INVESTIGATION OF HUMAN PROMOTER CPG CONTENT AND METHYLATION PROFILES AT DIFFERENT CONSERVATION LEVELS

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

BURAK DEMİRALAY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
BIOINFORMATICS

JULY 2012

INVESTIGATION OF HUMAN PROMOTER CPG CONTENT AND METHYLATION PROFILES AT DIFFERENT CONSERVATION LEVELS

Submitted by **BURAK DEMİRALAY** in partial fulfillment of the requirements for the degree of **Master of Science in Bioinformatics, Middle East Technical University** by,

Prof. Dr. Nazife Baykal

Director, **Informatics Institute**

Assist. Prof. Dr. Yeşim Aydın Son

Head of Department, **Medical Informatics**

Assist. Prof. Dr. Yeşim Aydın Son

Supervisor, **Medical Informatics**, **Middle East Technical University**

Examining Committee Members:

Assoc. Prof. Dr. Tolga Can

Computer Engineering, Middle East Technical University

Assist. Prof. Dr. Yeşim Aydın Son

Medical Informatics, Middle East Technical University

Assist. Prof. Dr.Bala Gür Dedeoğlu

Biotechnology, Ankara University

Assist. Prof. Dr.Hakan Öktem

Applied Mathematics, Middle East Technical University

Assist. Prof. Dr. Vilda Purutçuoğlu

Statistics, Middle East Technical University

Date: 27.06.2012

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: Burak Demiralay

Signature :

ABSTRACT

INVESTIGATION OF HUMAN PROMOTER CpG CONTENT AND METHYLATION PROFILES AT DIFFERENT CONSERVATION LEVELS

Demiralay, Burak
MSc. Bioinformatics Program
Supervisor: Assist. Prof. Dr. Yeşim Aydın Son

JULY 2012, 35 Pages

Methylation of CpG islands located at the promoter regions is a mechanism which controls gene silencing and expression. Hyper or hypo methylation of these sites on promoter sequences have been associated with many diseases, like cancer. Even though promoter CpG islands and their methylation profiles are important regulators of gene expression, the exact mechanism of gene silencing through methylation is not known. Here, we have investigated the status of promoter CpG methylation under various evolutionary pressures by calculating the differences in promoter CpG content and methylation profiles at different pass points. In order to determine the list of genes under each category we have analyzed and compared the orthologs among 58 genomes available through ENSEMBL. The total number of CpG

dinucleotides at the promoter regions of all groups of genes have been calculated and compared. Additionally, we have compared the experimentally determined methylation profiles of these CpG's between human blood cells and fibroblast cells. While the promoter CpG content changed through common to newer genes, the number of the CpG units methylated found to be consistent. Here, we present the functional level analysis of common gene lists at different pass points and report the differences of the promoter CpG content and the methylation profiles among these groups with distinct evolutionary conservation status. We have also observed the conservation status of individual methylated CpG units on the low and high methylated genes. Our analysis revealed that the surrounding methylation content had a positive effect on the conservation of individual CpG's.

Keywords: epigenetics, gene regulation, promoter methylation, CpG dinucleotides, molecular evolution

ÖZ

İNSAN PROMOTER CpG İÇERİĞİ VE METİLASYON PROFİLLERİNİN DEĞİŞİK EVRİMSEL SÜREÇLERDE İNCELENMESİ

Demiralay, Burak
Yüksek Lisans, Biyoenformatik Programı
Tez Yöneticisi: Yard. Doç. Dr. Yeşim Aydın Son

Temmuz 2012, 35 Sayfa

Promotor bölgelerinde bulunan CpG'lerin metilasyonu gen ekspresyonu ve gen susturulmasında kullanılan bir yöntemdir. Promotor bölgelerinin az ya da çok metilasyonu başta kanser olmak üzere birçok hastalıkla ilişkilendirilmiştir. Promotor CpG adalarındaki metilasyon, gen ekspresyonunu düzenleyen önemli bir profil olduğu halde, metilasyon aracığılıyla gen susturmanın tam mekanizması ve bireysel CpG'lerin rolü bilinmemektedir. Bu çalışma değişik evrimsel zaman dilimleri üzerinde promoter CpG sayısını ve metilasyon profilini hesaplayarak; evrim baskısının promotor CpG'leri üzerindeki etkisini araştırmaya yöneliktir.

Araştırdığımız her kategorideki genlerin listesini bulabilmek için, ENSEMBL veritabanından 58 canlının genomu üzerinde çalıştık. Promotor bölgelerindeki CpG'lerin toplam sayısı hesaplanıp karşılaştırılmıştır. Ek olarak, fibroblast ve beyaz kan hücrelerindeki promoter CpG'lerin metilasyon profili de karşılaştırılmıştır. Ortak genlerden yeni genlere doğru ilerlerken promoterların CpG içeriğinin değiştiği gözlenmiş ama metilasyona uğramış CpG'ler istikrarlı bulunmuştur. Bu çalışmada, değişik zaman ve korunmuşluk durumlarındaki ortak genlerin fonksiyonel analizi yapılmış ve değişik evrimsel korunma statüsündeki genlerin promoter CpG içeriğini ve metilasyon profillerini karşılaştırılmıştır. Bu çalışmada ayrıca her bir metilli CpG'nin az ve çok metilli gen prometerlarındaki korunma durumunu da incelenmiştir. Analiz sonuçları yüksek oranda metilli CpG'lerin etrafındaki diğer metilli CpG sayısının, CpG'lerin türler arasındaki korunması üzerinde olumlu etkisi olduğunu gösterdi.

Anahtar Kelimeler: epigenetik, gen regulasyonu, promoter metilasyonu, CpG dinükleotid, moleküler evrim

TABLE OF CONTENTS

ABSTRACT ……..iv
ÖZ…. vi
LIST OF TABLES x
LIST OF FIGURES xi
PREFACE..... xii

CHAPTERS

1 INTRODUCTION AND BACKGROUND…….. ……..1

1.1 Epigenetic Regulation and CpG Islands.....1
1.2 Methylation of Promoter CpG Units …….....2
1.3 Gene Silencing through Promoter Methylation and Cancer……...3
1.4 Homology and Ortholog Genes……..4

2 METHODS AND MATERIALS.....6

2.1 Databases and Tools.....6
2.2 Data.....7
2.2.1 Genome Sequence Data Used.....7
2.2.2 Whole Genome Methylation Profile Data.....9

2.3 Classification Of Human Genes 9

3 RESULTS AND DISCUSSION.....12

3.1 Classification of H.Sapiens Genes At Different Evolutionary Pass points……..……12
3.2 Total Promoter CpG Frequencies 13
3.3 Comparison of Methylation Profiles 14

3.4 Common Genes 16
3.4.1 Identification of Common Genes.....16
3.4.2 Functional Annotation of Common Genes.....18
3.5 Analysis of Conservation and Methylation of Individual CpG Units…..21

4 CONCLUSION 27
REFERENCES.....30
APPENDIX 33

LIST OF TABLES

Table 1: List of Species

Table 2; Distribution of Homo sapiens genes according to the passpoint they have emerged.

Table 3 Total Promoter CpG Count in genes with different emergence time

Table 4: Methylation Profiles in genes with different emergence time

Table 5 : Distribution of common genes exclusive for each subgroup.

Table 6: Methylated and Unmethylated CpG conversion count in Human Chromosome 1 vs. aligned Mouse sequences

Table 7: Methylated CpG conversion rate in different methylation profile of promoters

Table 8: Methylated CpG conversion count in different methylation profile of promoters

LIST OF FIGURES

Figure 1 : Epigenetic modifications

Figure 2: Schema of orthologs and paralogs

Figure 3: Evolutionary Tree of Life

Figure 4: Total CpG and methylated CpG in time

Figure 5: Network representation for the common and exclusive ranging from all 58 species to Homini

Figure 6: Functional annotation charts representing the biological processes and molecular function each group of common genes are involved.

PREFACE

Methylation of CpG units on promoter regions is one of the epigenetic mechanisms that controls gene silencing and activation of transcription. Even though there are many CpG Island Prediction tools it is still a challenge to point out single or groups of CpG units that are important in promoter methylation. In this study we have identified the genes with different levels of evolutionary conservation in order to analyze the promoter CpG unit distribution and methylation profiles; also we identified individual methylated CpGs' conservation status.

