain containing 1A
CGI	CpG island
Non-CGI	Non CpG island

ABSTRACT

The eukaryotic genome is said to be the most complex genome since it contains some of the most complicated modification mechanisms. Complexity has risen over time as a result of evolution and adaptation. The features of specific DNA segments are altered by several mechanisms and mutations, leading to evolution. DNA methylation followed by spontaneous deamination is one of the most important epigenetic modifications that distinguish certain regions known as CpG islands (CGI) from the rest of the genome. The DNA regions known as CpG islands have several unique qualities and properties that make them structurally as well as functionally important. Five genes were randomly selected for this study's assessment of CpG islands with regard to their higher eukaryotic species. The CGI and non-CGI sequences of these genes were identified and analysed for six different vertebrates. Based on CGI characteristics, the sequences were compared by constructing dendrograms. Comparison of dendrograms with phylogenetic trees showed poor overlap. The CGI sequences were compared based on their length, CG gaps, CG obs/exp and GC percentage. Some genes, like the Sox4 gene display very long CGIs and shortest CG gaps among all the genes studied in this work. It was observed that there is an inverse correlation between the length of CGIs and the gaps between the neighbouring CGs.

Keywords - DNA methylation, deamination, CGI sequences, non-CGI sequences, CG gaps, CG obs/exp and GC percentage

CHAPTER 1: INTRODUCTION

The genome can be widely described as the genetic blueprint of an organism. The important information that is necessary for the growth and development of the fully mature organism is passed to the next generation through the genome. The base of the genome is formed by the long sequence of bases that form DNA, chromosome which is a highly organised and compact structure. Genes are defined as small regions of the chromosome which encode RNA and proteins for the cell.

The higher eukaryotic genomes are several hundred million to more than a billion base pairs in size. The genetic information is stored in the genomes in the form of base sequences of DNA. Additionally, bases like 5-methylcytosine or hydroxyl methylcytosine may occasionally be found in an organism. DNA is made up of two strands in prokaryotic and eukaryotic organisms, as well as many viruses, with specific base pairs (A + T) or (G + C) at each location along the complementary strands (Johnson, 1985).

Research reveals that the eukaryotic genome is considered to be more complex due to gene expression and post-translational modifications as compared to the prokaryotic genome. In eukaryotic genomes, repeated elements can be very prevalent and are broadly classified into two large families i.e. "tandem repeats" and "dispersed repeats." Tandem repeats are the sequences that are clustered in the genome and split into two orientations, direct repeats (head-to-tail pattern) and inverted repeats (head-to-head). Satellite DNA (satellites, minisatellites, and microsatellites), tandem gene paralogues, and ribosomal RNA (rRNA) genes are the three main groups (subfamilies) of tandem repeats. On the other hand, Dispersed repeats are interspersed within the genome and the three other subfamilies of dispersed repeats include genes that encode transfer RNA (tRNA), transposons, retro elements, and retroviruses, as well as gene paralogues and gene families, LINEs, and SINEs. These repetitive sequences appear to have vital biological activities based on their high frequency and richness as well as evolutionary preservation (Sperling & Li, 2013).

DNA methylation is the basic mechanism of epigenetic inheritance. It is an epigenetic modification that does not change the DNA sequence but has an influence on gene activity (Xie *et al*., 2009). The mechanism involves is the enzymatic transfer of the methyl group from the S-Adenosyl methionine (SAM) to the fifth carbon of a cytosine residue to form ^{5m}C in vertebrates (Ciechomska *et al*., 2019) also ^{4m}C and m^6A (widely in prokaryotes) (Rodriguez *et al*., 2022).

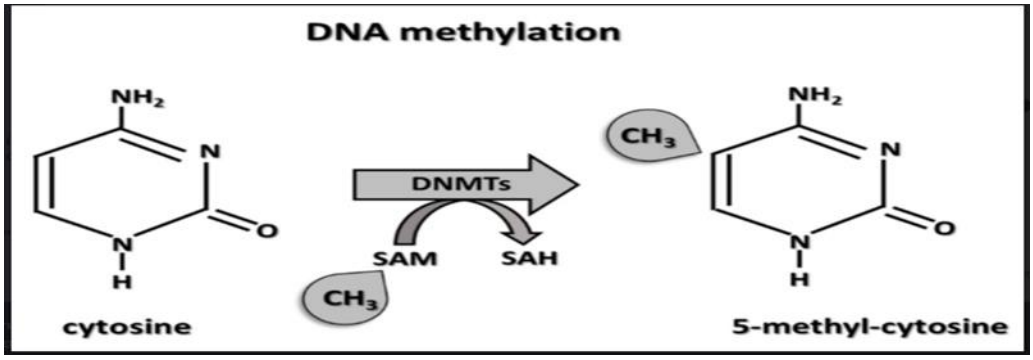


Fig.1- Mechanism of DNA methylation (Ciechomska *et al., 2019*)

Enzymes that catalyse the reaction is the family of DNA methyl transferases (DNMTs). Dnmt3a and Dnmt3b are known as *de novo* Dnmt as they establish new methylation patterns for unmodified DNA. On the other hand, dnmt1 is known as maintenance Dnmt as it maintains the methylation pattern in the newly synthesised daughter strand during DNA replication. All the three Dnmts are extensively involved in the development of an embryo but the expression keeps on reducing when the cells reach the terminal differentiation stage and finally become stable at the post-mitotic stage.

The compositional heterogeneity represented by CpG islands is distinct when compared to the rest of the genome as it exhibits unique characteristics. Originally known as HpaII Tiny Fragment (HTF) islands, they were discovered to be brief sections of mammalian DNA that were packed with HpaII restriction endonuclease sites. Short areas of 1-2 kb in size, combined makeup about 2% of the mammalian genome. Some of the preliminary criteria that distinguish some specific regions (CpG) from the bulk genome are GC-rich, unmethylated, and do not exhibit any CpG suppression. Contrary to what would be expected from its base makeup, bulk genomic DNA has a substantially lower GC content, is methylated at CpG, and has a much lower frequency of the CpG dinucleotide. All housekeeping genes and a significant number of genes with a tissue-restricted pattern of expression include CpG islands at their 5′ ends (Craig & Bickmore 1994).

Later on, the detailed and well-defined description was given by Gardiner-Garden and former i.e. CpG islands are the stretch of DNA which remains unmethylated and are found in regions of the methylated genome associated with the 5′ ends of all housekeeping genes, many tissue-specific genes and with 3′ ends of some tissue-specific genes (Gardiner-Garden & Frommer, 1987). They have high G+C content and a relatively higher abundance of GC dinucleotides (Gardiner-Garden & Frommer, 1987) and also they play a major role in regulating gene expression which includes cell type-specific expression, gene silencing, genomic imprinting and X- chromosome inactivation (Takai & Jones, 2002). Also for the scanning of CpG islands

in the genome parameters were given i.e. length of CGI should be 200bp, GC content at least 50% and CG_obs/exp should be 0.60. These were not applicable to all the genes and organisms also had some limitations so Takai and Jones amended the parameters in which length should be 500bp, GC content at least 55% and CG_Obs/exp should be 0.65.

CpG islands appear to have evolved to increase gene expression by controlling chromatin shape and transcription factor binding. Nucleosomes are tiny, packed pieces of DNA that are wrapped around histone proteins regularly. The DNA becomes less permissive for gene expression as it becomes more closely linked with histone proteins. CpG islands have fewer nucleosomes than other DNA regions, which is one of its most distinguishing characteristics (Tazi, 1990). CpG islands are connected with a small number of nucleosomes, and these nucleosomes frequently include histones with modifications that aid gene expression. CpG islands are believed to improve binding to transcriptional start sites since many transcription factor binding sites are GC rich. CpG islands improve DNA accessibility and transcription factor binding despite the lack of shared promoter elements (Tazi, 1990).

Followed by DNA methylation Cytosine is vulnerable to mutation and hence undergoes spontaneous deamination giving rise to Uracil and 5-Methyl-Cytosine gives rise to Thymine. The Repair mechanism of U (back to C) by uracil DNA glycosylases is much more efficient than the repair of T by the thymine DNA glycosylases. Over time, ^{5m}CpGs mutate to TpGs and CpAs but CpGs are repaired and remain in the genome (Xie *et al.,* 2009).

Nucleotide substitution takes place in somatic tissues but is not passed on to the progeny. In contrast, a methylated cytosine's hydrolytic deamination in germ cells is a heritable mutation that could eventually result in polymorphism frequencies and methylation-associated SNP. Changes to a single nucleotide called SNPs (single nucleotide polymorphisms) produce various sequences called alleles, with the less common allele having an abundance of 1% or greater (Brookes, 1999). They are the most prevalent types of genetic variation seen in humans. Several studies have been conducted to identify sets of SNPs that can be used as markers for multifactorial disease propensity or mapping linkage disequilibrium across all genomes (Souza *et al.,* 2020).

Compared to earlier transitions, the mutation rate from 5mC to T in humans is 10 to 50 times higher. (Fryxell & Moon, 2004). As a result, hyper mutability of CpG is the major cause of human genetic diseases and many of the somatic mutations that lead to cancer (Cooper and Youssoufian 1988; Cooper and Krawczak 1993). Thus estimation of mutation rates that result from the deamination of methylated cytosine could be done by using SNP data (Xie *et al.,* 2009)

CHAPTER 2: REVIEW OF LITERATURE

The genome is the complex of genetic information which is stored inside the nucleus of the cell in the form of a DNA base sequence. After the DNA sequence has been decoded, computational tools and algorithms can be used to access its structural and functional features. The genome's features and functions are easily altered by its base composition. Point mutations, deletions, insertions, and translocations are all examples of changes. In contrast, DNA methylation is a significant epigenetic mechanism that involves the direct chemical change of the DNA (Moore *et al*., 2012).

