

8-2011

PREDICTION OF DNA METHYLATION BASED ON GENOMIC ARCHITECTURE AND APPLICATIONS OF POSITIONAL WEIGHT MATRICES

Juan Gallegos

Follow this and additional works at: https://digitalcommons.library.tmc.edu/utgsbs_dissertations



Part of the [Biostatistics Commons](#)

Recommended Citation

Gallegos, Juan, "PREDICTION OF DNA METHYLATION BASED ON GENOMIC ARCHITECTURE AND APPLICATIONS OF POSITIONAL WEIGHT MATRICES" (2011). *The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access)*. 179.

https://digitalcommons.library.tmc.edu/utgsbs_dissertations/179

This Thesis (MS) is brought to you for free and open access by the The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences at DigitalCommons@TMC. It has been accepted for inclusion in The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact digitalcommons@library.tmc.edu.

PREDICTION OF DNA METHYLATION BASED ON GENOMIC ARCHITECTURE AND
APPLICATIONS OF POSITIONAL WEIGHT MATRICES

by

Juan Gallegos, B.S.

APPROVED:

Shoudan Liang, Ph.D

Thomas J. Goka, Ph.D

Yuan Ji, Ph.D

Peter Mueller, Ph.D

Allen R. White, Ph.D

APPROVED:

Dean, The University of Texas
Graduate School of Biomedical Sciences at Houston

PREDICTION OF DNA METHYLATION BASED ON GENOMIC ARCHITECTURE AND
APPLICATIONS OF POSITIONAL WEIGHT MATRICES

A

THESIS

Presented to the Faculty of
The University of Texas
Health Science Center at Houston
and
The University of Texas
M. D. Anderson Cancer Center
Graduate School of Biomedical Sciences
in Partial Fulfillment
of the Requirements
for the Degree of
MASTER OF SCIENCE

by

Juan Gallegos, B.S.
Houston, Texas

August 2011

DEDICATION

To: Dr. Shoudan Liang

An incredible advisor, friend, role model, and one of the great individuals I have ever had the pleasure to meet. I am in debt to you and I thank you for never giving up on me and always guiding me back to the path that has brought me here today. I will always hold you in my highest regards as there is no other person that has had as much compassion, care, and understanding for me than you. If I were asked today, who is the person I look the most up to, my answer would be you.

ACKNOWLEDGEMENTS

I would like to acknowledge Dr. Thomas J. Goka, Dr. Weiner, and Dr. Knutson for their efforts and always having the door open to speak to me when I needed it the most. I thank you all and especially Dr. Goka for always advocating on my behalf and always encouraging me to do my best. Finally, I would like to acknowledge Dr. Marcos R. Estecio who was instrumental in my development as a researcher. Marcos I am also in debt to you for all the time and patience you lend me during my time at GSBS.

PREDICTION OF DNA METHYLATION BASED ON GENOMIC ARCHITECTURE AND
APPLICATIONS OF POSITIONAL WEIGHT MATRICES

Publication No. _____

Juan Gallegos, M.S.

Supervisory Professor: Shoudan Liang, Ph.D.

Gene silencing due to epigenetic mechanisms shows evidence of significant contributions to cancer development. We hypothesis that the genetic architecture based on retrotransposon elements surrounding the transcription start site, plays an important role in the suppression and promotion of DNA methylation. In our investigation we found a high rate of SINE and LINEs retrotransposon elements near the transcription start site of unmethylated genes when compared to methylated genes. The presence of these elements were positively associated with promoter methylation, contrary to logical expectations, due to the malicious effects of retrotransposon elements which insert themselves randomly into the genome causing possible loss of gene function. In our genome wide analysis of human genes, results suggested that 22% of the genes in cancer were predicted to be methylation-prone; in cancer these genes are generally down-regulated and function in the development process. In summary, our investigation validated our hypothesis and showed that these widespread genomic elements in cancer are highly associated with promoter DNA methylation and may further participate in influencing epigenetic regulation.

TABLE OF CONTENTS

Approval Page	i
Title Page	ii
Dedication	iii
Acknowledgements	iv
Abstract	v
Table of Contents	vi
List of Illustrations	viii
List of Tables	iv
Chapter 1: Introduction	1
1.1 Epigenetics	1
1.2 CpG Islands	3
1.3 DNA Methylation	4
1.4 Retrotransposons	7
1.5 SINE and LINE	8
Chapter 2: Materials and Methods	10
2.1 Positional Weight Matrices	10
2.2 Calculation of Score to Predict Gene Promoter Predisposition to DNA Methylation	10
Chapter 3: Results and Discussion	15
Chapter 4: Conclusion and Future directions	37

4.1 Titile of Sub Chapter	1
4.2 Titile of Sub Chapter	1
4.3 Titile of Sub Chapter	1
4.3.1 Titile of Sub Chapter	1
4.4 Titile of Sub Chapter	1
References	39
Vita	44

List of Illustrations

CHAPTER 1

Figure 1: Epigenetic Modifications of DNA 4

Figure 2: Knudson's two-hit hypothesis 6

CHAPTER 2

Figure 3: SINE and LINE abundance score to predict gene predisposition to methylation in cancer. 11

Figure 4: Graphic representation of the presence of SINE and LINE in 1-KB bins surrounding the transcription start site (TSS) for methylation-prone and methylation-resistant genes 12

CHAPTER 3

Figure 5: Distribution of repetitive elements in methylation-prone versus methylation-resistant genes. 19

Figure 6: Prediction of gene predisposition and resistance to hypermethylation in cancer 24

Figure 7: Distribution of SINE, LINE and LTR repeats around the TSS of frequently hypermethylated and unmethylated genes identified by MCAM analysis of 28 cancer cell lines and 32 primary cancer tissues 26

Figure 8: Frequency of retroelements in methylation-prone and methylation-resistant genes identified in mouse cancer models and old mice.. 27

Figure 9: Genome-wide prediction of predisposition to DNA methylation in cancer 29

Figure 10: Genome architecture influences on PcG protein binding in embryonic and differentiated cells.. 34

CHAPTER 4 (no illustrations)

List of Tables

Table 1: Training Set	16
Table 2: Methylation profile of the training set in nine cancer cell lines from different tissue origin	17
Table 3: Methylation-prone genes in cancer used as test set.....	21
Table 4: Methylation-resistant genes in cancer used as test set	22
Table 5: Top 50 predicted Methylation-prone genes	31
Table 6: Top 50 predicted Methylation-resistant genes	32

INTRODUCTION

1.1 Epigenetics

Previously scientist have attributed a person phenotypic's characteristics based solely on the composition of the persons genomic DNA. It is through the genetic composition and alterations that we have studied the neoplastic evolution and were confounded to the idea that cancer was mainly a disease of genetics. Recently, the study of external influences on the DNA has come to question this idea in favor of a much larger complex mechanism known as epigenetics. Epigenetics was proposed by Conrad Waddington in the 1940's, and originally epigenetics focused on the study of how genes and proteins bring phenotypes into being. Nowadays, it primarily studies the mechanisms of how a cell becomes committed to particular functions and how those functional states can be inherited in cell lineages [1]. In short it can be described as heritable changes in gene expression that occur without changing a single DNA sequence in the genome. A more concise scientific definition of epigenetics is from Russel et. al: "The study of mitotically and/or meritoically heritable changes in gene function that cannot be explained by changes in DNA sequence" [2].