CHAPTER 1

1. INTRODUCTION AND BACKGROUND

1.1 Epigenetic Regulation and CpG Islands

Epigenetics is the study of inherited changes in phenotype (appearance) or gene expression caused by mechanisms other than changes in the underlying DNA sequence. Epigenetic mechanisms are shown to affect various biological features such as; mammalian X-chromosome inactivation, imprinting, gene silencing and carcinogenesis. As epigenetics is that important and not carried by sequence itself but by chemical modification on the genome, it is more complex and less understood compared to sequences carrying information. Chromatins are modified and genes are activated and silenced by the action of Polycomb-trithorax group (Pc-G/trx) protein complexes, topological confirmation of DNA in nucleus is inherently changed, and coding and non-coding regions of genome are methylated. [1,2] The importance of CpG's is that almost only Cytosine in CpG dinucleotides is methylated in genome.

In mammalian genome 60% to 90% of all CpG's are found to be methylated. One of the sites that undergoe heavy methylation is CpG islands[3]. Although there isn't a strict definition on CpG islands, it is accepted that GC percentage should be greater than 50% within a region of at least 200 bp and with an observed/expected CpG ratio of minimum 60%. [4] Yet, according to a relatively recent study that had been done on human chromosomes 21 and 22, DNA regions >500 bp with a GC content >55% and observed CpG/expected CpG of 0.65 were more likely to be the true CpG islands associated with the 5' regions of genes [5]

1.2 Methylation of Promoter CpG Units

Cytosine molecule in genomes is methylated and turned into 5-Methylcytosine by an enzyme called C-5 cytosine-specific DNA methylase. Although Cytosine can be methylated when it is part of dinucleotides CC, CA, CT; it is mostly methylated when it is paired with Guanine base in dinucleotide CG. Only <%0.2 of non CpG sites are methylated while >%68 of CpG sites are methylated. An interesting feature of Methylated CpG is that it turns into TpG mostly by spontaneous deamination of 5-methylcytosine and also by mCpG-specific base modification by mutagens and secondary factors [6].

Recent findings show that CpG's in intergenic regions of active genes in animals can be methylated proposedly to prevent intragenic gene transcription, however the most important CpG methylation centers for regulation are promoters [7] Promoter regions are relatively small (~1500 bp) regions located in 5' end upstream of genes that has functional sites. Analysis of CG dinucleotides in promoters in many studies revealed that genes that are expressed have unmethylated CpG's whereas unexpressed genes promoters have their CpG's methylated. [8]

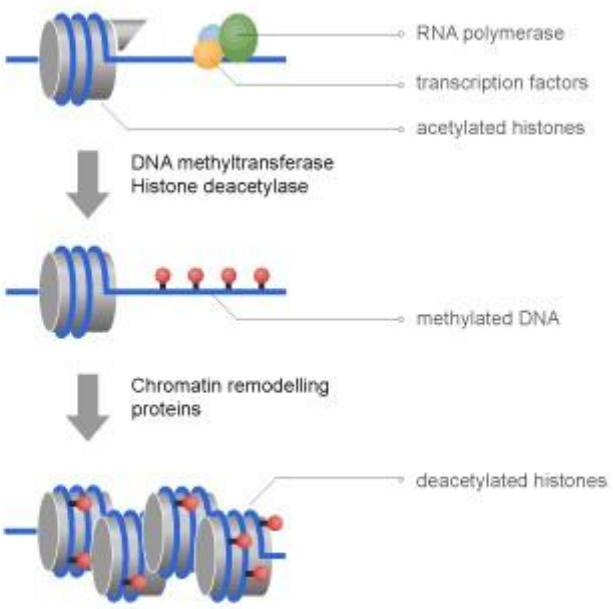


Figure 1: Epigenetic modifications that abolish gene expression [9]

CpG Methylation can prevent expression blocking the access of some transcription factors to their binding sites[10] and although CpG methylation is only one part of epigenetic chromatin remodeling, as depicted in figure 1, methylated CpG's attract transcription factor binding or the recruitment of methyl-binding proteins and their associated chromatin remodeling factors that are responsible for Histone deacetylation and in turn DNA condensation. So when DNA is condensed and compacted in a region, genes in that region are not expressed and in turn silenced [11,12,13]

1.3 Gene Silencing through Promoter Methylation and Cancer

Main importance of CpG is that if CpG dinucleotides are clustered 5' end of genes, they are shown to be associated with gene regulation and cancer. Methylation of these cytosine bases is a common chemical modification that is associated with transcriptional repression and silencing of tumor suppressor genes. As a result, the methylation profiles between cancer and normal cells can be distinguished and serve as a molecular biomarker.

In cancer, a global hypomethylation and CpG island specific regional hypermethylation is seen. Although CpG dicnuleotide methylation is seen as a consolidating event other than initiating event in cancer induced silencing of genes, its mechanism is different and not clear in different cancer cases. However if CpG's are hypermethylated in promoter region of tumor suppressing genes, this may result in cancer. [14] It has also been suggested that dynamic methylation and demethylation of CpG’s shift to methylation after silencing of genes in cancer and CpG methylation in CpG island help to silence genes by chromosome remodeling with secondary factors[15]

1.4 Homology and Ortholog Genes

When genes are descended from a common origin, they are called homolog genes. There are two types of homolog genes, orthologs and paralogs. Ortholog genes are arise by speciation while paralog genes are arise by duplication [16] When genes are duplicated in a genome of a species, they may have been modified by mutation, fusion, fission or other rearrangements and in turn, if they stay functional, they may have different functions in genome. These genes are called paralog genes. If a gene is duplicated in a species after a given speciation event, these paralog genes in a species are called in-paralogs, whereas if a gene is duplicated before a given speciation event, the non-vertical copies of paralog genes are called outparalogs of each other. [17]

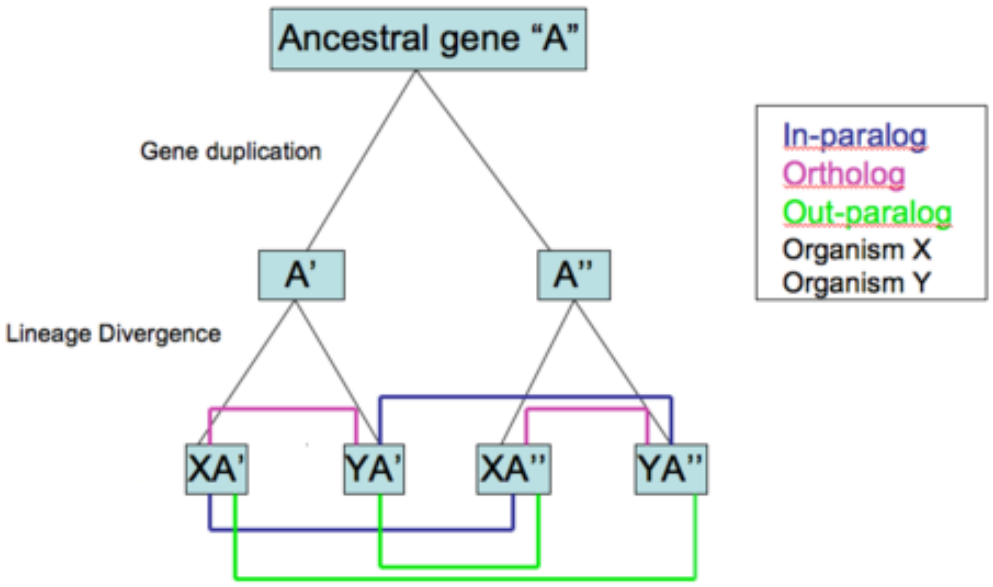


Figure 2: Schema of orthologs and paralogs. Ortholog genes are evolved after speciation and paralog genes are arise by duplication. [18]

We know evolution is change in traits of living organisms over generations including the emergence of new species. When speciation occurs, if gene is not lost during this phenomenon, two copies of same gene results in different species, they again may be modified with mutation or any rearrangements in genome. These genes are called ortholog genes. One or more ortholog gene(s) may be duplicated in one lineage; these in-paralog genes are called co-orthologs to the gene in other lineage.