2.1 Methylated CpGs as a mutational hotspot

CpG deficit is linked to DNA methylation; in human DNA, where the percentage of (G+C) is 0.4, we would anticipate CpG to occur with a frequency of 0.2 x 0.2 = 0.04, but the observed frequency is roughly 0.008. (Bird, 1980) therefore 5mC change would result in the loss of two CpGs and the gain of one TpG and one CpA, making mCpG a mutational hotspot. (Bird, 1980) also, CpG dinucleotide is statistically underrepresented, according to the study, and its extent is negatively associated with GC concentration (Bird 1980).

2.2 Rate of DNA methylation influenced by flanking sequences

The influence of flanking sequence on the catalytic activity of the Dnmt3a and Dnmt3b de novo DNA methyl transferases. High (5'-CTTGCGCAAG-3') and low (5'-TGTTCGGTGG-3') levels of methylation in human genomic DNA are marked by these sequences of up to +/-four base pairs surrounding the core CG site. Furthermore, AT-rich sides are favoured by GC-rich flanks. These experimental choices match the methylation patterns in the genome. As a result of the expanded experimental investigation, the methylation rates of the consensus sequences for high and low levels of methylation in the genome differed by more than 500-fold (Handa & Jeltsch, 2005).

2.3 Algorithms used to scan CGIs in the bulk genome

Gardiner-Garden and Frommer proposed the first technique for scanning CGIs in DNA sequences in 1987. This approach has been widely used in several assessments of CGIs in single genes or small groups of genomic sequences. It searches for GC content of at least 50%, Obs_{CpG}/Exp_{CpG} of at least 0.60, and a length of at least 200bp using three parameters. However, because many repetitions (e.g., Alu), which are prevalent in vertebrate genomes, also fit the parameters of this approach, the number of CGIs is significantly inflated. Takai and Jones used

a systematic examination of the three parameters of Gardiner-Garden and Frommer's algorithms to come up with an ideal set of parameters i.e. 55 per cent GC content, 0.65 Obs_{CpG}/Exp_{CpG}, and 500bp length to tackle this problem (Han & Zhao, 2008).

2.4 Sequence and chromatin feature of CGIs

Elevated CpG density and GC content, transcription factor binding sites (TFBS), and G quadruplex (G4) DNA sequences are all characteristics of CGIs. CGIs can swing between active/poised and repressive chromatin states depending on the activity of the gene they regulate. The complement of transcriptional activators (CBP/P300, SETD1, CFP1, TET1, KDM2A, and RNAP2) and repressors (PRC1, PRC2, and KDM2B) localised to CGIs influences these states. CGIs can be permanently silenced through DNA methylation (^{5m}C) and methyl-CpG binding proteins (MBDs) in extreme situations, such as imprinting control regions (ICRs) or cancer-testis antigen gene (CTA) promoters, a status maintained by continual DNA targeting.

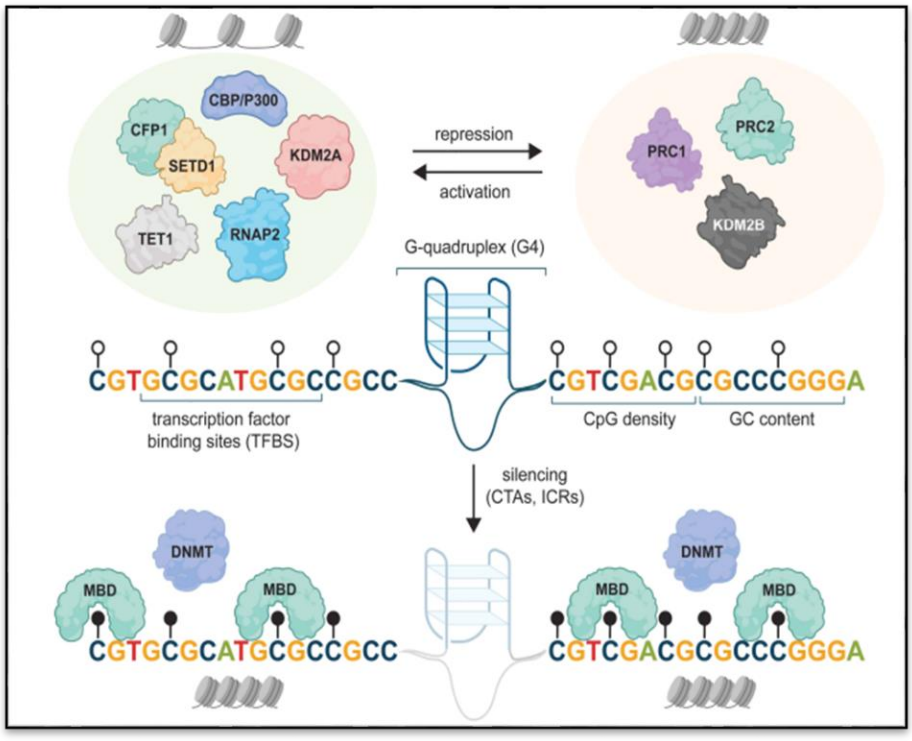


Fig-2: Sequence and chromatin feature of CGIs (Angeloni & Bogdanovic, 2021)

2.5 Diverse set of regulatory functions of CGIs in the genome

Studies reveal that CGIs are most typically examined in the setting of promoters, however, several studies have also shown that CGIs can regulate gene expression in a variety of genomic situations. Orphan CGIs (oCGIs) are nearly half of all discovered CGIs that are found in intergenic and intragenic areas. In some organisms like zebrafish, frog, and mouse embryos, orphan CGIs (oCGIs) overlap with developmental enhancers connected to critical developmental pathways. During the vertebrate phylotypic phase, these enhancers become developmentally activated by active DNA demethylation mediated by Ten-eleven translocation (TET) enzymes, gaining characteristic enhancer chromatin markers like H3K4me1 and H3K27ac. As a result, CGIs have a diverse set of regulatory functions in the genome, some of which have survived millions of years of divergent evolution (Angeloni & Bogdanovic, 2021).

2.6 Divergence of CpG island promoters, a consequence or cause of evolution

In previous research Long CGIs be particularly abundant in polycomb group (PcG) protein binding sites. The acquisition of long CpG islands may have facilitated the recruitment of polycomb complexes in the promoters of genes associated with placenta development, and thus contributed directly as a "cause" for the emergence of the placenta in mammals because loss or gain of cis-regulatory elements is known to have profound effects on gene expression (Sharif *et al.*, 2010).

CHAPTER 3: AIM OF THE STUDY

Comparison of CpG island sequences and Non-CpG island sequences among higher Eukaryotic organisms of different Phyla.

CHAPTER 4: MATERIAL AND METHODS

4.1 Selection of genes

A set of five genes were randomly selected with respect to their higher Eukaryotic organisms of the same Phylum.

Table 1: Genes with the corresponding organisms, chromosomes, and Accession number.

S. No	Gene	Organism	Chromosomes	Accession number
1	**Dnmt1**	*Homo sapiens*	19	NC_000019.10
2		*Pan troglodytes*	19	NC_036898.1
3		*Mus musculus*	9	NC_000075.7
4		*Gallus gallus*	30	NC_052561.1
5		*Xenopus tropicalis*	3	NC_030679.2
6		*Danio rerio*	3	NC_007114.7
1	**Gadd45a**	*Homo sapiens*	1	NC_000001.11
2		*Pan troglodytes*	1	NC_036879.1
3		*Mus musculus*	6	NC_000072.7
4		*Gallus gallus*	8	NC_052539.1
5		*Xenopus tropicalis*	4	NC_030680.2
6		*Danio rerio*	6	NC_007117.7
1	**Nanog**	*Homo sapiens*	12	NC_000012.12
2		*Pan troglodytes*	12	NC_036891.1
3		*Mus musculus*	6	NC_000072.7
4		*Gallus gallus*	1	NC_052532.1
5		*Danio rerio*	24	NC_007135.7
1	**Sox4**	*Homo sapiens*	6	NC_000006.12
2		*Pan troglodytes*	6	NC_036885.1
3		*Mus musculus*	13	NC_000079.7
4		*Gallus gallus*	2	NC_052533.1
5		*Xenopus tropicalis*	6	NC_030682.2
6		*Danio rerio*	19	NC_007130.7
1	**ZEYVE16**	*Homo sapiens*	5	NC_000005.10
2		*Pan troglodytes*	5	NC_036884.1
3		*Mus musculus*	13	NC_000079.7
4		*Gallus gallus*	Z	NC_052572.1
5		*Xenopus tropicalis*	1	NC_030677.2
6		*Danio rerio*	10	NC_007121.7

4.2 Gene Sequences and other Gene Information

- FASTA sequence of genes with respect to their higher eukaryotic organisms were retrieved from NCBI (https://www.ebi.ac.uk/Tools/seqstats/emboss_newcpgreport/) using the Accession number.
- CpG Islands were searched using EMBOSS Newcpgreport (https://www.ebi.ac.uk/Tools/seqstats/emboss_newcpgreport/) by taking three parameters GC content, $Obs_{CpG}/Exp_{CpG,}$ and length.

4.3 Algorithms for the Identification of CGIs

Using two techniques, we scanned CGIs in genomic sequences. First, we used the Takai and Jones technique, which is designed to find CGIs in human and other mammalian genomes connected with the 5′ ends of genes. Its search criteria are: 55 percent GC content, 0.65 Obs_{CpG}/Exp_{CpG}, and 500bp length. Second, we employed the Gardiner-Garden and Frommer approach, which requires a GC content of at least 50%, an Obs_{CpG}/Exp_{CpG} ratio of at least 0.60, and a length of at least 200bp.

4.4 Sequence Manipulation Suite: DNA Stats

DNA Stats is the web server which calculates the number of occurrences of each residue in the sequence and is used in both the cases CGI sequences as well as nonCGI sequences. The result is used to find the probability of each selected residue for further analysis. (https://www.bioinformatics.org/sms2/dna_stats.html)

4.5 The statistical tests applied in the analysis

The data was retrieved and various statistical (i.e., MID and Probability), as well as Analytical tools, were applied for further analysis.

CG_OBSERVED/EXPECTED: - The ratio of the observed value of CG to the Probability of (CG) multiple by a total of all four bases.