Epigenetics has developed and grown in interest through the years to become a hot topic in biology in cancer research, scientist have found that human tumors cells may undergo major disruptions in the pattern of DNA methylation and histone modification [3]. "The aberrant epigenetic landscape of the cancer cell is characterized by a massive genomic hypomethylation, CpG island promoter hypermethylation of tumor suppressor genes, (loss of imprinting, chromatin modification), an altered histone code for critical

genes and a global loss of monoacetylated and trimethylated histone H4 [4]" It is these collective modifications that make up the human epigenome.

Scientists studying epigenetics have made the analogy between the genome and epigenetics to the computer hardware and its software which runs the computer's operation. As mentioned by Dolinoy et al, "The two most extensively studied epigenetic mechanisms in mammals are methylation of cytosine at the carbon-5 position in CpG dinucleotides and chromatin packaging of DNA via histone variants and posttranslational histone modifications as well as subsequent nonhistone protein recruitment to specific regions of DNA. [5]" Dolinoy et al. further suggests that, "Both chromatin condensation and DNA methylation are generally associated with gene silencing. They are not necessarily independent events, but may act together to alter gene transcription. [6]" In addition, the aberrant epigenetic process can act as an alternative to DNA mutations to shut down tumor-suppressor genes and can mediate genetic alternation by inactivating DNA-repair genes (e.g. DNA hypermethylation in tumorigenesis). These superimposed epigenetic markers on the genome are areas of interest which may allow development of diagnostic, treatment, and preventive models to change the instruction of such malicious effects.

Through evolution the eukaryotic genome has continuously depleted itself from the dinucleotide CpG [7]. In the normal mammalian genome we find that the remaining CpG dinucleotides are methylated with a very high frequency, and this methylation in mammals only occurs at the 5' cytosines to guanines. It is suspected that this remaining high frequency of CpG dinucleotide may help in the arrangement of chromatin to repress the transcription in areas of repeated regions, such as transposons and Alu sequences [8].

1.2 CpG Islands

Although, targeted for depletion and present at lower than expected frequency, clusters of CpG's do exist throughout the genome and are known as CpG Islands. They are defined as regions of DNA with GC content above 0.5 UNIT and with a high observed/expected frequency of the occurrence of CpG in a region approximately 1 kb in length [9]. In the human genome, based on computation analysis, there are predicted to be around 29,000 CpG islands [10, 11], DNA Methylation and epigenetic memory and are usually found near the transcription start site in the promoter region of these genes. Previous studies have estimates of CpG island and human gene association at around 60%, of which the vast majority are unmethylated in all tissue types and at all stages of development [12]. Such finding and other literature support the belief that these regions of DNA are 'protected' from methylation. Therefore, as stated by Baylin and Herman, "this lack of methylation might be a prerequisite for active transcription [13]." They presented two classic examples in which certain alleles of the imprinted autosomal gene and multiple genes in the female inactive X-chromosome had been silenced through the full methylation of the CpG Island in their promoter region. Classic examples such as these and mounting literature have shown increasing support that the methylation of CpG Island within promoter sequence may serve as markers in the prediction of gene silencing.

1.3 DNA Methylation Silencing

Transcriptional silencing of a gene by DNA methylation of a CpG island is accomplished through the modification of chromatin that accompanies the base change, in doing so it affects structure and preventing the transcription of the gene [14].

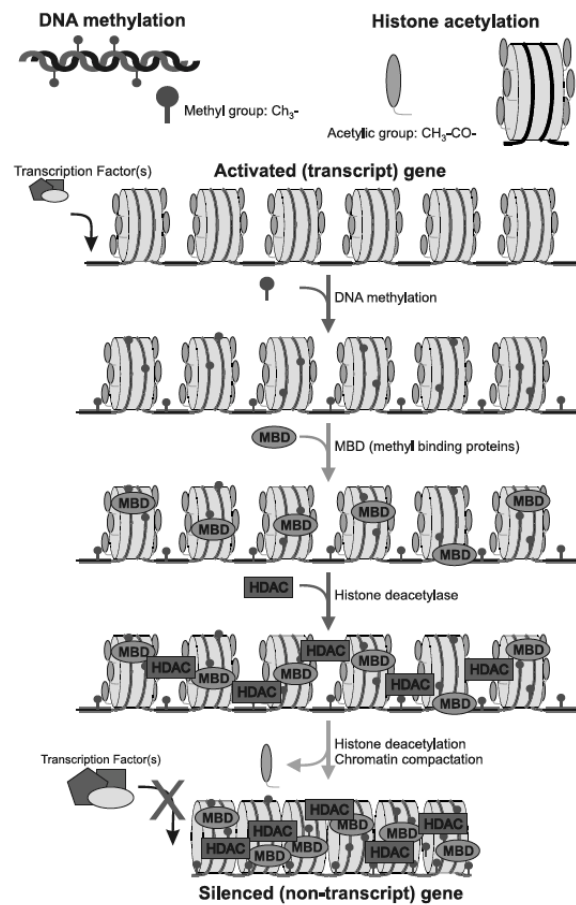


Fig. 1. “Epigenetic modifications of DNA. Euchromatic, transcriptionally-active DNA becomes silenced through DNA methylation, binding of Methyl Binding Proteins (MBD) and recruitment of Histone deacetylase (HDAC); this sequence of reactions lead to histone deacetylation and chromatin condensation with formation of genes stably silenced for the hindrance to the binding of transcription factors. DNA methylation and MBD, by themselves, can also transiently modulate transcription factors binding” Figure 1 was reproduced with PERMISSION from [15].

This transcriptional silencing of genes could then affect normal cell development, or result in abnormal cell growth. Therefore, in addition to genetic mutations, epigenetic silencing through DNA methylation could be viewed as an additional mechanism that contributes to the disruption of cell production particularly by the silencing of tumor suppressor genes.