In evolution, ortholog genes are more important than paralog genes, because they are in different species, so they provide information of their common ancestor gene. In theory, when time of common origin increases, ortholog genes become less similar due to different modifications and mutations. There are several methods for distance analyses of ortholog genes. These methods differ in parameters in the way one nucleotide changes into another, like different rates for different types of bases, C+G content bias or base frequencies of sequences. The distance of ortholog genes are used later on to construct phylogenetic trees in taxonomy. All in all ortholog sequences, particularly ortholog genes are the most important sequences in DNA evolution.

CHAPTER2

METHODS AND MATERIALS

2.1 Databases and Tools

The databases we used are ENSEMBL and NGSmethDB. ENSEMBL database is one of the most popular genome browsers and for biological life scientists. It provides information about genomes of hundreds of species and their variation and regulatory data. it has also tools for comparative genomics.

BIOMART, which is part of ENSEMBL database is a data mining tool for extracting sequences or particular information of interest. We used it to download list of orthologous genes of human in 57 species and these genes' positions in Human genome and these genes' biotype and gene status

NGSmethDB is genome wide single cytosine methylation database for human, mouse and arabidopsis. Methylation profile is found by next generation sequencing coupled with bisulfite treatment. NGSmethDB presents methylation data for different tissues and pathological conditions so it allows researchers to analyze methylation for different conditions. In this study we have used Blood Cell Methylation Data for analysis as it was one of the genome wide data available publicly.

Database for Annotation, Visualization and Integrated Discovery (DAVID) [19] and GeneMANIA[20] are functional annotation and association tools. Both tools allow researchers to upload large number of genes at once and then they analyze genes to find common pathways, protein complexes, co-expression, functionally related gene groups, protein-protein and genomic interactions of genes and related genes to reveal functional relations between the genes.

2.2 Data

2.2.1 Genome Sequence Data Used

In our analysis we have used all genomes available in Biomart. List of the species we used in our analysis are given in Table 1. The evolutionary relations between 58 species is updated according to the latest data based on The Tree of Life Web Project and current phylogenetic tree is represented Figure 3.

Ailuropoda melanoleuca	Monodelphis domestica
Anolis carolinensis	Mus musculus
Bos taurus	Myotis lucifugus
Caenorhabditis elegans	Nomascus leucogenys
Callithrix jacchus	Ochotona princeps
Canis familiaris	Oreochromis niloticus
Cavia porcellus	Ornithorhynchus anatinus
Choloepus hoffmanni	Oryctolagus cuniculus
Ciona intestinalis	Oryzias latipes
Ciona savignyi	Otolemur garnettii
Danio rerio	Pan troglodytes
Dasypus novemcinctus	Petromyzon marinus
Dipodomys ordii	Pongo abelii
Drosophila melanogaster	Procavia capensis
Echinops telfairi	Pteropus vampyrus
Equus caballus	Rattus norvegicus
Erinaceus europaeus	Saccharomyces cerevisiae
Felis catus	Sarcophilus harrisii
Gadus morhua	Sorex araneus
Gallus gallus	Spermophilus tridecemlineatus
Gasterosteus aculeatus	Sus scrofa
Gorilla gorilla	Taeniopygia guttata
Homo sapiens	Takifugu rubripes
Latimeria chalumnae	Tarsius syrichta
Loxodonta africana	Tetraodon nigroviridis
Macaca mulatta	Tupaia belangeri
Macropus eugenii	Tursiops truncatus
Meleagris gallopavo	Vicugna pacos
Microcebus murinus	Xenopus tropicalis

Table 1: List of Species available for download from BioMart as of June 2012

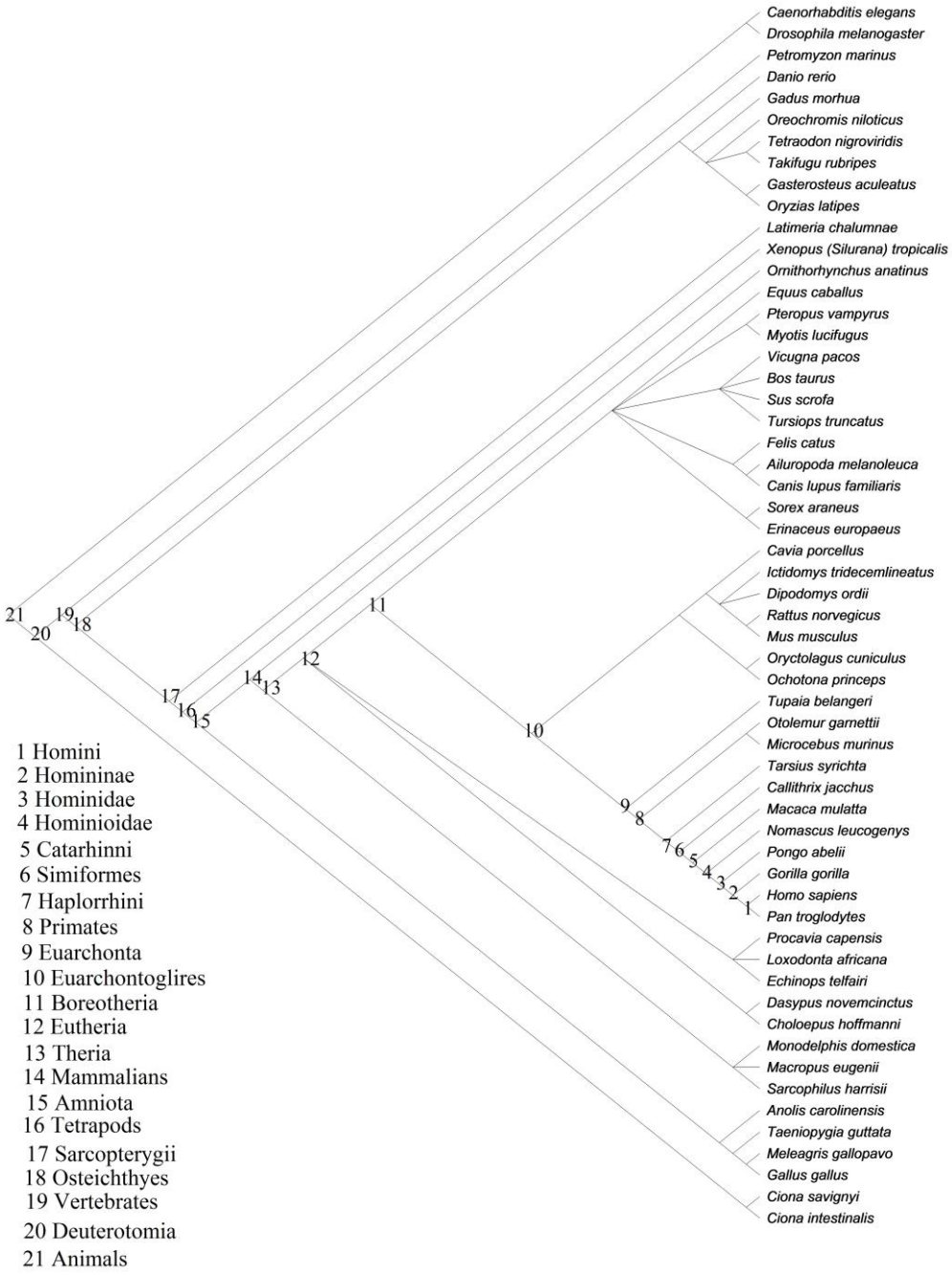


Figure 3: Evolutionary Tree of Life (reconstructed with the information gathered from The Tree of Life Web Project)

2.2.2 Whole Genome Methylation Profile Data

We used two different cell types for methylation analysis. First data coming from Peripheral blood mononuclear cells that have average 24.7 fold coverage (12.3-fold per strand). Methylated CpG'sare found by whole genome bisulfite sequencing. DNA is treated with bisulfite which converts unmethylated cytosines to uracil and leaves methylated cytosines unchanged. Then following PCR converts uracils to thymines. Converted sequence data are read by high throughput sequencing and then aligned to reference genome [21]. In our analysis, over 28,162,537 all CpG's in whole genome, 16,493,829 CpG's of them have more than five coverage were used. [22]

Second data is originated from Newborn human foreskin fibroblasts. This data is generated by bisulfite treatment followed by paired end sequencing too. It covers 24,745,733 individual CpG's that have more than 5 fold coverage. Median coverage is 9. [23]

The additional data available in NGSmethDB weren't preferred for analysis because they were either not genome wide or they covered embryonic cells or they came from same dataset we already included in our study.