CG_OBS/EXP: - Instead of using the CG obs/exp value we introduced a new index, CG obs/exp + 10. *

(CG_O-E)/ (TG_O+CA_O-TG_E-CA_E): - Instead of using the (CG_O-E)/ (TG_O+CA_O-TG_E-CA_E) we introduced a new index, (CG_O-E)/ (TG_O+CA_O-TG_E-CA_E) + 10.*

*This was done to get the final value to be greater than zero.

4.6 Dendrogram construction and Analysis

Dendro UPGMA is a web server that uses the UPGMA (Unweighted Pair Group Method with Arithmetic Mean) to create dendrograms. A dendrogram is a diagram that shows how distinct clusters formed by hierarchical clustering are arranged. It's made up of U-shaped lines that connect data points in a tree hierarchy. The distance between two connected data points is represented by the length of lines. (http://genomes.urv.cat/UPGMA/)

4.7 Multiple Sequence Alignment and Phylogenetic Analysis

The gene sequences with respect to the particular list of genes selected were run in the CLUSTAL W web server also phylogenetic trees were constructed for further analysis. (https://www.genome.jp/tools-bin/clustalw)

CHAPTER 5: RESULT

5.1 Comparison of the bases of frequency of bases and dinucleotides.

Several statistical and analytical tools were used to further analyse the gene sequences that have been obtained from the NCBI with regard to their organism. The frequency of bases and dinucleotides in CGI and non-CGI sequences is then determined using the results.

Table-2-a: Frequency of bases and dinucleotides in Dnmt1 gene

ORGANISM & PATTERN	*Homo sapiens*			*Pan troglodytes*			*Mus musculus*			*Gallus gallus*			*Xenopus tropicalis*			*Danio rerio*		
	Gene	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**
G	15417	136	15281	15605	186	15419	13756	133	13623	5998	116	5882	3564	74	3490	3346	136	3210
A	13649	32	13617	13653	88	13565	11508	49	11459	4482	115	4367	4516	39	4477	4365	106	4259
T	16940	59	16881	16981	152	16829	14691	79	14612	5406	93	5313	4881	66	4815	5214	126	5088
C	15602	112	15490	15821	225	15596	12729	130	12599	7364	244	7120	3212	82	3130	3085	124	2961
TG	5188	17	9	5202	41	5161	4930	34	4896	1773	17	1756	1316	26	1290	1389	39	1350
CG	1348	22	14	1428	59	1369	925	38	887	844	44	800	221	20	201	324	37	287
CA	4357	7	11	4358	34	4324	3622	22	3600	1731	44	1687	1061	8	1053	1042	28	1014
TOTAL (G+A+T+C)		**339**	**61269**		**651**	**61409**		**391**	**52293**		**568**	**22682**		**261**	**15912**		**492**	**15518**
P(G)		0.40	0.24		0.28	0.25		0.34	0.26		0.20	0.25		0.28	0.21		0.27	0.20
P(A)		0.09	0.22		0.13	0.22		0.12	0.21		0.20	0.19		0.14	0.28		0.21	0.27
P(T)		0.17	0.27		0.23	0.27		0.20	0.27		0.16	0.23		0.25	0.30		0.25	0.32
P(C)		0.33	0.25		0.34	0.25		0.33	0.24		0.43	0.31		0.31	0.19		0.25	0.19
P(TG)		0.07	0.06		0.06	0.06		0.06	0.07		0.03	0.06		0.07	0.06		0.07	0.06
P(CA)		0.03	0.05		0.04	0.05		0.04	0.05		0.08	0.06		0.04	0.05		0.05	0.05
P(CG)		0.13	0.06		0.09	0.06		0.11	0.06		0.08	0.08		0.08	0.03		0.07	0.03
EXP(TG)		23.6	4210.2		43.4	4225.5		26.8	3806.6		18.9	1377.7		18.7	1056.0		34.8	1052.4

ORGANISM & PATTERN	*Homo sapiens*			*Pan troglodytes*			*Mus musculus*			*Gallus gallus*			*Xenopus tropicalis*			*Danio rerio*		
	Gene	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**
EXP(CA) EXP(CG)		10.5 44.9	3442.6 3863.3		30.4 64.2	3445.0 3915.9		16.2 44.2	2760.8 3282.2		49.4 49.8	1370.8 1846.3		12.2 23.2	380.6 586.5		26.7 34.2	812.6 612.5
CG_Obs/ Exp+10		10.80	10.34		10.91	10.35		10.85	10.27		10.88	10.43		10.86	10.29		11.07	10.46
(CG_O-E)/ (TG_O+ CA_O-TG_E-CA_E)+10		9.08	8.61		5.43	8.59		9.51	8.75		10.78	8.49		8.92	8.08		10.49	9.34

Table-2-b: Frequency of bases and dinucleotides in Gadd45a gene

ORGANISM & PATTERN	*Homo sapiens*			*Pan troglodytes*			*Mus musculus*			*Gallus gallus*			*Xenopus tropicalis*			*Danio rerio*		
	Gene	**CGI**	**Non CGI**	**Gene**	**CGI**	**Non CGI**	**Gene**	**CGI**	**Non CGI**	**Gene**	**CGI**	**Non CGI**	**Gene**	**CGI**	**Non CGI**	**Gene**	**CGI**	**Non CGI**
G	907	195	712	952	198	754	625	279	346	474	246	228	1039	61	978	873	82	791
A	726	80	646	752	78	674	564	136	428	339	116	223	1265	53	1212	1191	59	1132
T	758	88	670	788	85	703	578	124	454	316	122	194	1470	33	1437	1313	71	1242
C	743	195	548	770	193	577	545	262	283	455	246	209	907	45	862	733	60	673
TG	242	44	198	258	44	214	179	63	116	105	54	51	423	12	411	366	31	335
CG	152	57	95	152	56	96	110	81	29	116	67	49	61	10	51	91	14	77
CA	170	27	143	181	26	155	124	37	87	96	36	60	324	15	309	291	18	273
TOTAL (G+A+T+C)		**558**	**2576**		**554**	**2708**		**801**	**1511**		**730**	**854**		**192**	**4489**		**272**	**3838**
P(G) P(A) P(T)		0.34 0.14 0.15	0.27 0.25 0.26		0.35 0.14 0.15	0.27 0.24 0.26		0.34 0.17 0.15	0.22 0.28 0.30		0.33 0.15 0.16	0.26 0.26 0.22		0.31 0.27 0.17	0.21 0.27 0.32		0.30 0.21 0.26	0.20 0.29 0.32

ORGANISM & PATTERN	*Homo sapiens*			*Pan troglodytes*			*Mus musculus*			*Gallus gallus*			*Xenopus tropicalis*			*Danio rerio*		
	Gene	**CGI**	**Non CGI**	**Gene**	**CGI**	**Non CGI**	**Gene**	**CGI**	**Non CGI**	**Gene**	**CGI**	**Non CGI**	**Gene**	**CGI**	**Non CGI**	**Gene**	**CGI**	**Non CGI**
P(C)		0.34	0.21		0.34	0.21		0.32	0.18		0.33	0.24		0.23	0.19		0.22	0.17
P(TG)		0.05	0.07		0.05	0.07		0.05	0.06		0.05	0.06		0.05	0.07		0.07	0.06
P(CA)		0.05	0.05		0.04	0.05		0.05	0.05		0.05	0.06		0.06	0.05		0.04	0.05
P(CG)		0.12	0.05		0.12	0.05		0.11	0.04		0.11	0.06		0.07	0.04		0.05	0.05
EXP(TG)		30.7	185.1		30.3	195.7		43.1	103.9		41.11	51.7		10.48	313.0		21.4	255.9
EXP(CA)		27.9	137.4		27.1	143.6		44.4	80.1		39.09	54.5		12.42	232.7		13.0	198.5
EXP(CG)		68.1	151.4		68.9	160.6		91.2	64.8		82.90	55.8		14.30	187.8		15.6	217.7
CG_Obs/Exp		10.83	10.62		10.81	10.59		10.88	10.44		10.80	10.87		10.69	10.27		10.89	10.35
(CG_O-E)/(TG_O+CA_O-TG_E-CA_E)		9.09	6.92		8.95	7.81		9.16	8.10		8.37	8.53		8.95	9.21		9.88	9.08

Table-2-c: Frequency of bases and dinucleotides in NanoG gene

ORGANISM & PATTERN	*Homo sapiens*			*Pan troglodytes*			*Mus musculus*			*Gallus gallus*			*Danio rerio*		
	Gene	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**
G	2095	78	2017	1241	35	1206	1632	50	1582	1078	47	1031	973	66	907
A	2518	49	2469	1548	26	1522	1816	8	1808	1147	27	1120	1489	61	1428
T	2965	72	2893	1809	18	1791	2166	33	2133	1349	28	1321	1638	44	1594
C	2167	96	2071	1348	29	1319	1531	18	1513	1010	56	954	980	91	889
TG	740	19	721	461	6	455	608	27	581	411	11	400	379	11	368
CG	142	19	123	73	7	66	90	7	83	78	12	66	115	22	93
CA	643	17	626	419	9	410	435	5	430	309	12	297	354	27	327
TOTAL (G+A+T+C)		**295**	**9450**		**108**	**5838**		**109**	**7036**		**158**	**4426**		**262**	**4818**
P(G)		0.26	0.21		0.32	0.20		0.45	0.22		0.29	0.23		0.25	0.18
P(A)		0.16	0.26		0.24	0.26		0.07	0.25		0.17	0.25		0.23	0.29