For nearly all cases of human cancer, the silencing of an entire tumor suppressor gene has been correctly shown to require Knudson's Two Hit Hypothesis [16]. To date most attention has been centered on two pathways that promote the disabling of tumor suppressor genes. The two pathways for disabling are intragenic mutations (i.e. loss of heterozygosity) and loss of chromosomal material (homozygous deletion). Yet, literature also states transcriptional silencing may be caused through methylation of CpG Islands in the promoter regions of genes. Many researchers armed with the knowledge that DNA methylation patterns are abnormal in cancer cells, have suggested that this methylation abnormality in the promoter regions of tumor suppressor genes might be associated to human cancer [16, Figure 2].

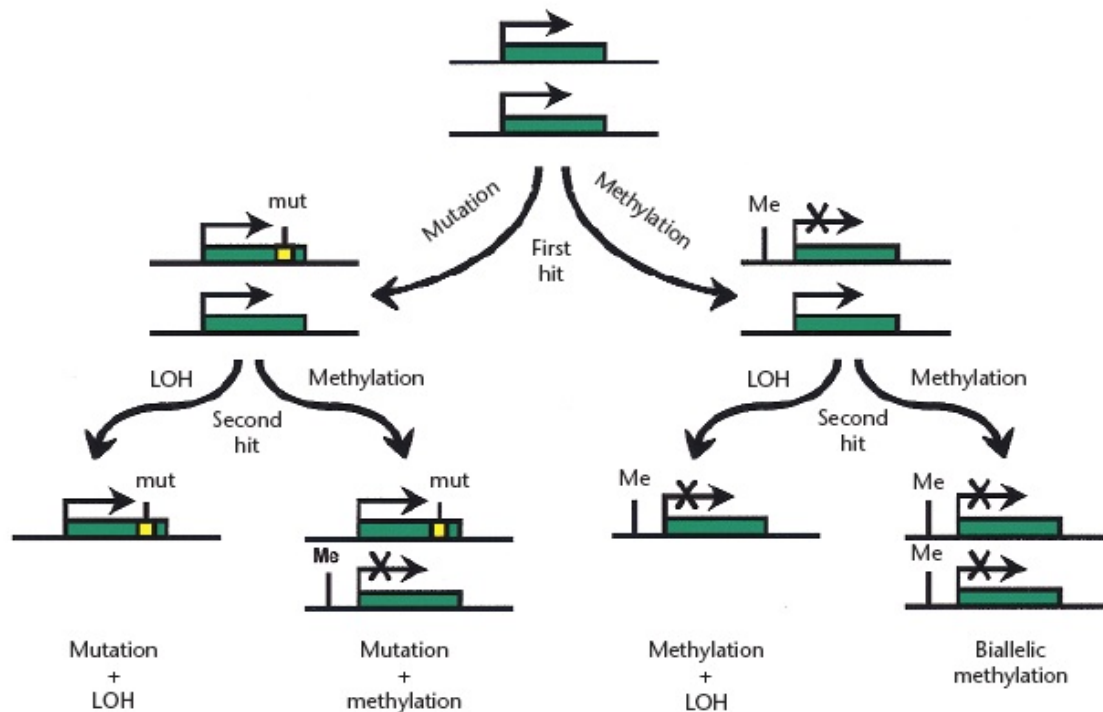


Fig. 2 “Knudson’s two-hit hypothesis revised. Two active alleles of a tumor suppressor gene are indicated by the two green boxes shown at the top. The first step of gene inactivation is shown as a localized mutation on the left or by transcriptional repression by DNA methylation on the right. The second hit is shown by either LOH or transcriptional silencing” Figure 2 was reproduced with permission from [16].

As explained in the figure above, both alleles in a tumor suppressor gene must usually be disrupted and silenced in order to cause gene function loss [17]. In Knudson's model, a tumor suppressor gene in sporadic cancer may lose its function for example, when an allele of a gene encounters a mutation (First Hit) and the other allele then experiences a deletion (Second Hit). Similarly, function loss of a tumor suppressor gene can be associated with abnormal methylation of the promoter CpG island. This occurs, when an allele of tumor suppressor gene is methylated and second allele experience a deletion. The loss of function through the mutation of both alleles is uncommon; it is more often the case that both alleles of a gene are inactivated through the association of

DNA methylation. Additionally, Knudson's model describes tumor suppressor gene inactivation in inherited cancer when an allele of a tumor suppressor gene suffers a germ line mutation, and the second becomes inactivated because of a chromosomal deletion. As described above, the second hit in inherited cancer may also be described in the methylation of promoter regions with high a concentration of CpG's. As a result there is an association which is exhibited in cancer when the loss of function is caused by the aberrant methylation of a gene promoter region.

1.4 Retrotransposons

Sequences of DNA with the ability to move within the human genome of an individual cell, are called transposons or transposable elements first discovered by Barbara McClintock, for which she was later awarded a Nobel prize in 1983. The process of movement for transposable elements between different locations is called transposition. During transposition, transposable elements can affect the genome by causing mutations and changing the number of bases in the DNA. These elements were once known as jumping genes, due to their ability to move within the genome, and are classic examples of what are known as mobile genetic elements.

There are several mobile genetic elements, and they are thus classified based on their mechanism of transposition.

- Retrotransposons are the first class (Class I) of these mobile genetic elements. Their method of transposition is accomplished by first transcribing itself to RNA; then using reverse transcriptase to reverse transcribe itself back to DNA; and finally, re-inserting itself into a different position in the genome.

- The second class of mobile genetic elements (Class II) accomplishes transposition using the enzyme transposase to "cut and paste" elements so that they will directly move from one position to another within the genome.

A large portion of a eukaryote's genome is made up transposable elements. Initially, this large volume of what was thought to be useless material perplexed scientist, and thus label "junk DNA [18]." Further studies have shown that these elements actually do play important roles, amongst them development. They are indeed now viewed as useful information to researchers, their role in DNA alterations have provided them with many clues on inter workings inside the DNA of a living organism [18].

1.5 SINES AND LINES

As mentioned previously, once viewed as "junk DNA" transposons have gained importance in recent years, specifically Class I, or retrotransposons. Within this class two members of the family known as SINEs and LINEs exist. Their names are acronyms for Short Interspersed Nuclear Elements (SINEs) and Long Interspersed Nuclear Elements (LINEs). They are particularly important because their presence can lead to genetic instability [39]. Therefore, it is critical that retrotransposons remain silent, and a key mechanism to accomplish this is DNA methylation.