2.3 Classification Of Human Genes

Our goal was to compare CpG content and methylation differences in promoters of genes that have evolved at different timelines. We followed pass points in time through the homosapiens lineage. For example after analyzing split of mammals, we went on our analysis through placental mammals' branch where homo sapiens belong to.

All human genes that have at least one ortholog have been grouped according to the passpoint when they have first emerged. So, to find which set of genes that have evolved in a certain time, first we looked at the genes that are present in first split subgroups. Then we extracted genes that are present in other species.

For example Eutherians are thought to split into three groups around same time and then evolved separately. These subgroups are called boreoeutheria, xnartha and afrotheria. Logic is; if a gene is present in any one of the boreoeutheria species and is present in any one of the xnartha species and is present in any one of the afrotheria species; this gene must be present before split. If this gene is absent in all species of non-Eutheria, then this gene must have evolved after common ancestor of Eutheria evolved and before Eutheria lineage split.

The passpoints for the *Homo Sapiens's* phylogeny tree and subgroups evolved after the point that have been used to classify these genes are as listed below:

Metazoa ;	animals
Deuterotomia ;	animals with two openings
Vertebrates ;	animals with backbone
Osteichthyes ;	jawed vertebrates
Tetrapods ;	four-limbed vertebrates
Amniota ;	tetrapods that has terrestially adapted egg
Mammalians;	mammals
Theria ;	mammals giving birth without a shelled egg
Eutheria ;	placental mammals
Boreotheria;	humans, monkeys, rats, hares, cattle, whales, bats, cats..
Euarchontoglires;	primates, three shrew, rodents..
Euarchonta ;	true ancestors consisting of primates, tree shrew, flying lemurs
Primates ;	Lemurs, tarsiers, monkeys, apes, and humans
Hominidae;	Great Apes

Recall that while we were generating gene lists of species from Biomart, we took genes that are ortholog to HomoSapiens genes. Reason behind this is we are interested in CG content and methylation of Human promoter sequences . Note that if more than one gene has orthologous relationship to a gene in HomoSapiens, we calculated the average value. So if a gene has evolved during a certain time and not present in HomoSapiens and it is not included in our study. Also if a gene is present in only one species, that gene has most likely evolved from common ancestor of that species and human. In order to be more confident, by looking at both split subgroups we excluded these genes. The list of genes at each time point is provided in Appendix

Afterconstructing the list of genes, we trimmed them to be more confident about our future analysis. Putative genes, genes that are not assigned in a chromosome but in a contig and non-coding RNA genes except tRNA and rRNA genes are trimmed. Also we have excluded genes coding for

Mt-tRNA (transfer RNA located in the mitochondrial genome)
scRNA (small cytoplasmic RNA)
snRNA (small nuclear RNA)
snoRNA (small nucleolar RNA)
miRNA (microRNA precursors)
misc_RNA (miscellaneous other RNA)
lincRNA (Long intergenic non-coding RNAs)

These genes are almost exclusively annotated computationally and none of them are manually checked. And more importantly, the databases used for annotation for noncoding RNAs are not complete and mostly model organisms are studied. So especially in closer branches to human, these genes gave a significant number of false positives.

CHAPTER 3

RESULTS AND DISCUSSION

3.1 Classification of H.Sapiens Genes At Different Evolutionary Passpoints

Table 2; Distribution of H.sapiens genes according to the passpoint they have emerged.

Split	**Time (million years)**	**# of genes in H.sapiens**
Metazoa*	590	6959
Deuterotomia	530	1751
Vertebrates	505	1304
Osteichthyes	420	1850
Tetrapods	395	281
Amniota	340	435
Mammalians	220	355
Theria	160	505
Eutheria	125	228
Boreotheria		181
Euarchontoglires	100	45
Euarchonta		9
Primates	75	40
Hominidae	15	49

*Note that the species Biomart provides has only one non-animal species, that is fungi S.cerevisiae. So at least few of the 6959 genes that belong to metazoan must be present in other unicellular eukaryotes or even in prokaryotes. Although the dating of these genes is not reliable as others, we know they are “at least” 590 million years old, not newer.

3.2 Total Promoter CpG Frequencies

After the gene lists are generated, we found their positions (chromosome, strand, start-end positions) in human genome; we counted their CpG number in their promoters where 1500 bp upstream from the transcription start site is considered as the promoter region. A Perl code is written to search and count promoter CpG's according to the latest Human Genome assembly GRCh37 Ensembl release 67.

Table 3: Total Promoter CpG Count in genes with different emergence time

Split	Time (million years)	# of genes	Average number of CpG's in promoter
Metazoa	590	6959	50,91
Deuterotomia	530	1751	45,63
Vertebrates	505	1304	43,59
Osteichthyes	420	1850	40,90
Tetrapods	395	281	24,85
Amniota	340	435	26,82
Mammalians	220	355	18,93
Theria	160	505	22,98
Eutheria	125	228	27,27
Boreotheria		181	38,19
Euarchontoglires	100	45	31,89
Euarchonta		9	34,11
Primates	75	40	34,45
Hominidae	15	49	35,14

3.3 Comparison of Methylation Profiles

Next, the methylation profile for each gene promoter is studied. For the analysis of methylation sites in blood cell lines we have counted a Cytosine as methylated if it is read more than 5 times and its methylation ratio is over 0.5 (Table 4 and Figure 4). Although human genome has about 28 million CpG dinucleotides, about 16 million of them is read more than 5, so the analyses was performed over these 16 million CpG's.

Table 4: Methylation Profiles in genes with different emergence time in blood cell line

	Split Time (million years)	# of genes	Average # of CpG's in promoter	Average # of methylated CpG's in promoter
Metazoa	590	6959	50,91	7,32
Deuterotomia	530	1751	45,63	7,73
Vertebrates	505	1304	43,59	7,47
Osteichthyes	420	1850	40,90	7,63
Tetrapods	395	281	24,85	7,13
Amniota	340	435	26,82	7,75
Mammalians	220	355	18,93	6,59
Theria	160	505	22,98	6,83
Eutheria	125	228	27,27	7,49
Boreotheria		181	38,19	7,92
Euarchontoglires	100	45	31,89	3,18
Euarchonta		9	34,11	10,33
Primates	75	40	34,45	9,20
Hominidae	15	49	35,14	6,82

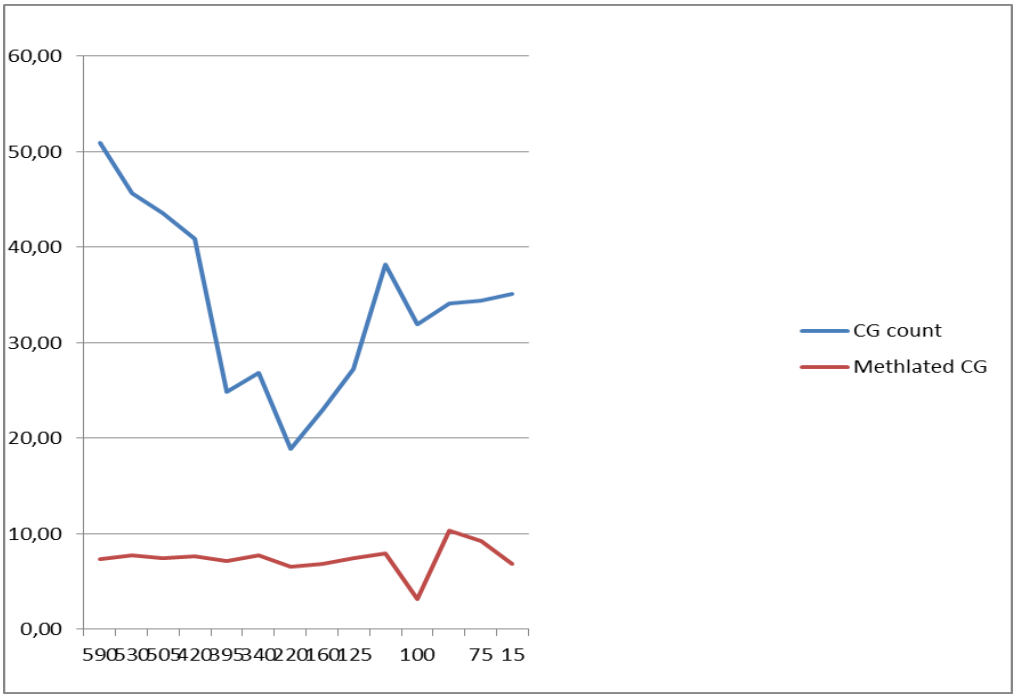


Figure 4 : Total CpG and methylated CpG in time

3.4 COMMON GENES

3.4.1 Identification of Common Genes

The genes that are old and common among in all species under a group, those genes are more likely a fundamental function. After looking at CpG number and methylated CpG number in promoters of genes that were categorized by time, we then analyzed common genes that are present in all species in a group, in order to reveal any relation to CpG count, methylation status and functionality.