ORGANISM & PATTERN	*Homo sapiens*			*Pan troglodytes*			*Mus musculus*			*Gallus gallus*			*Danio rerio*		
	Gene	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**
P(T) P(C)		0.24 0.32	0.30 0.21		0.16 0.26	0.30 0.22		0.30 0.16	0.30 0.21		0.17 0.35	0.29 0.21		0.16 0.347	0.33 0.185
P(TG) P(CA) P(CG)		0.06 0.05 0.08	0.06 0.05 0.04		0.05 0.06 0.08	0.06 0.05 0.04		0.13 0.01 0.07	0.06 0.05 0.04		0.05 0.06 0.10	0.07 0.05 0.05		0.04 0.08 0.08	0.06 0.05 0.03
EXP(TG) EXP(CA) EXP(CG)		19.04 15.95 25.38	617.4 541.0 442.0		5.83 6.98 9.39	369.9 343.8 272.4		15.1 1.32 8.25	479.5 388.7 340.1		8.32 9.57 16.65	307.7 241.4 222.2		11.08 21.19 22.92	300.0 263.4 167.3
CG_Obs/Exp		10.74	10.27		10.74	10.24		10.84	10.24		10.72	10.29		10.96	10.55
(CG_O-E)/(TG_O+CA_O-TG_E-CA_E)		3.72	8.30		8.90	8.63		9.91	8.19		9.08	8.94		9.83	9.43

Table-2-d: Frequency of bases and dinucleotides in Sox4 gene

ORGANISM & PATTERN	*Homo sapiens*			*Pan troglodytes*			*Mus musculus*			*Gallus gallus*			*Xenopus tropicalis*			*Danio rerio*		
	Gene	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**
G	1309	543	766	1499	541	958	1241	404	837	556	439	117	906	91	815	790	204	586
A	1143	249	894	1452	250	1202	1163	225	938	226	178	48	1015	81	934	854	186	668
T	1169	206	963	1438	204	1234	1175	151	1024	149	118	31	1055	56	999	854	100	754
C	1248	588	660	1460	589	871	1202	364	838	558	464	94	896	128	768	763	235	528
TG	277	66	211	330	64	266	282	54	228	49	33	16	261	11	250	233	27	206
CG	315	208	107	337	209	128	259	119	140	222	182	40	102	28	74	173	69	104
CA	240	84	156	322	85	237	261	77	184	66	53	13	270	28	242	236	63	173

ORGANISM & PATTERN	*Homo sapiens*			*Pan troglodytes*			*Mus musculus*			*Gallus gallus*			*Xenopus tropicalis*			*Danio rerio*		
	Gene	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**
TOTAL (G+A+T+C)		**1586**	**3283**		**1584**	**4265**		**1144**	**3637**		**1199**	**290**		**356**	**3516**		**725**	**2536**
P(G) P(A) P(T) P(C)		0.34 0.15 0.13 0.37	0.23 0.27 0.29 0.20		0.34 0.15 0.12 0.37	0.22 0.28 0.28 0.20		0.35 0.19 0.13 0.31	0.23 0.25 0.28 0.23		0.36 0.14 0.09 0.38	0.40 0.16 0.10 0.32		0.25 0.22 0.15 0.36	0.23 0.26 0.28 0.21		0.28 0.25 0.13 0.32	0.23 0.26 0.29 0.20
P(TG) P(CA) P(CG)		0.04 0.05 0.12	0.06 0.05 0.04		0.04 0.05 0.12	0.06 0.05 0.04		0.04 0.06 0.11	0.06 0.05 0.05		0.03 0.05 0.14	0.04 0.05 0.13		0.04 0.08 0.09	0.06 0.05 0.05		0.03 0.08 0.09	0.06 0.05 0.04
EXP(TG) EXP(CA) EXP(CG)		70.5 92.3 201.3	224.6 179.7 153.9		69.6 92.9 201.1	277.1 245.4 195.6		53.3 71.5 128.5	235.6 216.1 192.8		43.2 68.8 169.8	12.51 15.56 37.92		14.3 29.1 32.7	231.5 204.0 178.0		28.1 60.2 66.1	174.2 139.0 122.0
CG_Obs/ Exp		11.03	10.69		11.03	10.65		10.92	10.72		11.07	11.05		10.85	10.41		11.04	10.85
(CG_O-E)/(TG_O+CA_O-TG_E-CA_E)		9.47	11.25		9.42	13.44		8.43	11.32		9.53	12.22		11.06	8.15		11.82	9.72

Table-2-e: Frequency of bases and dinucleotides in ZFYVE16 gene

ORGANISM & PATTERN	*Homo sapiens*			*Pan troglodytes*			*Mus musculus*			*Gallus gallus*			*Xenopus tropicalis*			*Danio rerio*		
	Gene	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**	**Gene**	**CGI**	**Non-CGI**
G	14339	176	14163	13477	269	13208	8394	122	8272	5986	108	5878	6943	114	6829	9231	67	9164
A	21895	103	21792	20465	135	20330	12113	33	12080	10046	60	9986	10515	86	10429	15476	47	15429
T	25170	74	25096	24173	119	24054	14422	65	14357	8870	59	8811	11470	82	11388	16082	41	16041
C	14366	151	14215	13621	237	13384	8133	129	8004	6739	109	6630	6385	117	6268	9762	68	9694
TG	5534	23	5511	5257	39	5218	3559	21	3538	1994	19	1975	2759	17	2742	3540	6	3534
CG	669	46	623	621	73	548	312	43	269	381	36	345	384	36	348	1109	28	1081
CA	5025	22	5003	4728	34	4694	2930	9	2921	2685	18	2667	2339	24	2315	3771	21	3750
TOTAL (G+A+T+C)		**504**	**75266**		**523**	**70976**		**349**	**42713**		**336**	**31305**		**399**	**34914**		**223**	**50328**
P(G) P(A) P(T) P(C)		0.34 0.20 0.14 0.30	0.18 0.29 0.33 0.18		0.51 0.25 0.22 0.45	0.18 0.28 0.33 0.18		0.35 0.09 0.18 0.37	0.19 0.28 0.33 0.18		0.32 170.1 70.32	0.18 0.31 0.28 0.21		0.28 0.21 0.20 0.29	0.19 0.29 0.32 0.18		0.30 0.21 0.18 0.30	0.18 0.30 0.31 0.19
P(TG) P(CA) P(CG)		0.05 0.06 0.10	0.06 0.05 0.03		0.11 0.11 0.23	0.06 0.05 0.03		0.06 0.03 0.12	0.06 0.05 0.03		0.05 0.05 0.10	0.05 0.06 0.04		0.05 0.06 0.08	0.06 0.05 0.03		0.05 0.06 0.09	0.05 0.05 0.03
EXP(TG) EXP(CA) EXP(CG)		52.7 25.8 30.8	2674.8 4722.3 4115.7		121.9 61.2 61.2	2490.6 4476.2 3833.6		45.0 22.7 12.2	1550.0 2780.4 2263.6		35.0 18.9 19.4	1244.8 1654.4 2114.9		33.4 23.4 25.2	1225.9 2227.4 1872.2		20.4 12.3 14.3	1765.1 2920.8 2971.8
CG_Obs/ Exp		10.87	10.23		10.59	10.22		10.95	10.17		11.02	10.27		11.07	10.28		11.37	10.61
(CG_O-E)/ (TG_O+CA_O-TG_E-CA_E)		12.36	7.39		12.20	7.38		11.21	8.30		37.0	7.19		9.60	8.29		8.80	8.88

5.2 Multiple sequence alignment and phylogenetic analysis of genes

The gene sequences with respect to its orthologues in higher eukaryotic organisms were retrieved from the NCBI web server, and further multiple sequence alignment was done followed by Phylogenetic trees constructed. The phylogenetic trees show that *homo sapiens* and *Pan troglodytes* are the most closely related species and *Mus musculus* shares the common ancestor in most of the selected genes but with one exception in Dnmt1 gene in which *Gallus gallus* shares a closer ancestral relationship. The rest of the organisms in all the trees constructed of different genes shows the same evolutionary hierarchy.