In most eukaryotes these elements are the byproduct of an amplification process which depends on the reverse transcription of an RNA intermediate [19]. The amount of genome that these elements take up vary from species, from around 35% in humans [19] to 60% and greater in specific plants such as maize [19]. Such large amounts of amplification are a real threat to the host genome, since insertion of these elements into

new sites can bring about abnormal mutations [19]. In order to subdue such malicious effects, the host cells counteract transposon mobility by a combining several strategies to directly suppress one or several steps of these elements mobility process or on targeting them away from genes [19].

MATERIALS AND METHODS

2.1 Positional Weight Matrix

Positional Weight Matrices are considered the basis of motif finding algorithm and are used as an algorithm to analyze and predict DNA binding sites. This common analysis and prediction of DNA binding sites can be divided into two problems. Problem one involves developing a representation of binding sites when given a collection of sites. These representations are then used to locate new sequences and predict the location where other binding sites occur. The second problem is that given a set of known sequences containing binding sites for a common factors (e.g. retrotransposons elements), but not knowing where the sites are located, one must identify the location of the sites in each given sequence and representation for the specificity of the protein [40]. In molecular biology, A major objective is to understand sequence-specific binding of transcription factors. A Positional weight matrix may be viewed as a way to represent a motif of interest. It specifies the probability of viewing a certain base of interest at each index position of a motif.

2.2 Calculation of a Score to Predict Gene Promoter Predisposition to DNA

Methylation

The presence of SINE and LINE retrotransposons were annotated by dividing the promoter sequence of 36 methylation-resistant and 36 methylation-prone genes from into 10 bins downstream and 10 bins upstream of 1-kb sequence size for each gene's transcription start site (Figure 3A/4-A).

Figure 3: SINE and LINE abundance score to predict gene predisposition to methylation in cancer.

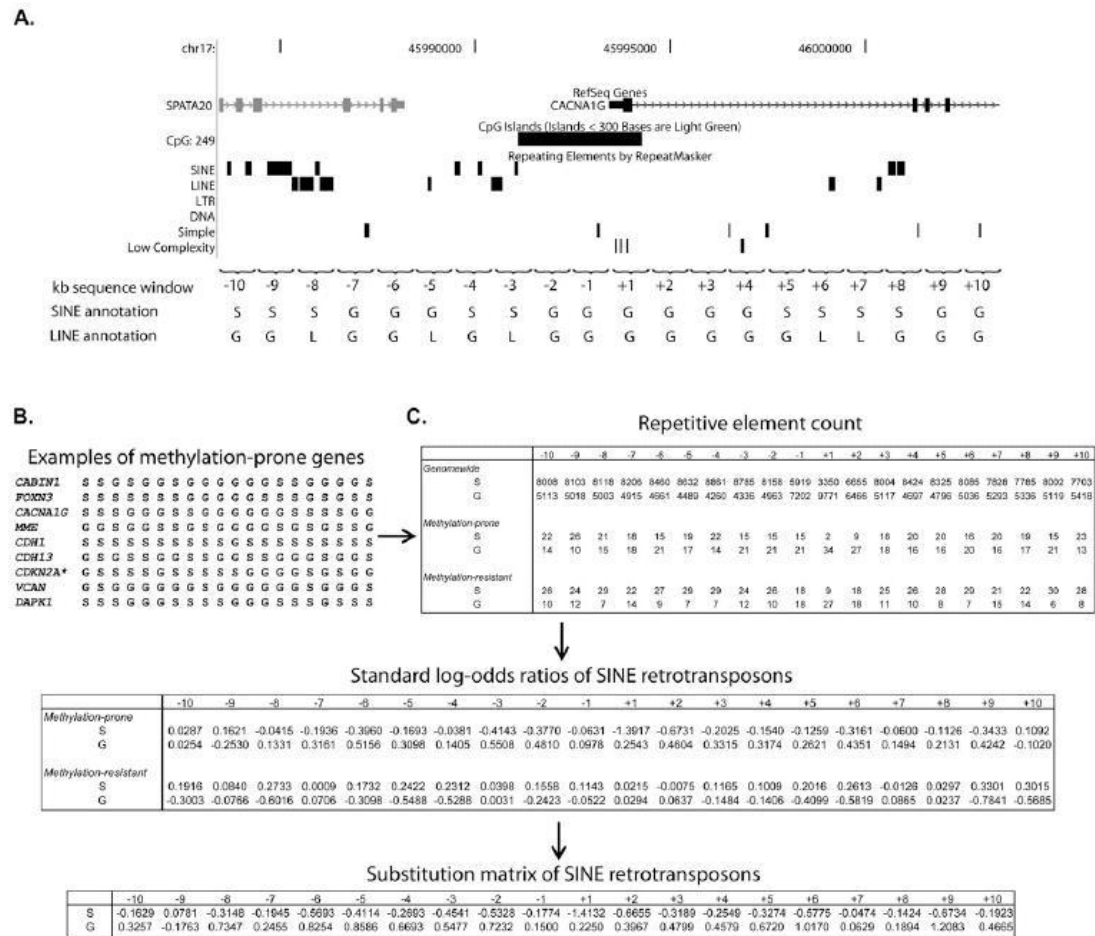


Figure 3. SINE and LINE abundance score to predict gene predisposition to methylation in cancer. (A) Annotation of SINE and LINE retrotransposons near the promoter sequence of a representative methylation-prone gene (in this example, the CACNA1G gene). The promoter sequence was divided into 20 bins of 1-kb sequence each (10 bins upstream and 10 bins downstream of each gene TSS), and the presence of SINE and LINE retrotransposons was annotated for each bin. Note that each element was annotated to just one bin (the closest to TSS). The same procedure was followed for all human genes with CpG islands overlapping or no more than 200 bp from their TSS. (B) Example of a 20-letter acronym representing SINE retrotransposon abundance in a

collection of methylation-prone genes. (C) Counting of SINE presence (S) and absence (G) in all human genes with a promoter CpG island (genome-wide) and the training set of methylation-prone and methylation-resistant genes. SINE abundance was converted to standard log-odds ratios, as described in the Methods section, and the final substitution matrix for SINE retrotransposons is presented (bottom table). The same calculation was done for LINE retrotransposons. *Transcript variant coding for the P16INK4A protein.” Figure 3 was reproduced with permission from [38].