To test this, first we grouped species again and picked the genes which are present in all species in that group. But if a common gene in a group is present in another species, that gene is discarded. For example all vertebrates have also all genes that are common in animals, and all primates also have common genes that are common in vertebrates. So we discarded genes that are present in any species that are not in a particular group and the result was not only common but also exclusive genes for that group. Again we traced all nodes of branches that *Homo Sapiens* belong to. For genes' promoters, we found average CpG number, average methylated Cytosine in both blood cell line and fibroblast cell line.

Table 5 : Distribution of common genes exclusive for each subgroup**.**

Common Genes among	Shared among number of species	number of genes	Average number of CpG in promoter region	Average number of methylated CpG in promoter in	
				Blood Cell	Fibroblast Cell
Fungi + animals	**58**	**234**	**45.31**	**7.60**	**10.98**
animals	**57**	**174**	**54.41**	**6.88**	**9.19**
Deuterotomia	**55**	**56**	**48.10**	**8.17**	**11.23**
Vertebrates	**53**	**71**	**40.05**	**7.38**	**9.47**
Osteichthyes	**52**	**33**	**43.54**	**7.90**	**9.87**
Sarcopterygii	45	1	80	4	1
Tetrapods	44	0	0	0	0
Amniota	43	3	22.33	15.33	16
Mammalians	39	2	33.50	4.50	7.50
Theria	38	2	6.50	4	5
Eutheria	35	1	68	11	18
Boreotheria	30	0	0	0	0
Euarchontoglires	18	0	0	0	0
Euarchonta	11	0	0	0	0
Primates	10	0	0	0	0
Haplorrhini	8	0	0	0	0
Simiformes	7	6	14.83	7.50	9.67
Catarhinni	6	3	18.33	10	10
Hominioidae	**5**	**17**	**37.52**	**7.88**	**9.82**
Hominidae	**4**	**31**	**34.90**	**7.35**	**12.61**
Homininae	**3**	**36**	**48.25**	**8.27**	**13.55**
Homini	**2**	**52**	**47.78**	**5.88**	**10.86**

After finding gene lists we again trimmed them to be confident about our analysis, all pseudogenes, noncoding RNA genes (except tRNA and rRNA), genes that are not assigned to a chromosome and putative genes are discarded. Because chimpanzee and human are studied more than many organisms, non-coding RNAs gave false positives especially in Homini branch. There is little to none common genes after Osteichthyes split and until Apes appear, so we have disregarded those groups of genes for this analysis..

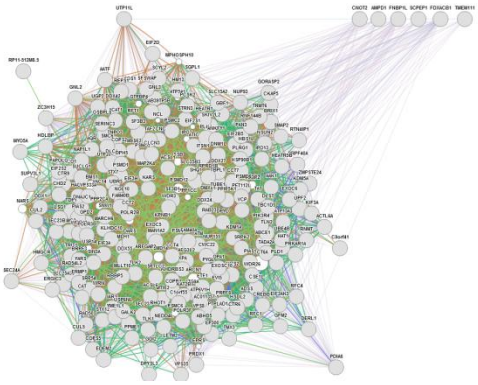
CpG methylation counts were collected from the data for two different cell types, blood and fibroblast. Methylation profiles seem to be similar in two different tissues. So results indicate that whether a gene is old and common among almost all species, or new and common among very few species, does not seem to affect percentage of CpG units on their promoter sequences or their methylation profile.

3.4.2 Functional Annotation of Common Genes

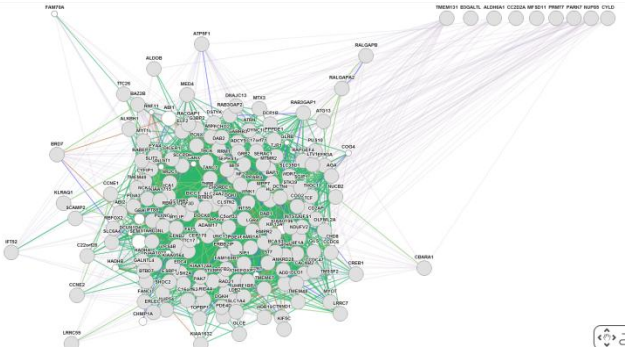
As expected the genes common between all 58 species were group of fundamental genes involved protein phosphorylation, ribonucleotide binding genes (Figure 3). Along with other fundamental processes common genes exclusive to animals showed genes with cytoskeletal function. In vertebrates we observed emerging of glycosylation and signaling processes.

When the overall interactions within each group of genes are analyzed we have seen that the connectivity of the networks (co-expression, common domain, pathways etc…) decreased as we have move towards more recent species. Many genes common to Hominidae and Homini were not functionally annotated. We believe that further investigation of these genes computationally and experimentally could reveal important information about biological processes emerged through the evolution of homo sapiens.

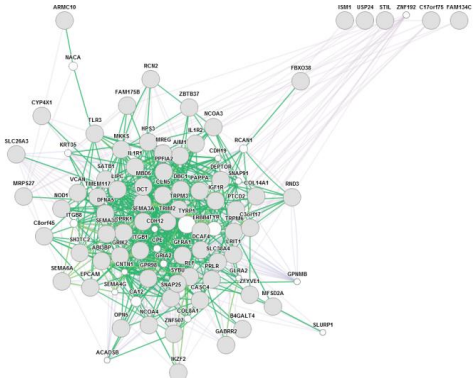
All Common



Animal



Vertebrates



Hominioidae

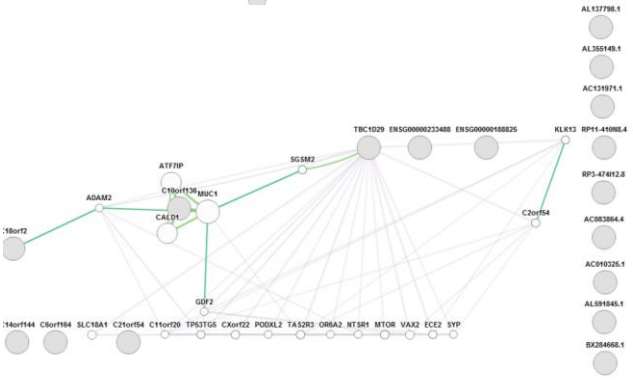
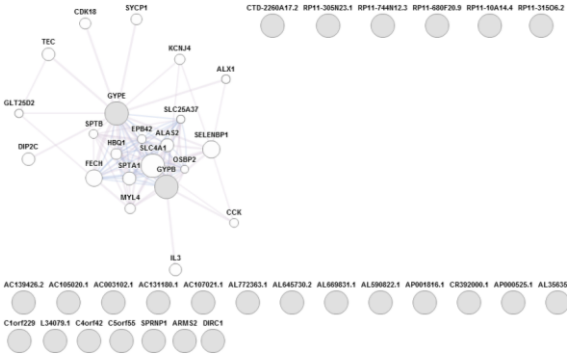


Figure 5: Network Representation for the common and exclusive ranging from all 58 species to Homini

Figure 5 (cont.)