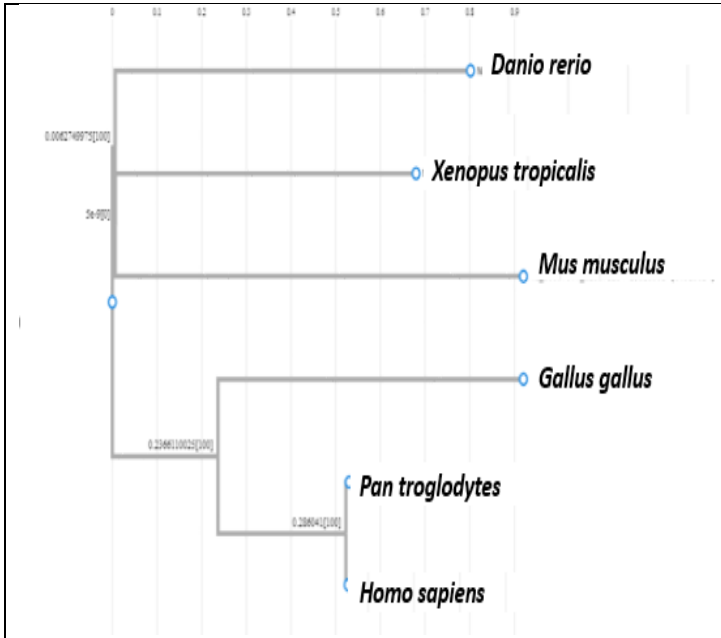


Fig-3 Phylogenetic tree of Dnmt1 gene sequence

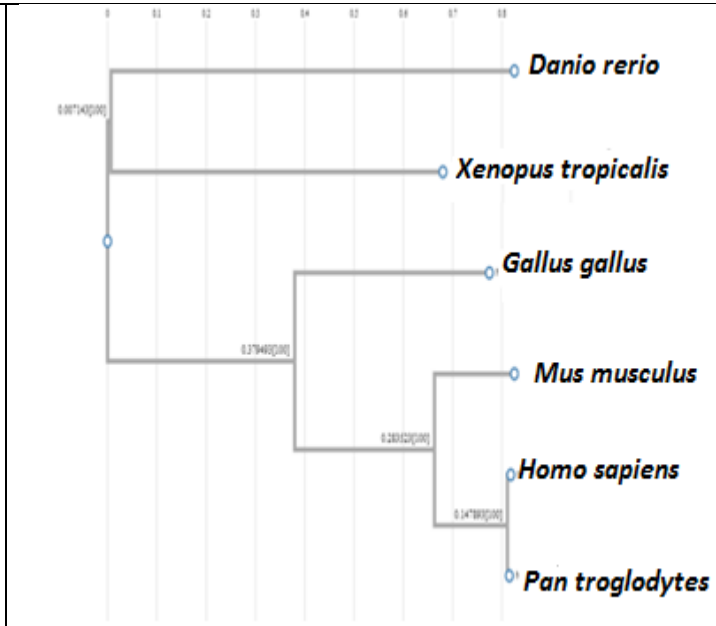


Fig-4 Phylogenetic tree of Gadd45a gene sequence

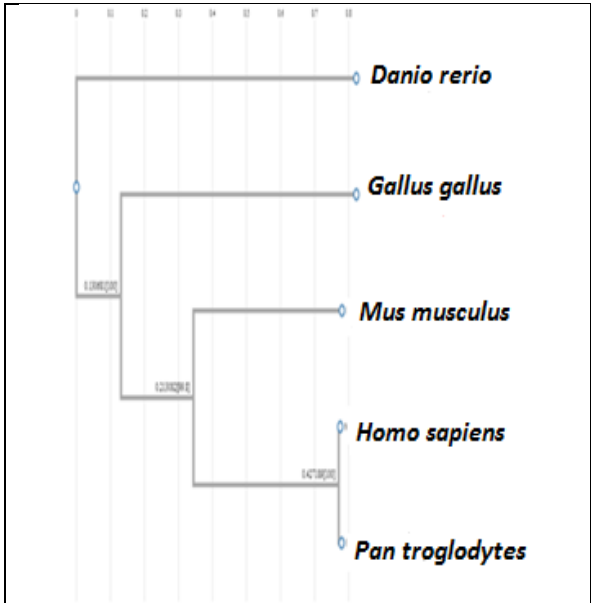


Fig-5 Phylogenetic tree of NanoG gene sequence

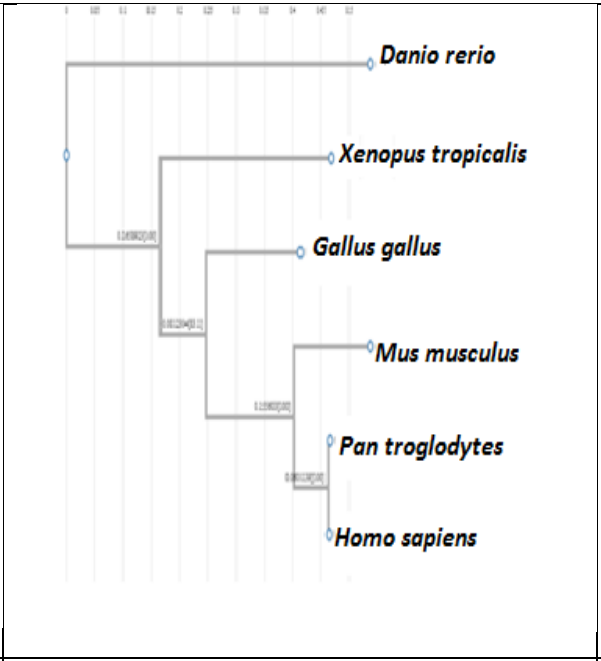


Fig-6 Phylogenetic tree of SOX4 gene sequence

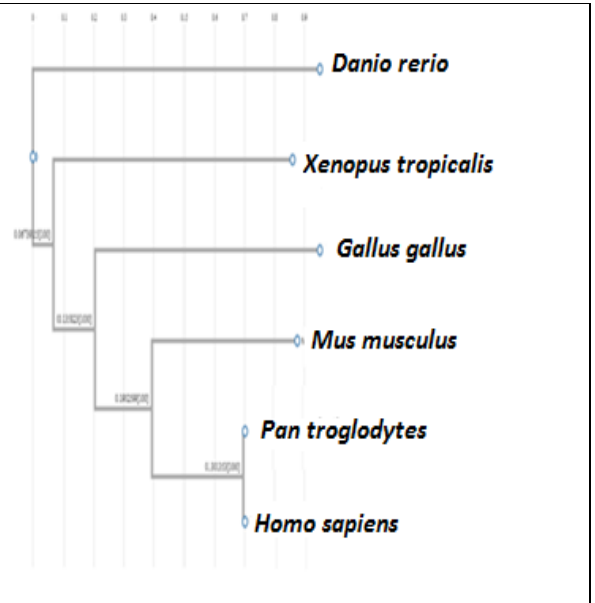


Fig-7 Phylogenetic tree of ZFYVE16 gene sequence

5.3 Comparison based on CG_obs/exp + 10 ratio with corresponding gene and organism for CGI sequence and Non-CGI sequence

Table-3-a CG_OBS/EXP + 10 for CGI sequence

	Hsa	Ptr	Mmu	Gga	Xtr	Dre
Dnmt1	10.80	10.91	10.85	10.88	10.86	11.07
Gadd45a	10.83	10.81	10.88	10.80	10.69	10.89
Nanog	10.74	10.74	10.84	10.72	Nf*	10.96
Sox4	11.03	11.03	10.92	11.07	10.85	11.04
ZFYVE16	10.87	10.59	10.95	11.02	11.07	11.37

*Gene not found

Table-3-b CG_OBS/EXP + 10 for non-CGI sequence

	Hsa	Ptr	Mmu	Gga	Xtr	Dre
Dnmt1	10.34	10.35	10.27	10.43	10.29	10.46
Gadd45a	10.62	10.59	10.44	10.87	10.27	10.35
Nanog	10.27	10.24	10.24	10.29	Nf*	10.55
Sox4	10.69	10.65	10.72	11.05	10.41	10.85
ZFYVE16	10.23	10.22	10.17	10.27	10.28	10.61

*Gene not found

The data from the formulated index of CG_Obs/exp + 10 clearly depict that the value in CGI sequence is greater as compare to the non-CGI sequences of all the genes with respect to the selected higher eukaryotic organism.

5.4 Comparison based on (CG_O-E)/ (TG_O+CA_O-TG_E-CA_E) + 10 ratio with corresponding gene and organism for CGI sequence and Non-CGI sequence

Table-4-a (CG_O-E)/ (TG_O+CA_O-TG_E-CA_E) +10 for CGI sequence

	Hsa	Ptr	Mmu	Gga	Xtr	Dre
Dnmt1	9.08	5.43	9.51	10.78	8.92	10.49
Gadd45a	9.09	8.95	9.16	8.37	8.95	9.88
Nanog	3.72	8.90	9.91	9.08	Nf*	9.83
Sox4	9.47	9.42	8.43	9.53	11.06	11.82
ZFYVE16	12.36	12.20	11.21	37.00	9.60	8.80

*Gene not found

Table-4-b (CG_O-E)/ (TG_O+CA_O-TG_E-CA_E) +10 for non-CGI sequence

	Hsa	Ptr	Mmu	Gga	Xtr	Dre
Dnmt1	8.61	8.59	8.75	8.49	8.80	9.34
Gadd45a	6.92	7.81	8.10	8.53	9.21	9.08
Nanog	8.30	8.63	8.19	8.94	Nf*	9.43
Sox4	11.25	13.44	11.32	12.22	8.15	9.72
ZFYVE16	7.39	7.38	8.30	7.19	8.29	8.88

*Gene not found

Since CG/CG mutates to TG/CA, the CG suppression is known to have moderate increase in TGs and CAs. Based on this fact, another index was formulated to compare CGI and non-CGIs. However the results are inconclusive. It shows that the index is not suitable for genome sequence analysis.

5.5 Dendrogram constructed from the above data in tables 3.1, 3.2, 4.1, and 4.2 for CGI and Non-CGI sequence

The dendrogram illustrates the relationship between higher eukaryotic organisms of the same phylum but different classes by displaying the ratios of CG obs/exp and (CG O-E)/ (TG O+CA O-TG E-CA E) in CGI and non-CGI sequences. The results are inconclusive as they do not shows the evolutionary correlation.

The abbreviations used in the below figures *Danio rario* (Dre), *Homo sapiens* (Has), *Pan troglodytes* (Ptr), *Gallus gallus* (Gga), *Mus musculus* (Mmu), *Xenopus tropicalis* (Xtr).