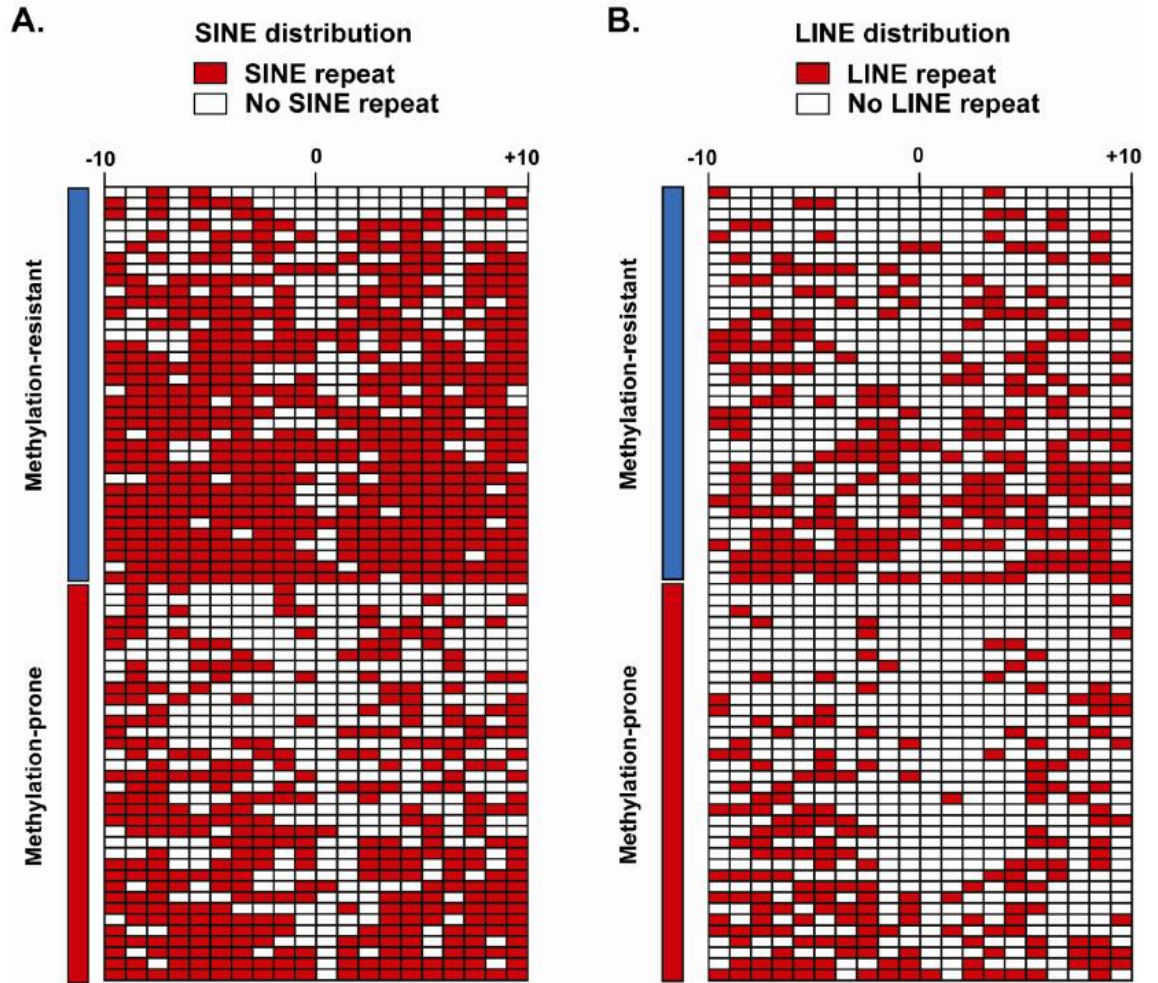


Figure 4. “Graphic representation of the presence of (A) SINE and (B) LINE retrotransposons in each 1-kb bin from transcription start site (TSS) of methylation-prone and methylation-resistant genes.” Figure 4 was reproduced with permission from [38].

“Each element was annotated to a single bin based on the start site of SINE or LINE repeat, i.e. repeats belong to the bin in which their start sites fell. In doing so, we created a 20-letter acronym representation for each gene based on the presence of SINE and LINE. Two independent acronym sequences for SINE and LINE for each gene were generated; locations free of the element were marked as G representing the absence of that element (Figure 3B/ 4-B). Using this information we compared the average abundance of LINE and SINE repeat elements per bin in methylation-resistant and methylation-prone genes to their average abundance genomewide in the entire collection of human promoter CpG islands, and interpreted the preference to retrotransposon repeats to a score that distinguished two kinds of promoters. This score

is the standard log-odd ratio which is the sum over the bin score, $s_{i,r} = \ln\left(\frac{q_{i,r}}{p_r}\right)$

where p_r is the background frequency for the repeat r , and $q_{i,r}$ is the frequency of observing the repeat of type r for the i -th bin for the promoters known to be methylated.

To account for the low count and avoid taking logarithm of zero, $q_{i,r}$ is replaced

by $Q_{i,r} = \frac{C_{i,r} + f_r}{N + 1}$ derived from ‘pseudo-count’ where f_r is the fraction of the repeat that

is type $r : \sum_r f_r = 1$

N is the total number of promoters with known methylation status; $C_{i,r}$ is the number of

repeat of type r in the i -th bin: $\sum_r c_{i,r} = N$.

The final value for each letter in the 20-letter acronym represent the abundance of SINE and LINE elements was calculated as the difference between its value in methylation-prone and methylation-resistant genes (for example, , where Smp is the SINE standard log-odd ratio in methylation-prone genes and Smr is the SINE standard log-odd ratio in methylation-resistant genes). The calculation of the log-odd ratios for SINE elements is illustrated in Figure 3-C” [38].

RESULTS AND DISCUSSION

In order to identify sequence features that are associated with the predisposition of DNA methylation in cancer, Dr. Marcos Estacio of the Department of Leukemia compared the DNA sequence promoter region in the 4-kb region surrounding the transcription start site of a training set. This set consisted of 36 methylation-resistant and 36 methylation-prone genes. The methylation analysis of the promoter region for the genes used were accomplished through quantitative methods (bisulfate-PCR followed by Cobra or pyrosequencing analysis) in nine cell lines. These genes, as well as the nine cell lines and the peripheral blood mononuclear cell DNA from a healthy individual that was used as a control, can be viewed in Table 1 and 2.

Table 1: Training Set (reproduced with permission from [38]).