Homini



ALL COMMON

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	SP_PIR_KEYWORDS	phosphoprotein	RT		157	67.7	2.1E-20	2.9E-18
	SP_PIR_KEYWORDS	acetylation	RT		92	39.7	1.4E-22	3.7E-20
	SP_PIR_KEYWORDS	nucleus	RT		84	36.2	1.4E-6	7.6E-5
	GOTERM_MF_FAT	nucleotide binding	RT		75	32.3	1.8E-13	2.5E-11
	GOTERM_MF_FAT	purine nucleotide binding	RT		72	31.0	2.0E-15	8.6E-13
	GOTERM_MF_FAT	purine ribonucleotide binding	RT		70	30.2	2.7E-15	5.7E-13
	GOTERM_MF_FAT	ribonucleotide binding	RT		70	30.2	2.7E-15	5.7E-13
	SP_PIR_KEYWORDS	nucleotide-binding	RT		66	28.4	5.9E-18	5.3E-16
	SP_PIR_KEYWORDS	cytoplasm	RT		66	28.4	2.8E-5	1.2E-3
	GOTERM_CC_FAT	membrane-enclosed lumen	RT		64	27.6	2.1E-11	5.7E-9
	GOTERM_CC_FAT	intracellular organelle lumen	RT		62	26.7	3.3E-11	4.6E-9
	GOTERM_CC_FAT	organelle lumen	RT		62	26.7	8.6E-11	7.9E-9
	GOTERM_MF_FAT	nucleoside binding	RT		61	26.3	6.3E-13	6.8E-11
	GOTERM_MF_FAT	purine nucleoside binding	RT		60	25.9	1.6E-12	1.4E-10
	GOTERM_MF_FAT	adenyl nucleotide binding	RT		58	25.0	1.0E-11	7.3E-10
	GOTERM_MF_FAT	adenyl ribonucleotide binding	RT		56	24.1	1.4E-11	8.8E-10
	GOTERM_MF_FAT	ATP binding	RT		55	23.7	2.8E-11	1.5E-9
	GOTERM_CC_FAT	nuclear lumen	RT		53	22.8	3.3E-10	2.3E-8
	SP_PIR_KEYWORDS	atp-binding	RT		50	21.6	1.1E-12	7.2E-11
	GOTERM_CC_FAT	intracellular non-membrane-bounded organelle	RT		50	21.6	2.7E-2	7.8E-1
	GOTERM_CC_FAT	non-membrane-bounded organelle	RT		50	21.6	2.7E-2	7.8E-1

ANIMALS

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	SP_PIR_KEYWORDS	phosphoprotein	RT		116	66.7	1.5E-14	3.9E-12
	UP_SEQ_FEATURE	splice variant	RT		110	63.2	1.3E-10	9.0E-8
	SP_PIR_KEYWORDS	alternative splicing	RT		109	62.6	3.0E-10	3.9E-8
	SP_PIR_KEYWORDS	cytoplasm	RT		47	27.0	1.4E-3	1.2E-1
	GOTERM_CC_FAT	intracellular non-membrane-bounded organelle	RT		36	20.7	1.5E-2	8.4E-1
	GOTERM_CC_FAT	non-membrane-bounded organelle	RT		36	20.7	1.5E-2	8.4E-1
	SP_PIR_KEYWORDS	coiled coil	RT		31	17.8	4.0E-3	2.3E-1
	UP_SEQ_FEATURE	mutagenesis site	RT		31	17.8	5.3E-3	8.4E-1
	GOTERM_CC_FAT	cytoskeleton	RT		26	14.9	1.0E-3	2.3E-1
	SP_PIR_KEYWORDS	disease mutation	RT		26	14.9	4.3E-3	2.0E-1
	GOTERM_CC_FAT	membrane-enclosed lumen	RT		26	14.9	4.1E-2	9.7E-1
	GOTERM_CC_FAT	organelle lumen	RT		25	14.4	5.5E-2	9.7E-1
	GOTERM_CC_FAT	intracellular organelle lumen	RT		24	13.8	7.2E-2	9.8E-1
	SP_PIR_KEYWORDS	transport	RT		22	12.6	7.3E-2	9.6E-1

Figure 6: Functional annotation charts representing the biological processes and molecular function each group of common genes are involved

Figure 6 (cont.)

VERTEBRATE

Sublist	Category	Term	RT	Genes	Count	%	P-Value	Benjamini
	SP_PIR_KEYWORDS	polymorphism	RT		49	69.0	8.9E-2	9.0E-1
	SP_PIR_KEYWORDS	phosphoprotein	RT		36	50.7	2.4E-2	5.0E-1
	SP_PIR_KEYWORDS	membrane	RT		35	49.3	3.8E-3	1.2E-1
	SP_PIR_KEYWORDS	alternative splicing	RT		34	47.9	9.9E-2	8.9E-1
	GOTERM_CC_FAT	intrinsic to membrane	RT		33	46.5	7.5E-4	9.3E-2
	GOTERM_CC_FAT	integral to membrane	RT		32	45.1	1.0E-3	6.3E-2
	SP_PIR_KEYWORDS	transmembrane	RT		30	42.3	3.2E-3	1.3E-1
	UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	RT		29	40.8	3.8E-4	6.5E-2
	SP_PIR_KEYWORDS	glycoprotein	RT		29	40.8	7.2E-4	5.9E-2
	UP_SEQ_FEATURE	transmembrane region	RT		29	40.8	6.0E-3	4.2E-1
	SP_PIR_KEYWORDS	signal	RT		26	36.6	1.1E-4	1.9E-2
	UP_SEQ_FEATURE	signal peptide	RT		26	36.6	1.3E-4	4.4E-2
	SP_PIR_KEYWORDS	disulfide bond	RT		21	29.6	3.1E-3	1.6E-1
	UP_SEQ_FEATURE	topological domain:Cytoplasmic	RT		21	29.6	1.6E-2	6.9E-1
	UP_SEQ_FEATURE	disulfide bond	RT		20	28.2	5.1E-3	4.6E-1

3.5 Analysis of Conservation and Methylation of Individual CpG Units

After analyzing promoters' CG content and methylation profile with their genes' appearance time and conservation status, we followed a different approach for investigation of individual CpG dinucleotides. Our goal was to show whether methylated Cytosines are conserved throughout time and if there is any condition that supports the conservation of methylated Cytosines.

Firstly we investigated if there is any difference in mutation frequencies of CpG's between hypo and hyper methylated cytosines. In this analysis we have only compared orthologous human and mouse promoter sequences on human chromosome 1. We have identified all methylated and unmethylated cytosines in all human gene promoters that have an ortholog in mouse. Then for each individual CpG dinucleotide, we looked at the mouse counterpart at the exact location based on ENSMEBL Human vs. Mouse Genome alignment.

For example; Human gene FURIN (Ensembl version ENSG00000140564.6) has a mouse ortholog. We took the promoter of FURIN and blasted that sequence.

Exact match is :

homo_sapiens:15 > chromosome:GRCh37:15:91425425:91426925:1
mus_musculus:7 > chromosome:NCBIM37:7:87533994:87535311:-1
chromosome:NCBIM37:7:87533558:87533739:-1

```
homo_sapiens/      TGGGATTCCTGACCCAGGCCGCAGCTCT-----
mus_musculus/      --GAGTCCCTGA-CCAGGCTGCAGCCCTGCCCC

homo_sapiens/      TGCCCTTCCCTGTCCCTC-TAAAGCAATAATGG
mus_musculus/      TTCCCTTCCCTGCCCCTCAGAAAGCAATAATGG

homo_sapiens/      -TCCCATCCAGGCAGTCGGG-GGCTGGCCTAGG
mus_musculus/      TTCCCATCCAGGCAACAGGGAGGCTGGCC-AGG
```

Here the first 99 bases of Homo Sapiens FURIN gene promoter i s shown and its ortholog sequence in mouse. After finding alignment, we found which individual CpG's are methylated and which are unmethylated. In the above alignment, yellow CpG are methylated and CG – TG is calculated in methylated CpG analysis, on the other hand blue CpG are not methylated in humans and CG – AG are calculated in nonmethylated analysis.

Human vs. Mouse Whole Genome Alignment data is downloaded from Ensembl and it is done by using Blastz and Lastz [24,25,26]. For this analysis, we accepted a cytosine as methylated if its methylation ratio is over 0.8 and accepted a cytosine as unmethylated if its methylation ratio is below 0.2.