Fig-8 CG_OBS/EXP + 10 for CGI sequence

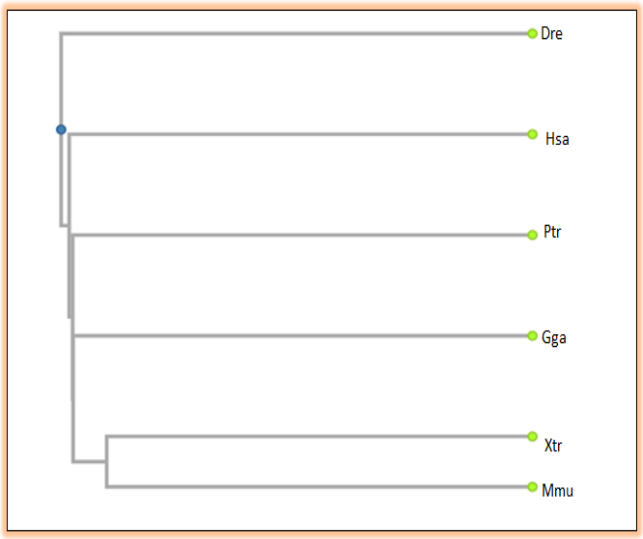


Fig-9 CG_OBS/EXP + 10 for non-CGI sequence

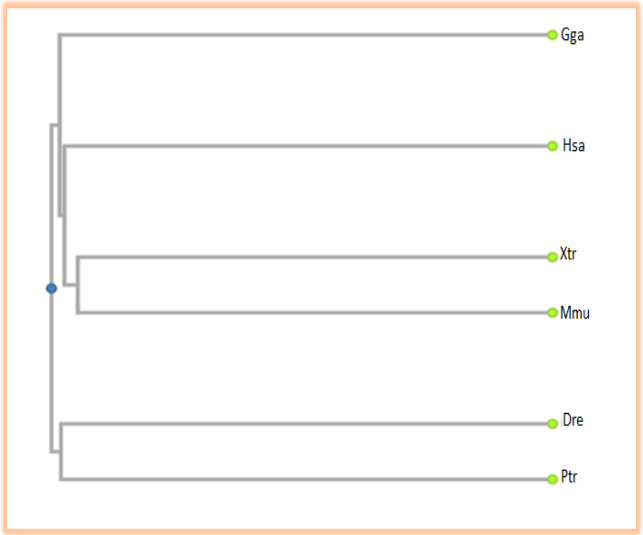


Fig-10 (CG_O-E)/ (TG_O+CA_O-TG_E-CA_E) + 10 for CGI sequence

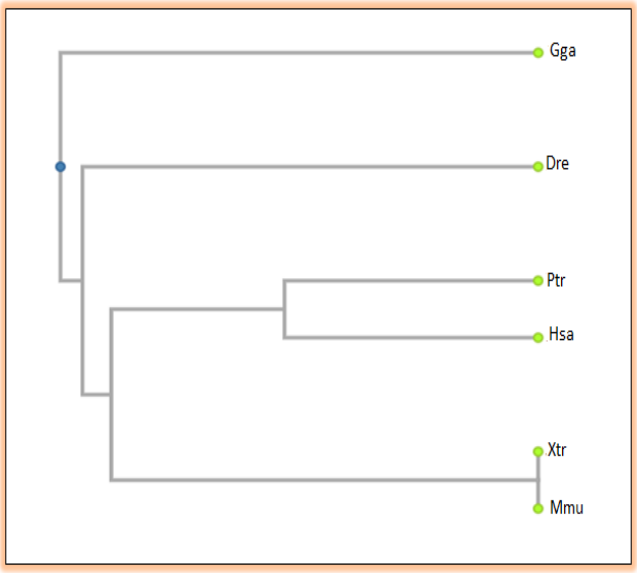
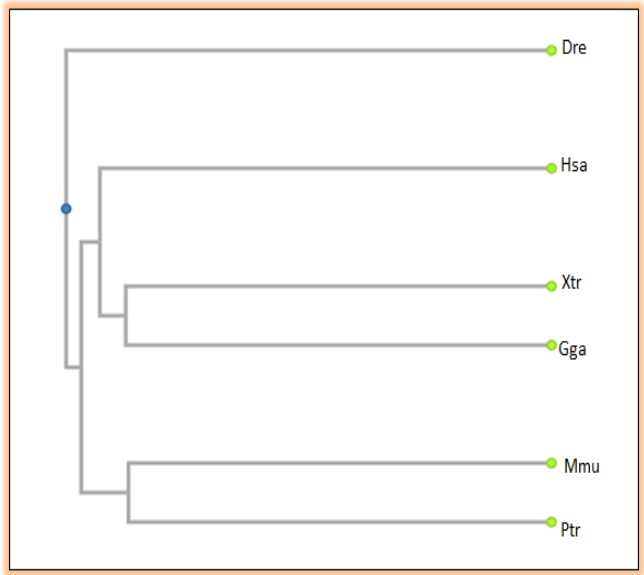


Fig-11 (CG_O-E)/ (TG_O+CA_O-TG_E-CA_E) + 10 for non-CGI sequence



5.6 Comparison of Mean length of CGI sequence of a gene with respect to their higher eukaryotic organisms.

Table: 5-a Illustrating the Mean Length of CGI sequence

ORGANISM	Has	Ptr	Mmu	Gga	Xtr	Dre
	656.4 ± 475.1	731.4 ± 480.5	558.8 ± 367.6	598.2 ± 358.3	302 ± 80.8	492.8 ± 202.2
GENES	**Dnmt1**	**Gadd45a**	**Nanog**	**Sox4**	**ZFYVE16**	
	450 ± 134	518 ± 222	186 ± 78	1099 ± 443	510 ± 169	

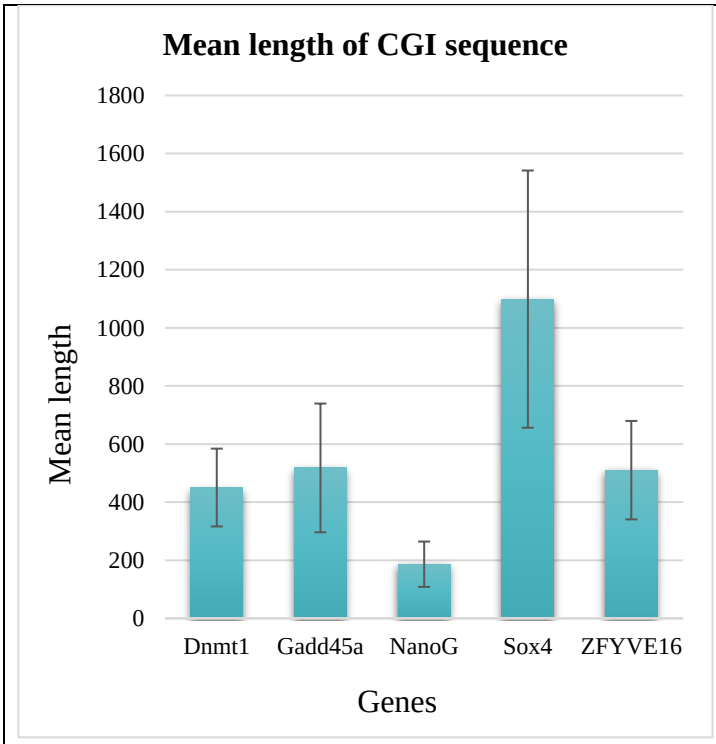


Fig-12 Mean Length of CGI sequence with respect to gene

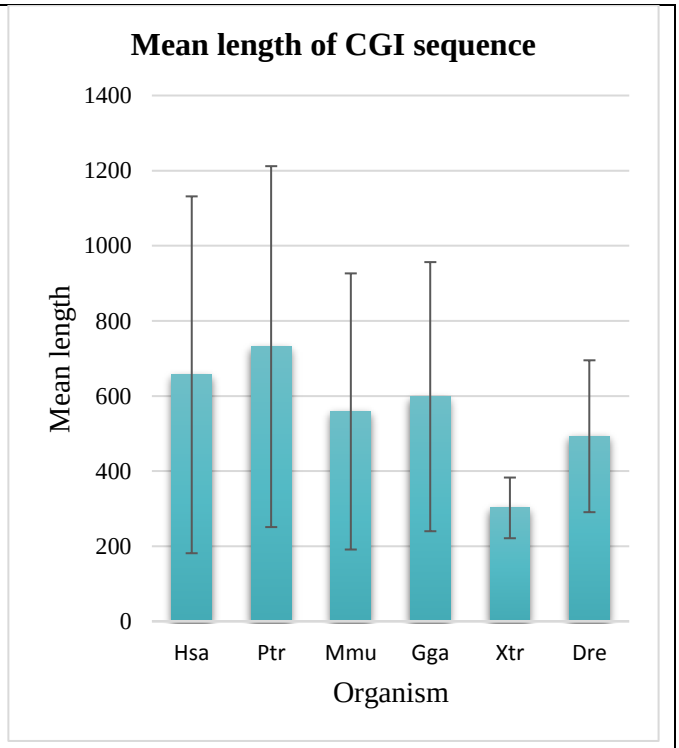


Fig-13 Mean Length of CGI sequence with respect to the higher eukaryotic organism

GC content and CpG obs/exp values define CGIs and are expected to play important role in their function of remaining unmethylated. In another experiment, it was attempted to study another parameter in comparison of CGI sequences. CG gaps were defined as the space between adjacent CGs in the CGIs. Figure -12 shows the Sox4 gene has the longest length of all the selected genes, Nanog gene has the shortest with the absence seen in *Xenopus tropicalis*. Figure-13 demonstrates that, aside from *Xenopus tropicalis*, which has the shortest length of CGI, *Pan troglodytes* have the longest.

5.7 Comparison of Mean CG gap in CGI sequence of a gene with respect to their higher eukaryotic organisms

Table: 5-b Illustrating the Mean CG gap in CGI sequence

ORGANISM	Hsa	Ptr	Mmu	Gga	Xtr	Dre
	8.73 ± 7.57	8.62 ± 7.36	9.54 ± 8.28	8.64 ± 7.60	12.67 ± 10.27	10.66 ± 9.60
GENES	**Dnmt1**	**Gadd45a**	**Nanog**	**Sox4**	**ZFYVE16**	
	10.6 ± 9.0	12.4 ± 9.9	13.1 ± 9.1	9.3 ± 7.7	9.5 ± 7.6	