Gene Symbol	Gene Name	Chrom	Tx Start	RefSeq ID
Methylation prone				
CABIN1	calcineurin binding protein 1	22	22881972	NM_012295
CACNA1G	calcium channel, voltage-dependent, T type, alpha 1G subunit	17	45993447	NM_198397
CDH1	cadherin 1, type 1, E-cadherin (epithelial)	16	67326695	NM_004360
CDH13	cadherin 13, H-cadherin (heart)	16	81218078	NM_001257
CDKN1C	cyclin-dependent kinase inhibitor 1C (p57, Kip2)	11	2863537	NM_000076
CDKN2A*	cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)	9	21965038	NM_058197
DAPK1	death-associated protein kinase 1	9	89302575	NM_004938
DPH2	DPH2 homolog (S. cerevisiae)	1	44208239	NM_001384
ESR1	estrogen receptor 1	6	152170378	NM_000125
FGFR2	fibroblast growth factor receptor 2	10	123347962	NM_022970
FHT	fragile histidine triad gene	3	61212164	NM_002012
FLT1	fms-related tyrosine kinase 1 (vascular endothelial growth factor/vascular permeability factor receptor)	13	27967265	NM_002019
FOXN3	forkhead box N3	14	88953207	NM_005197
GATA2	GATA binding protein 2	3	129694718	NM_032638
GDNF	glial cell derived neurotrophic factor	5	37871350	NM_199231
GPX1	glutathione peroxidase 1	3	49370795	NM_000581
HAND1	heart and neural crest derivatives expressed 1	5	153838017	NM_004821
HCN2	hyperpolarization activated cyclic nucleotide-gated potassium channel 2	19	540892	NM_001194
HRK	harakiri, BCL2 interacting protein (contains only BH3 domain)	12	115803615	NM_003806
MGMT	O-6-methylguanine-DNA methyltransferase	10	131155455	NM_002412
MLH3	mutL homolog 3 (E. coli)	14	74587988	NM_014381
MEPE	membrane metallo-endopeptidase	3	156280129	NM_000902
NKX2-3	NK2 transcription factor related, locus 3 (Drosophila)	10	101282679	NM_145285
PAX5	paired box 5	9	37024476	NM_016734
PDLIM4	PDZ and LIM domain 4	5	131621285	NM_003687
PPP1R1A	protein phosphatase 1, regulatory (inhibitor) subunit 1A	12	53268710	NM_006741
RB1	retinoblastoma 1	13	47775883	NM_000321
RUNX3	runt-related transcription factor 3	1	25164088	NM_004350
RYR1	ryanodine receptor 1 (skeletal)	19	43616179	NM_000540
SIN3A	SIN3 homolog A, transcription regulator (yeast)	15	73530979	NM_015477
TEAD2	TEA domain family member 2	19	54557526	NM_003598
TEC	tec protein tyrosine kinase	4	47966571	NM_003215
TP73	tumor protein p73	1	3658988	NM_005427
VCAN	versican	5	82803338	NM_004385
VRK2	vaccinia related kinase 2	2	58127223	NM_006296
ZNF160	zinc finger protein 160	19	58298499	NM_198893
Methylation-resistant				
ABL1	c-abl oncogene 1, receptor tyrosine kinase	9	132700651	NM_005157
AIP	aryl hydrocarbon receptor interacting protein	11	67007096	NM_003977
ARSA	arylsulfatase A	22	49413473	NM_000487
ARSB	arylsulfatase B	5	78318113	NM_000046
BNIP3	BCL2/adenovirus E1B 19kDa interacting protein 3	10	133645425	NM_004052
CD82	CD82 molecule	11	44543716	NM_002231
CDKN2A**	cyclin-dependent kinase inhibitor 2A (melanoma, p16, inhibits CDK4)	9	21965038	NM_058195
CLCN6	chloride channel 6	1	11788793	NM_021736
CKK	deoxycytidine kinase	4	72078255	NM_000788
FOX1	ferredoxin 1	11	109805803	NM_004109
FOXP2	forkhead box P2	7	113842287	NM_148898
GAS8	growth arrest-specific 8	16	88616508	NM_001481
GYS1	glycogen synthase 1 (muscle)	19	54188361	NM_002103
HMGGA2	high mobility group AT-hook 2	12	64504506	NM_003483
ITGAV	integrin, alpha V (vitronectin receptor, alpha polypeptide, antigen CD51)	2	187163044	NM_002210
KCNK6	potassium channel, subfamily K, member 6	19	43502323	NM_004823
MKLN1	muskelin 1, intracellular mediator containing kelch motifs	7	130663177	NM_013255
MSH2	mutS homolog 2, colon cancer, nonpolyposis type 1 (E. coli)	2	47483766	NM_000251
MSH6	mutS homolog 6 (E. coli)	2	47863789	NM_000179
NEK4	NIMA (never in mitosis gene a)-related kinase 4	3	52779991	NM_003157
NSD1	nuclear receptor binding SET domain protein 1	5	176494690	NM_172349
PDE8A	phosphodiesterase 8A	15	83326208	NM_002605
PIK3R3	phosphoinositide-3-kinase, regulatory subunit 3 (gamma)	1	46370901	NM_003629
PIP4K2B	phosphatidylinositol-5-phosphate 4-kinase, type II, beta	17	34209684	NM_003559
PXN	paxillin	12	119187892	NM_002859
RAD23A	RAD23 homolog A (S. cerevisiae)	19	12917653	NM_005053
RASA2	RAS p21 protein activator 2	3	142688615	NM_005506
SHFM1	split hand/foot malformation (ectrodactyly) type 1	7	96177139	NM_006304
STAT3	signal transducer and activator of transcription 3 (acute-phase response factor)	17	37794039	NM_003150
TAF11	TAF11 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 28kDa	6	34963797	NM_005643
TDG	thymine-DNA glycosylase	12	102883746	NM_003211
TGFBR2	transforming growth factor, beta receptor II (70/80kDa)	3	30622997	NM_003242
TOP1	topoisomerase (DNA) I	20	39090875	NM_003286
TOP2A	topoisomerase (DNA) II alpha 170kDa	17	35827695	NM_001067
TRIM27	tripartite motif-containing 27	6	28999747	NM_006510
VHL	von Hippel-Lindau tumor suppressor	3	10158318	NM_000551

Table 2: Methylation profile of the training set in nine cancer cell lines from different tissue origin (reproduced with permission from [38]).