Results are shown in Table 6. (first16 numbers, other very few conversions are not shown) First columns show which dinucleotide mouse has at the aligned position of human CG. ".." means no alignment is found and "—"means gap. It is clear that unmethylated CpG's are conserved compared to methylated CpG's . Only a very small fraction of methylated CpG's is same between human and mouse.

Table 6: Methylated and Unmethylated CpG conversion count in Human Chromosome 1 vs. aligned Mouse sequences

Methylated CpG conversion	
..	4287
TG	544
CA	497
--	348
CG	312
CT	190
AG	182
GG	150
TA	135
CC	130
AA	70
-G	64
C-	54
TT	53
GA	47
TC	42

Unmethylated CpG conversion	
CG	3820
..	3306
TG	1400
CA	1364
--	1316
GG	702
CC	680
AG	652
CT	645
TA	272
-G	240
C-	230
AA	164
TT	155
GA	134
TC	124

In order to search if there could be some important methylated CpG's, we decided to investigate methylated CpG's conservation on promoter sequences with different total methylated CpG number. Our main goal here was to reveal if the methylation status of the neighboring regions had an effect on conservation of methylated CpG's. We have examined the differences in the conservation status of methylated CpG's according to the its presence on a promoter with low number of highly methylated CpG's or high number of highly methylated CpG's.

We have grouped promoters according to their number of highly methylated CpG numbers in a genome wide fashion without taking account of total CG number in a promoter or lowly methylated CG number in a promoter. First group of promoters had 1 to 5 highly methylated CpG's; second group 6-10 and so on and the fifth group had more than 21 highly methylated CpG's (Table 7)

We again looked at individual CpGs' counterpart in mouse human alignment and results are as follows

Table 7: Methylated CpG conversion count in different methylation profile of promoters.

# of highly methylated CpG's in a promoter	# of mouse counterpart of highly methylated human CpG dinucleotide				
	no alignment	TG	CA	gap	CG
1-5	13180	1846	1824	1147	837
6-10	18563	1935	1808	1120	951
11-15	7498	614	671	382	442
16-20	1777	124	115	65	124
>20*	337	47	28	30	60

*we took all promoters that have methylated CpG number over 20 together because there are only a few of them. They could not be split more for the sake of statistical analysis

Table 8: Methylated CpG conversion rate in different methylation profile of promoters

# of highly methylated CpG's in a promoter	% of mouse counterpart of highly methylated Human CpG dinucleotides					
	CG	TG	CA	CT	AG	TA
1-5	0,082366	0,181657	0,179492	0,068392	0,066424	0,051565
6-10	0,090202	0,183534	0,171488	0,068292	0,06336	0,049322
11-15	0,119621	0,166171	0,181597	0,064953	0,067659	0,040866
16-20	0,174648	0,174648	0,161972	0,060563	0,067606	0,025352
>20	0,26087	0,204348	0,121739	0,047826	0,03913	0,021739

The Table 8 represents the ratios of conversion of methylated CpGs into other dinucleotides. Only first 6 dinucleotides that have the highest ratios are shown. Because we are interested in dinucleotide conversion, while calculating ratios, we didn't consider the –no alignment‖, as they don't provide any information on conversion rates. As seen, probability of CpG conservation increases while methylated CpG increases and its conservation ratio increases more when there are high numbers of methylated CpG s. The statistical significance of the differences observed have been analyzed with the Friedman test.

The result of Friedman rank sum test:
Friedman chi-squared = 13.25, df = 4, p-value =0.01012

As p-value is <0.05 we can say that there is a significant difference between the groups. Next the multiple comparison tests are conducted to determine exactly which groups are significantly different.

Multiple comparisons between groups after Friedman test: p.value: 0.05

Comparisons

	obs.dif	critical.dif	difference
1-2	3	25.10687	FALSE
1-3	3	25.10687	FALSE
1-4	9	25.10687	FALSE
1-5	26	25.10687	**TRUE**
2-3	6	25.10687	FALSE
2-4	12	25.10687	FALSE
2-5	29	25.10687	**TRUE**
3-4	6	25.10687	FALSE
3-5	23	25.10687	FALSE
4-5	17	25.10687	FALSE

This analysis shows that the most significant difference is between groups 5-1 and 5- 2. Group5 is the promoter group that has more than 20 highly methylated CpG's. So we can say that as methylated CG dinucleotides increases, the conservation of CpG's increases and this phenomenon is significant when there are more than 20 methylated CpG in a 1500 base pair length DNA.

A very interesting observation that worth a notice is TpG ratio doesn't decrease in Group 5 in contrast to every other dinucleotide including CpA. We know that CpG dinucleotides can be converted to TpG by oxidative deamination without a foreign agent[24], so continuous TpG formation from CpGs coupled with decreased mutation rate of all dinucleotides may again mean that the promoter regions that have a high number of highly methylated CpG are in a closed formation.

CHAPTER 4

CONCLUSION

Among the 16 dinucleotides, CG covers only %0.09 of a human genome, it has been an area of research for a long time. Their role is found to be prevalent in human promoters, we have decided to analyze if their distribution varies in different group of genes with an evolutionary perspective.

Our first approach was to group genes according to their emergence time. We found out that old genes tend to have more unmethylated CpG's. CpG content of promoters tend to decrease gradually until mammalians appear, then increases again. However there is almost no significant difference in average number of methylated CpG except the genes appeared about 100 million years ago. As the 100 mya mark remains as an outlier, overall tendency of constant methylation of genes regardless of time and total CpG count implies that time and total CpG content are not major factors in methylation profile of genes. We know that hyper methylation of promoters induces silencing of genes, so it is reasonable that many different types of genes must have appeared in a period so we don't see a pattern. Yet, why the genes that appeared during era of common ancestor of rodents and primates have so little methylation can be addressed in future studies.

After observing that emergence time of genes has little to no effect on methylation, we thought that conservation status of a gene may have been effecting methylation. The logic was if a gene is common among many species, that gene must have an important function and those genes' promoter methylation status may be different. Important genes for this analysis are the most common genes compared to new ones. We used newer genes to compare with the most common ones, however again there was no significant difference. Results show that average unmethylated CG number of most common genes' promoters is more than the average unmethylated CG number in promoters of genes closest to Homo Sapiens branch. There is like an average threshold that genes got methylated up to. This is almost same result we got from our initial analysis based on time. So here we can conclude that whether a random gene is common and have an important function or not, it may have a similar epigenetic regulation with another random gene.

While making these two analyses, because of unavailability of data, we couldn't consider promoter sequences separate from coding sequences. We know that genes and promoters may have evolved at different time points. But again methylation is related to expression of genes, so not time but being common and having a fundamental function must have shown a difference if there was any, even without analyzing promoters individually.

We then analyzed individual methylated CpG's. Every highly methylated CpG in a promoter is assigned to their corresponding genes and aligned to mouse genome. After the analysis, we have found that if not lost, methylated CpG dinucleotides in human have turned to mainly TG and CA dinucleotides along with other bases as expected. In contrast unmethylated CpG's tend not to change if not lost. This observation was confirmative with the current literature.

Next, we tested conservation of CpG's on promoters with different methylation profiles. Without considering its age or conservation among species, we grouped genes according to methylation content of their promoters. We thought that if we could find difference in mutation frequencies of individual CpG's, we could also present an evidence for the influence of methylation and evolution to each other.

In this study, we have showed that transition rates of CpG's in lowly methylated promoters are different from highly methylated promoters. We suggest that this is due to interaction of histone deacetylation proteins and DNA Methylases. High methylated regions will condense and as it is shown here; their mutation frequency will be lower. That also means that any bases around this area will be less prone to mutation as well. Many molecular DNA evolution models provide generic rates for all parts of genome, although it is known that genome evolves at different rates at different areas [28]. If studied extensively, these methylated CpG rich areas provide different rates which will help a better understanding of evolution timeline. Varying rates for different areas will provide better results and these methylated CpG rich areas mutates slowly which makes them a very good candidate for long time evolution analysis. Furthermore the topology of methylated CpG's and nonmethylated CpG's can also be a research topic to find location of nucleosomes.