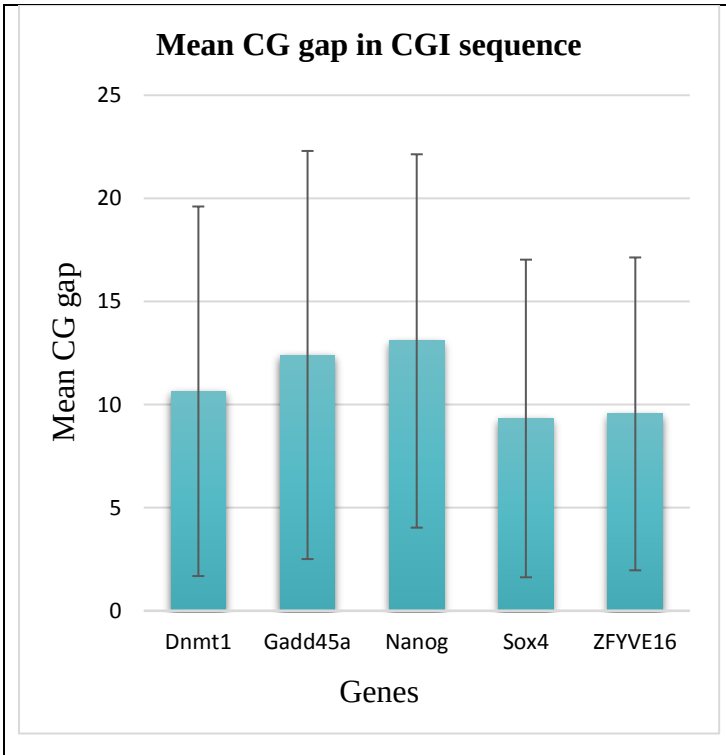


Fig-14 Mean CG gap in CGI sequence with respect to genes

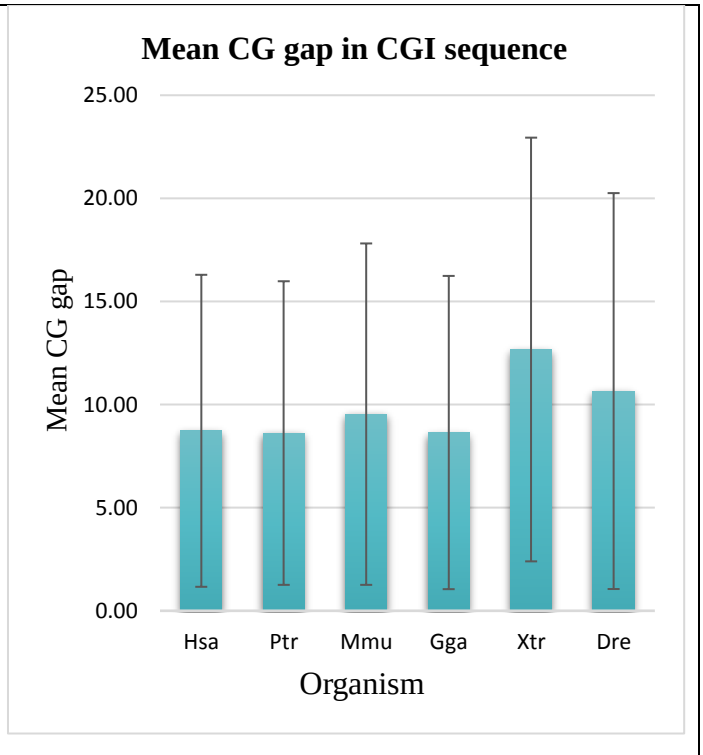


Fig-15 Mean CG gap in CGI sequence with respect to the organism

The above graphs depict that Nanog gene and Xenopus tropicalis organism shows the highest CG gap as compared to the other gene and organisms selected in which Sox4 gene and Gallus gallus shows the lowest degree of CG gaps in their CGI sequences. It can be concluded from the above data that the reduction in the CG gaps is because of DNA methylation which majorly attacks the CG dinucleotides of the genome.

5.8 Comparison of Mean CG_obs/exp of CGI sequence of a gene with respect to their higher eukaryotic organisms

MEAN CG_obs/exp IN CGI SEQUENCE						
	Hsa	Ptr	Mmu	Gga	Dre	Xtr
Dnmt1	0.80	0.92	0.86	0.88	0.08	0.86
Gadd45a	0.84	0.81	0.89	0.81	0.89	0.70
Sox4	1.03	1.04	0.93	1.07	1.04	0.86
ZFYVE16	0.87	0.60	0.95	1.03	1.37	1.08
Nanog	0.75	0.75	0.85	0.72	0.96	*Nf

Table:5-c Illustrating the Mean CG_obs/exp of CGI sequence

*Gene not found

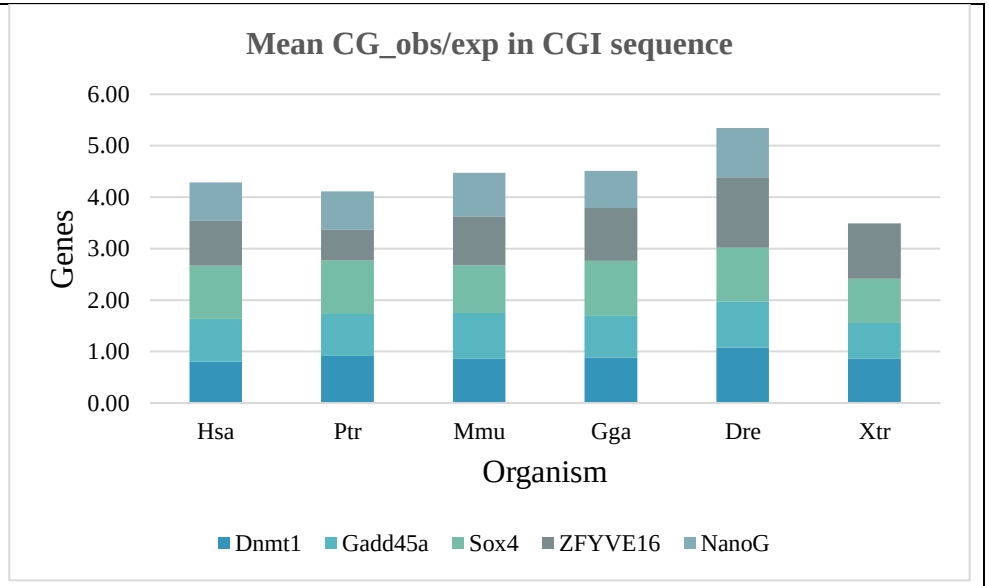


Fig-16 Mean CG_obs/exp of CGI sequence

The Mean CG obs/exp in CGI sequences graph illustrates the strong correlation between CG obs/exp in CGI sequences. In the first and second cases, *Homo sapiens* and *Pan troglodytes* both have greater values of CG obs/exp in the Sox4 gene, with other genes showing a 1 to 1.5-fold drop. Additionally, the ZFYVE16 gene has a higher value in other organisms like *Danio rerio*, *Gallus gallus*, *Xenopus tropicalis*, and *Mus musculus*. Surprisingly, if we look at the entire dataset, *Danio rerio* has the highest value for three genes (Dnmt1, Sox4 and ZFYVE16 gene).

5.9 Comparison of GC percentage of CGI sequence of a gene with respect to their higher eukaryotic organisms

GC percentage of CGI sequence						
	Hsa	Ptr	Mmu	Gga	Dre	Xtr
Dnmt1	73.41	69.9	67.01	63.38	50.38	59.77
Gadd45a	69.89	70.58	67.54	67.26	50.36	54.69
Sox4	71.25	71.28	67.0	75.23	60.55	55.08
ZFYVE16	64.68	64.97	71.63	64.58	60.54	57.64
Nanog	58.98	58.33	62.39	64.56	59.54	*Nf

Table:5-d Illustrating the GC percentage of CGI sequence

*Gene not found

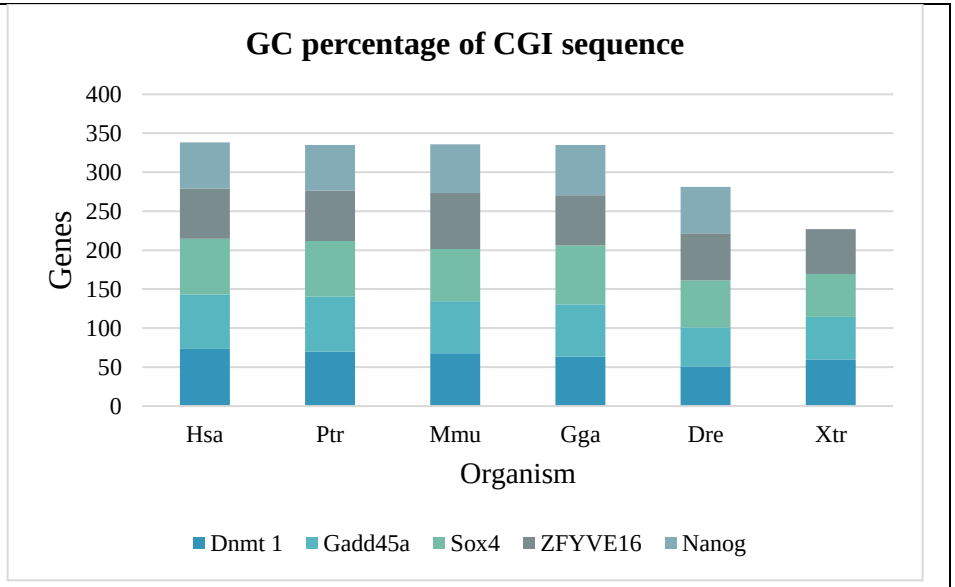


Fig-17 GC percentage of CGI sequence

The GC percentage of the CGI sequence in relation to higher eukaryotic organisms is displayed in the graph above. With the exception of *Xenopus tropicalis*, which lacks Nanog gene has the lowest GC percentage and *Homo sapiens* has the highest as compared to the other organisms.

CHAPTER 6: DISCUSSION

DNA methylation in higher eukaryotes is one of the most important epigenetic modifications that plays important role in the regulation of gene expression, gene silencing, genomic imprinting and X-chromosome inactivation. DNA methylation takes place at position-5 of cytosine in CpG dinucleotides by the action of enzymes known as DNA methyl transferases. 5-methyl cytosine is mutagenic and that has led to the suppression of CGs in the methylated genomes. The spontaneous deamination of 5-methylcytosine to thymine is to blame for the decrease in the frequency of CpG dinucleotides in the genome. Not all CpGs are methylated in the genome. In particular, there are certain GC-rich regions known as CpG islands that are usually not methylated and are distinguish in the rest of the genome. These regions have some unique characteristics i.e. high C+G content and relatively higher abundance of CG dinucleotides which make them structurally as well as functionally important. They are positioned at the 5′ ends of the housekeeping genes which are the transcriptional start sites that keep them transcriptionally active.

These features of CpG islands incline the interest of various scientists also to enhance our knowledge and study some hidden evolutionary characteristics of the CpG islands the hypothesis has been designed in which Comparison of CpG islands in Genomes with different evolutionary lineage has been done keeping in mind the characteristics of the different regions of the genome. Their comparison started by selecting five different genes with respect to the higher eukaryotic organisms of the same phylum but different classes, followed by comparing the CGI sequences and the non-CGI sequences of different genes. Firstly, from the gene sequences, CGI regions were retrieved using the standards algorithms given by Gardiner-Garden and Frommer's and Takai and Jones. Later on, DNA sequence analysis was performed to compare the CGI and non-CGI sequences based on CG obs/exp values as well as TG and CA o/e values. As expected the CG obs/exp + 10 value of CGIs was higher than those in non-CGIs in all the genes of the studied organisms. Multiple sequence alignment, followed by phylogenetic analysis, was performed. Dendrograms and phylogenetic trees were constructed keeping in mind some of the hypotheses for the comparison of the CGI sequences and non-CGI sequences.