Gene Symbol	NSCLC NCI-H322	PROSTATE PC3	BREAST MB 435	RENAL TK-10	CNS SNB-19	OVARIAN OVCAR-5	MELANOMA LOX IMUI	LEUK MOLT-4	COLON HCT 116
<i>Methylation-prone</i>									
CABIN1	100	0	0	0	0	0	0	0	0
CACNA1G	38	32	62	80	57	24	58	7	58
CDH1	0	0	100	63	68	0	100	100	0
CDH13	6	20	88	75	3	79	4	ND	97
CDKN1C	0	0	36	54	0	0	27	51	0
CDKN2A*	Deleted	89	0	81	Deleted	0	0	Deleted	57
DAPK1	59	15	74	0	8	0	64	54	10
DPH2	0	0	0	44	32	56	67	45	34
ESR1	0	43	27	44	29	66	55	66	57
FLT1	0	20	0	0	0	60	0	0	31
FOXP3	0	0	0	0	0	0	74	0	0
GATA2	0	0	68	0	0	0	0	0	53
GDNF	79	3	75	50	39	93	44	77	93
GPX1	54	0	0	0	0	0	0	55	0
HAND1	24	15	69	69	52	78	33	47	47
HRK	14	45	13	29	4	54	96	4	7
MGMT	58	3	93	64	65	0	91	0	79
MLH3	6	2	12	2	54	2	2	77	9
MME	0	0	49	44	0	11	2	81	76
PAX5	54	2	87	84	37	54	85	3	64
PDLIM4	92	0	53	0	0	0	92	76	56
PPP1R1A	0	18	27	0	0	0	20	0	28
RYR1	0	0	35	0	0	32	0	0	0
TEAD2	20	0	0	0	0	0	0	48	0
TEC	4	0	42	0	0	0	0	20	0
TP73	0	28	100	0	0	0	100	30	0
VRK2	0	0	0	0	90	5	0	15	0
ZNF160	ND	0	6	0	7	90	74	5	0
<i>Methylation-resistant</i>									
ABL1	0	0	0	0	0	0	0	0	0
AIP	0	0	0	0	0	0	0	0	0
ARSA	ND	3	5	ND	5	12	5	ND	10
ARSB	8	12	12	7	7	8	10	6	7
BNIP3	0	0	0	0	0	27	0	0	0
CD82	2	4	17	1	2	1	2	1	7
CDKN2A**	0	0	0	13	Deleted	0	0	0	15
CLCN6	0	0	0	0	0	0	0	0	0
DCK	4	3	3	12	2	4	2	4	11
FDX1	0	0	0	0	0	0	0	0	0
FOXP2	0	0	14	14	0	0	0	18	0
GAS8	50	0	0	0	0	0	0	0	0
GYS1	0	0	0	0	0	0	0	0	0
HMG42	0	0	0	0	0	0	0	0	0
ITGAV	0	0	0	0	0	0	0	0	0
KCNK6	0	0	0	2	12	0	2	0	0
MKLN1	14	0	0	0	0	0	0	0	0
MSH2	2	5	3	2	2	2	2	2	11
MSH6	4	6	3	2	3	2	3	2	13
NEK4	0	0	0	0	0	0	0	0	0
NSD1	0	0	0	10	0	0	0	0	0
PDE8A	0	0	0	0	0	0	0	0	0
PIK3R3	0	0	0	0	0	0	0	0	0
PIP4K2B	0	0	0	0	0	0	0	0	0
PXN	0	0	0	0	0	0	0	0	0
RAD23A	0	0	0	0	0	0	0	0	0
RAS42	10	12	7	4	3	9	2	2	14
SHFM1	0	0	0	0	0	0	0	0	0
STAT3	0	0	0	0	0	0	0	0	0
TAF11	2	3	2	2	2	2	1	2	7
TDG	0	0	0	0	0	0	0	0	0
TGFBR2	2	2	2	2	2	2	2	3	9
TOP1	0	0	0	0	0	0	0	0	0
TOP2A	1	2	2	2	1	1	1	2	6
TRIM27	6	0	14	5	5	0	0	0	0
VHL	6	8	8	5	4	4	5	8	18

These particular 9 cancer lines were used because they have been identified as heavily methylated in previous studies [20]. Therefore, non-methylated genes found in these cell lines were unlikely to be found methylated elsewhere. Due to epigenetic modifications targeting retrotransposons to suppress their mobilization [22], we asked the question of whether methylation-prone genes have a different distribution of such elements when compared to genes that were classified as methylation resistant. As illustrated in Figure 5A, SINE and LINE repeats were approximately half as common in methylation-prone as in methylation resistant genes. Annotations of other repeats were analyzed but their distributions between methylation-resistant and methylation-prone genes were not significantly different. Other features such as GC content, CpG Island length, and CpG ratio have previously been shown to be associated with methylation status in somatic tissues (Weber et. al, 2007), yet in our observation these variables were not significantly different between methylation-prone and methylation-resistant genes in cancer. (Fig 5B)