REFERENCES

1. Adrian Bird (2002) DNA methylation patterns and epigenetic memory Genes Dev. 2002 16: 6-21

2. Brero, A., Easwaran, H.P., Nowak, D., Grunewald, I., Cremer, T.,Leonhardt, H. and Cardoso, M.C. (2005) Methyl CpG-binding proteins induce large-scale chromatin reorganization during terminal differentiation. J. Cell Biol., 169, 733–743

3. Kangaspeska S, Stride B, Métivier R, Polycarpou-Schwarz M, Ibberson D, Carmouche RP, Benes V, Gannon F, Reid G (2008). Transient cyclical methylation of promoter DNA. Nature. 2008 Mar 6;452(7183):112-5

4. Gardiner-Garden M, Frommer M. (1987) CpG islands in vertebrate genomes. J Mol Biol. Jul 20;196(2):261-82.

5. Takai D, Jones PA. (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. Proc Natl Acad Sci U S A. 2002 Mar 19;99(6):3740-5. Epub 2002 Mar 12.

6. G. P. Pfeifer (2006) Mutagenesis at Methylated CpG Sequences. CTMI (2006) 301:259–281

7. Miho M. Suzuki1 & Adrian Bird (2008) DNA methylation landscapes: provocative insights from epigenomics. Nature Reviews Genetics 9, 465-476

8. J Boyes and A Bird (1992) Repression of genes by DNA methylation depends on CpG density and promoter strength: evidence for involvement of a methyl-CpG binding protein. EMBO J. 1992 January; 11(1): 327–333

9. Courtesy of Richard Twyman, Wellcome Trust

10. Tate PH, Bird AP. (1993) Effects of DNA methylation on DNA-binding proteins and gene expression. Curr Opin Genet Dev. 1993 Apr;3(2):226-31.

11. Nan,X., Campoy,F.J. and Bird,A. (1997) MeCP2 is a transcriptional repressor with abundant binding sites in genomic chromatin. Cell, 88, 471–481.

12. Jones,P.L., Veenstra,G.J.C., Wade,P.A., Vermaak,D., Kass,S.U., Landsberger,N., Strouboulis,J. and Wolffe,A.P. (1998) Methylated DNA and MeCP2 recruit histone deacetylase to repress transcription. Nature Genet., 19, 187–191.

13. Nan,X., Ng,H.-H., Johnson,C.A., Laherty,C.D., Turner,B.M., Eisenman,R.N. and Bird,A. (1998) Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature, 393, 386–389.

14. Keith D. Robertson and Peter A.Jones (2000) DNA methylation: past, present and future directions. Carcinogenesis, 21 (3): 461-467.

15. Clark SJ, Melki J. (2002) DNA methylation and gene silencing in cancer: which is the guilty party? Oncogene. 2002 Aug 12;21(35):5380-7.

16. Fitch WM. 1970. Distinguishing homologous from analogous proteins. Syst. Zool.19:99–10

17. Eugene V. Koonin (2005) Orthologs, Paralogs, and Evolutionary Genomics. Annu. Rev. Genet.2005. 39:309–38

18. Courtesy of http://genomevolution.org/

19. http://david.abcc.ncifcrf.gov/

20. Genemania.org

21. Pauline A. Callinan and Andrew P. Feinberg. (2006) The emerging science of epigenomics. Human Molecular Genetics, 2006, Vol. 15, Review Issue No. 1

22. Yingrui Li, Jingde Zhu, Geng Tian, Ning Li, Qibin Li, Mingzhi Ye, Hancheng Zheng, Jian Yu, Honglong Wu, Jihua Sun, Hongyu Zhang, Quan Chen, Ruibang Luo, Minfeng Chen, Yinghua He, Xin Jin, Qinghui Zhang, Chang Yu, Guangyu Zhou, Jinfeng Sun, Yebo Huang, Huisong Zheng, Hongzhi Cao, Xiaoyu Zhou, Shicheng Guo, Xueda Hu, Xin Li, Karsten Kristiansen, Lars Bolund, Jiujin Xu, Wen Wang, Huanming Yang, Jian Wang, Ruiqiang Li, Stephan Beck, Jun Wang, Xiuqing Zhang (2010) The DNA Methylome of Human Peripheral Blood Mononuclear Cells. Plos Biol. 2010 Nov9:8(11):E1000533

23. Louise Laurent,Eleanor Wong, Guoliang Li,Tien Huynh, Aristotelis Tsirigos,Chin Thing Ong,Hwee Meng Low, Ken Wing Kin Sung,Isidore Rigoutsos, Jeanne Loring, and Chia-Lin Wei (2010) Dynamic changes in the human methylome during differentiation . Genome Res. 2010 March; 20(3): 320–331.

24. W. James Kent,Robert Baertsch, Angie Hinrichs, Webb Miller and David Haussler (2003) Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. PNAS vol. 100 no. 20 11484-11489

25. Scott Schwartz, W. James Kent, Arian Smit, Zheng Zhang, Robert Baertsch, Ross C. Hardison, David Haussler, and Webb Miller (2002) Human–Mouse Alignments with BLASTZ. Genome Res. 2003. 13: 103-107

26. Harris, R.S. (2007) Improved pairwise alignment of genomic DNA. Ph.D. Thesis, The Pennsylvania State University.

27. Saura C. Sahu, Toxicology and Epigenetics. 1.Edition 2012, P.280

28. Liqing Zhang and Wen-Hsiung Li (2004) Mammalian Housekeeping Genes Evolve More Slowly than Tissue-Specific Genes. Mol Biol Evol (2004) 21 (2): 236-239

APPENDIX

The supplementary material of this study is provided in elecctronic format in one CD, organization of the contents in the CD is as follows

Folder A : human ortholog genes of all species that belong to a classification (for all nodes in Figure 1) and human ortholog genes of all species that do not belong to a classification (for all nodes in Figure 1)

Folder B: common genes in all animals in a classification (for all nodes in Figure 1)

Folder C: genes which are present in any one of the species that does not belong to a classification (for all nodes in Figure 1)
- for example: for Deuteromia; all genes that is present in any of C.elegans, D.melanogater and S.cerevisiae

Folder D: genes which are present in any one of the species that "does" belong to a classification (for all nodes in Figure 1)

Folder E: genes that emerged at a given node (genes that are shown in Chapter 3.1)

Folder F: all CG and methylated CG number of promoters of genes that emerged at a given time (genes that are shown in chapter 3.2 and 3.3)

Folder G: genes that are common among and exclusive to a classification and their promoter CG content (coloumn 4 of Table 5 in chapter 3.4). Column 5 and 6 are calculated by calculating which ones of the CG's are methylated in these promoters of exclusive genes (positions of methylated genes are taken from NGSmethDB)

Folder H: list of human genes that have orthlogs in mouse and their positions in genome (we used these lists for the analyses in chapter 3.5). Mouse human genome wide alignment we used for these analyses is about 5,5 GB and latest version is avaliable in http://www.ensembl.org/info/docs/compara/index.html

Latest methylation scores of all single CpG's in genome in blood and fibroblast tissue of homosapiens is avaliable at bioinfo2.ugr.es/meth/NGSmethDB.php?see=content

TEZ FOTOKOPİ İZİN FORMU

ENSTİTÜ

Fen Bilimleri Enstitüsü ☐

Sosyal Bilimler Enstitüsü ☐

Uygulamalı Matematik Enstitüsü ☐

Enformatik Enstitüsü ☐

Deniz Bilimleri Enstitüsü ☐

YAZARIN

Soyadı : Burak
Adı : Demiralay
Bölümü: Bioenformatik

TEZİN ADI (İngilizce) :

Investigation Of Human Promoter CpG Content And Methylation Profiles At Different Conservation Levels

TEZİN TÜRÜ : Yüksek Lisans ☐ Doktora ☐

1. Tezimin tamamı dünya çapında erişime açılsın ve kaynak gösterilmek şartıyla tezimin bir kısmı veya tamamının fotokopisi alınsın. ☐
2. Tezimin tamamı yalnızca Orta Doğu Teknik Üniversitesi kullanıcılarının erişimine açılsın. (Bu seçenekle tezinizin fotokopisi ya da elektronik kopyası Kütüphane aracılığı ile ODTÜ dışına dağıtılmayacaktır.) ☐
3. Tezim bir (1) yıl süreyle erişime kapalı olsun. (Bu seçenekle tezinizin fotokopisi ya da elektronik kopyası Kütüphane aracılığı ile ODTÜ dışına dağıtılmayacaktır.) ☐

Yazarın imzası ..... Tarih .....