The phylogenetic analysis shows that *Homo sapiens* and *Pan troglodytes* were closely related species and share common ancestors. Furthermore, it is seen that *Mus musculus* has a close

relationship with *Homo sapiens* and *Pan troglodytes*, followed by *Gallus gallus, Xenopus tropicalis, and Danio rerio*.

The CGI and non-CGI sequences based parameters of (CG_obs/exp and TG & CA) were classified using UPGMA and represented in the form of dendrograms. The dendrograms representing the CG_obs/exp + 10 ratio and the (CG_O-E)/ (TG_O+CA_O-TG_E-CA_E) + 10 ratio in CGI and non-CGI sequences were compared with the phylogenetic trees. No appreciable association was observed between two results. It may be inferred that CGIs evolution is strictly in accordance with that of the overall gene sequences. It may be attributed to the high mutagenic nature of CGs and their varying rates of mutation in CGIs and non-CGIs.

Further CGIs were compared on the basis of other sequence attributes such as length, GC percentage and CG_obs/exp in CGI sequences. It was observed that the mean length of CGIs can vary considerably among genes and it was evident from the case of Sox4 that its CGI is several fold longer that nanog CGIs. The mean CG gaps in CGIs also vary considerably among genes. When CG gaps in CGIs were compared among genomes of different evolutionary lineages, it was observed that gap was conspicuously smaller in mammals and chicken when compared to fish and amphibian. It may be inferred that CGI characteristics are related to the complexity of genomes.

The current work may be extended by using more genes and organisms to get a clearer picture of CGI evolution. Statistical analysis can make the finding more robust.

CHAPTER: 7 CONCLUSIONS

Following the completion of our research, we used the hypothesis while keeping in mind the distinctive characteristics of CpG islands i.e. they usually remain unmethylated, have high C+G content, and relatively higher number of CG dinucleotides also DNA methylation which is the primary epigenetic mechanism involved in the evolution of CpG islands. In comparison to higher eukaryotic organisms belonging to the same phylum, CGI and non-CGI sequences have been compared. The comparison was followed by a multiple sequence alignment and phylogenetic analysis, from which different findings were generated. These results were further analysed and compared based on the GC content, the average length of CGI sequences, CG obs/exp ratio and CG gaps.

We inferred from the data that CGI sequences differ from non-CGI sequences in that they have a high GC percentage, a relatively higher mean length, and a GC obs/exp ratio. The phylogenetic analysis shows that *Homo sapiens* and *Pan troglodytes* are the most closely related species followed by mammals and birds but the UPGMA dendrogram results do not show the evolutionary correlation. CG gaps are the space between adjacent CGs in the CGIs. After examination, it was identified that the mean length and CG gaps are the peculiar features, Sox4 gene has some influential characteristics since it has the largest length, shows the lowest CG gaps and relatively higher GC percentage as compared to all the chosen genes. Also, nanog gene and *Xenopus tropicalis* organism show the highest CG gaps and lowest length in the above parameters selected for the comparison because some evolution might have taken place with time.

REFERENCES

Angeloni, A., & Bogdanovic, O. (2021). Sequence determinants, function, and evolution of CpG islands. Biochemical Society Transactions, 49(3), 1109-1119. https://doi.org/10.1042/bst20200695

Bird, A. P. (1980). DNA methylation and the frequency of CpG in animal DNA. *Nucleic acids research*, *8*(7), 1499-1504.

Ciechomska, M., Roszkowski, L. and Maslinski, W., 2019. DNA Methylation as a Future Therapeutic and Diagnostic Target in Rheumatoid Arthritis. *Cells*, 8(9), p.953.

Cooper, D. N., and H. Youssoufian. 1988. The CpG dinucleotide and human genetic disease. Hum. Genet.78:151–155.

Cooper, D. N., and M. Krawczak. 1993. Human gene mutation. BIOS Scientific Publishers Limited, Oxford.

Craig, J., & Bickmore, W. (1994). The distribution of CpG islands in mammalian chromosomes. *Nature Genetics*, *7*(3), 376-382. https://doi.org/10.1038/ng0794-376

Fryxell, K., & Moon, W. (2004). CpG Mutation Rates in the Human Genome Are Highly Dependent on Local GC Content. Molecular Biology And Evolution, 22(3), 650-658. https://doi.org/10.1093/molbev/msi043

Gardiner-Garden, M., & Frommer, M. (1987). CpG islands in vertebrate genomes. *Journal of molecular biology*, *196*(2), 261-282.

Han, L., & Zhao, Z. (2008). Comparative Analysis of CpG Islands in Four Fish Genomes. Comparative And Functional Genomics, 2008, 1-6. https://doi.org/10.1155/2008/565631

Handa, V., & Jeltsch, A. (2005). Profound flanking sequence preference of Dnmt3a and Dnmt3b mammalian DNA methyl transferases shape the human epigenome. *Journal of molecular biology*, *348*(5), 1103-1112.

Johnson, J. (1985). 1 Determination of DNA Base Composition. *Methods In Microbiology*, 1-31. https://doi.org/10.1016/s0580-9517(08)70470-7

Moore, L., Le, T., & Fan, G. (2012). DNA Methylation and Its Basic Function. Neuropsychopharmacology, 38(1), 23-38. https://doi.org/10.1038/npp.2012.112

Rodriguez, F., Yushenova, I. A., DiCorpo, D., & Arkhipova, I. R. (2022). Bacterial N4-methylcytosine as an epigenetic mark in eukaryotic DNA. *Nature communications*, *13*(1), 1-17.

Sharif, J., Endo, T., Toyoda, T., & Koseki, H. (2010). Divergence of CpG island promoters: A consequence or cause of evolution?. *Development, Growth &Amp; Differentiation*, *52*(6), 545-554. https://doi.org/10.1111/j.1440-169x.2010.01193.x

Souza, A., Lopes, O., Liberato, A., Oliveira, P., Herrero, S., & Nascimento, A. et al. (2020). Association between SNPs and Loss of Methylation Site on the CpG Island of the Promoter Region of the Smoothened Gene, Potential Molecular Markers for Susceptibility to the Development of Basal Cell Carcinoma in the Brazilian Population. Asian Pacific Journal Of Cancer Prevention, 21(1), 25-29. https://doi.org/10.31557/apjcp.2020.21.1.25

Sperling, A., & Li, R. (2013). Repetitive Sequences. *Brenner's Encyclopedia of Genetics*, 150-154. https://doi.org/10.1016/b978-0-12-374984-0.01297-3

Takai, D., & Jones, P. A. (2002). Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the national academy of sciences*, *99*(6), 3740-3745.

Tazi, J. (1990). Alternative chromatin structure at CpG islands. *Cell*, *60*(6), 909-920. https://doi.org/10.1016/0092-8674(90)90339-g

Xie, H., Wang, M., Bischof, J., de Fatima Bonaldo, M., & Soares, M. B. (2009). SNP-based prediction of the human germ cell methylation landscape. *Genomics*, *93*(5), 434-440.

Zhao, Z., & Zhang, F. (2006). Sequence context analysis of 8.2 million single nucleotide polymorphisms in the human genome. *Gene*, *366*(2), 316-324.

Ouriginal
by Turnitin

Document Information

Analyzed document	mehakpreet final thesis.pdf (D142457294)
Submitted	7/27/2022 10:47:00 AM
Submitted by	Shreya Gupta
Submitter email	sgupta_msc18@thapar.edu
Similarity	4%
Analysis address	sgupta_msc18.thapar@analysis.urkund.com

Sources included in the report

W	URL: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2008-9-5-r79 Fetched: 6/22/2021 11:31:48 AM	2
W	URL: https://doi.org/10.1042/bst20200695 Fetched: 7/27/2022 10:47:00 AM	11
W	URL: https://www.bioinformatics.org/sms2/dna_stats.html Fetched: 7/27/2022 10:47:00 AM	1
W	URL: https://doi.org/10.1093/molbev/msi043 Fetched: 7/27/2022 10:47:00 AM	3
W	URL: https://www.nature.com/articles/srep24666 Fetched: 2/16/2020 10:23:20 PM	1

Entire Document

i Comparison of CpG islands in Genomes with different evolutionary lineage A Thesis Submitted in partial fulfilment of the requirements for the award of the degree of Master of Science In Biotechnology By Mehakpreet Kaur (Reg no: 302001018) Under the Supervision of Dr. Vikas Handa Assistant Professor Department of Biotechnology Thapar Institute of Engineering and Technology, Patiala, Punjab India June 2022

ii Certificate This is to certify that the thesis entitled, Comparison of CpG islands in Genomes with different evolutionary lineage being submitted by Mehakpreet Kaur (Reg. No.302001018), in partial fulfilment of the requirements for the award of the degree of Master of Science in Biotechnology, Thapar Institute of Engineering and Technology, Patiala, Punjab is a bonafide work carried out under the guidance and conception of Dr. Vikas Handa and that no part of this thesis has been submitted for the award of any other degree. Date: 27/07/2022 Dr. Vikas Handa Supervisor

iii Candidate Declaration I hereby certify that the project work entitled, Comparison of CpG islands in Genomes with different evolutionary lineage in a partial fulfilment of the requirements for the award of the degree of Master of Science in Biotechnology and submitted is an authentic record of my work carried out during the period January 2022 to June 2022 under the guidance of Dr. Vikas Handa, Assistant professor, Department of Biotechnology, Thapar Institute of Engineering and Technology, Patiala, Punjab, India. Date: 27/07/2022 Mehakpreet Kaur This is to certify that the above statement made by the student is correct to the best of our knowledge and belief. Date: 27/07/2022 Dr. Vikas Handa Supervisor