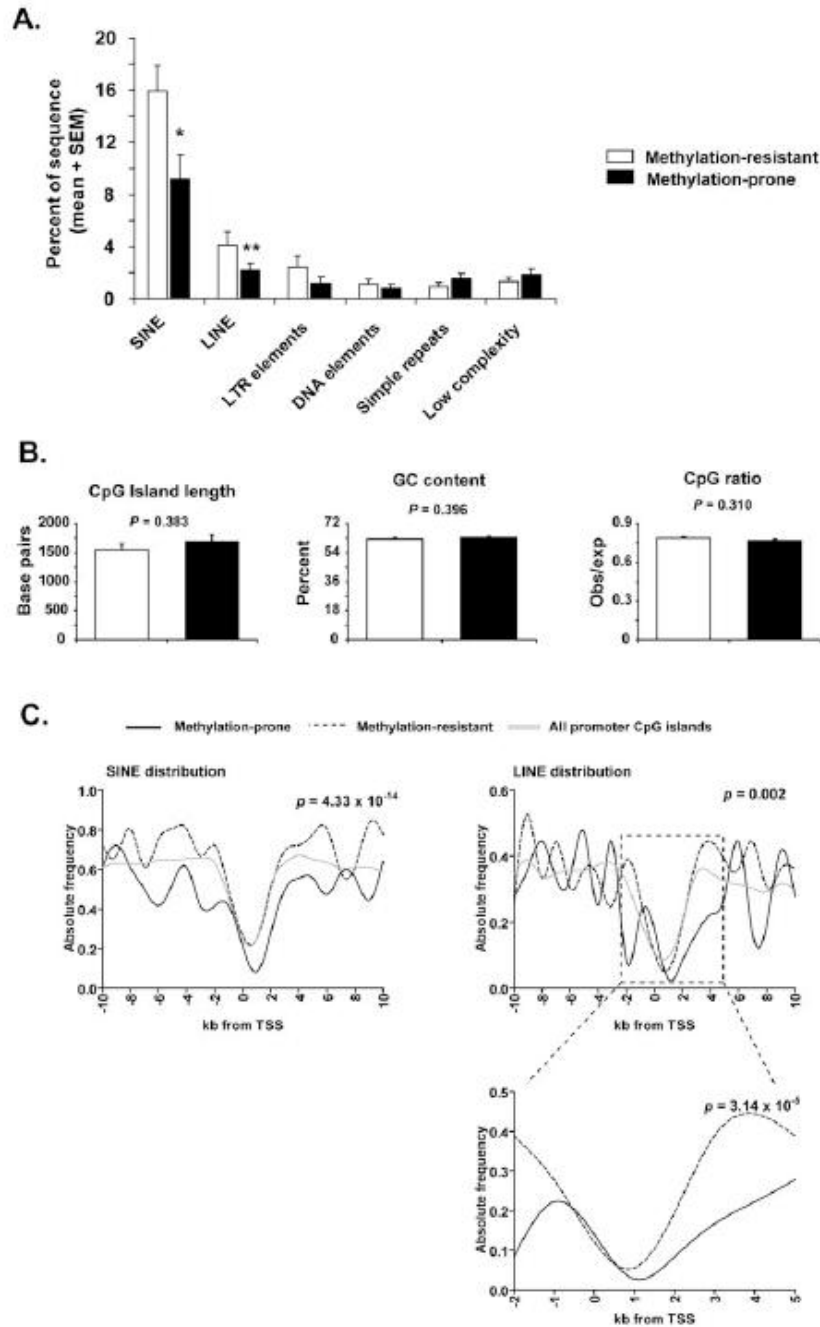


Figure 5. “Distribution of repetitive elements in methylation-prone versus methylation-resistant genes. (A) The abundance of repetitive elements of different classes was determined for the 4-kb sequence window centered in the TSS of 36 methylation-resistant (white) and 36 methylation-prone (black) genes. Retrotransposons of the SINE and LINE classes were found to be depleted in methylation-prone genes. * $P < 0.02$; ** $P < 0.01$ (Student’s t-test). (B) Average length, GC content, and CpG ratio of CpG islands were not significantly different between methylation-prone and methylation-resistant genes. Error bars represent SEM. (C) Abundance of SINE and LINE retrotransposons in the 20-kb sequence window centered in the TSS of 36 methylation-prone and 36 methylation-resistant genes. The abundance of SINE and LINE retrotransposons

in all promoter CpG islands in the human genome is shown in gray. Note that the depletion of LINE retrotransposons is more significant in the -2-kb to +5-kb sequence window.” Figure 5 was reproduced with permission from [38]

To investigate the effects of window size on the distribution of SINE and LINE retrotransposons repeats between methylation-resistant and methylation-prone genes, we increased our 4-kb window to a 20-kb region centered on the TSS and annotated for every 1-kb non-overlapping window. This analysis identified a near depletion of SINE repeats spanning the 20-kb region, the same did not hold for LINE repeats which only showed a depletion occurring mainly in the -2kb to 5 kb window (Fig 5C). For every 1-kb window, the log-odds score was calculated for SINE and LINE based on the full list of human promoter CpG Island and their distribution in the training set. Taking the sum of these scores in the 20-kb window enable us to measure the similarity in the distribution of LINE and SINE retrotransposons in an individual gene promoter in contrast to the average distribution of LINE and SINE in methylation-resistant and methylation-prone genes (Fig. 4). Using this information, we can identify three groups in the training set: (i) genes predicted as methylation-prone, i.e. genes depleted of both SINE and LINE, (ii) genes predicted to be methylation-resistant, i.e. genes enriched both SINE and LINE, and (iii) genes that showed enrichment of either SINE or LINE but not both. Comparing our results with the methylation data for each individual of the 72 observed genes showed that 19 of 23 (83%) genes predicted to be prone to methylation were indeed hypermethylated in cancer, 23 and 25 (92%) genes predicted to be methylation resistant were actually never or rarely hypermethylated in cancer (Fig. 6A). The 24 of 72 (34%) genes showing a depletion of only one type of repeat represented a class of genes of intermediate predisposition to methylation.

In order to validate our predictive method, we examined 68 methylation-resistant and 74 methylation prone genes with available promoter methylation data in cancer (Table 3 and 4) from a collection of tissues (Leukemia, colon, breast, and lung among others).

Table 3: Methylation-prone genes in cancer used as test set (reproduced with permission from [38]).

Gene Symbol	Gene Name	RefSeq	Chrom	Tx Start	S score	L score	Prediction	Source
ANKHD1-EIF4EBP3	ANKHD1-EIF4EBP3 readthrough	NM_020690	5	139761612	-5.322	2.160	intermediate	MCAM
BARHL1	Barhl-like homeobox 1	NM_020064	9	134447813	3.776	0.379	prone	Bis-PCR
BIN3	bridging integrator 3	NM_018688	8	22582606	1.374	-1.956	intermediate	MCAM
BMP7	bone morphogenetic protein 7	NM_001719	20	55274708	4.903	0.701	prone	18
C10orf70	chromosome 19 open reading frame 70	NM_205767	19	5631911	-3.435	0.235	intermediate	MCAM
CALCA	calcitonin-related polypeptide alpha	NM_001741	11	14950408	2.834	0.680	prone	2
CEBPD	CCAAT/enhancer binding protein (C/EBP), delta	NM_005195	8	48813279	1.265	2.055	prone	23
CHFR	checkpoint with forkhead and ring finger domains	NM_018223	12	131974257	-2.538	-0.969	resistant	5
CLDN3	claudin 3	NM_001306	7	72822512	-5.427	-1.544	resistant	13
CNNM1	cyclin M1	NM_020348	10	101080022	-1.270	-2.175	resistant	10
COL2A1	collagen, type II, alpha 1	NM_001844	12	46684552	6.333	1.245	prone	10
CPT1C	carnitine palmitoyltransferase 1C	NM_152359	19	54886218	-1.354	0.124	intermediate	10
DAB2	disabled homolog 2, mitogen-responsive phosphoprotein (Drosophila)	NM_001343	5	39461092	2.058	0.533	prone	4
DDIT4L	DNA-damage-inducible transcript 4-like	NM_145244	4	101330636	4.524	-1.492	intermediate	10
DERL3	Der1-like domain family, member 3	NM_198440	22	22511201	2.592	0.052	prone	10
DHR33	dehydrogenase/reductase (SDR family) member 3	NM_004753	1	12600407	0.700	1.099	prone	10
DNK1	dickkopf homolog 1 (Xenopus laevis)	NM_012242	10	53744046	5.597	1.351	prone	1
DLC1	deleted in liver cancer 1	NM_024767	8	13416766	1.425	0.132	prone	28
DLEC1	deleted in lung and esophageal cancer 1	NM_007337	3	38055699	-1.228	-2.988	resistant	20
DPYS	dihydropyrimidinase	NM_001385	8	105548453	4.745	1.418	prone	10
DSC3	desmocollin 3	NM_024423	18	26876779	3.530	-0.565	intermediate	19
EDIL3	EGF-like repeats and discoidin I-like domains 3	NM_005711	5	83716367	4.375	2.502	prone	MCAM
EDNRB	endothelin receptor type B	NM_000115	13	77447665	6.416	-1.733	intermediate	6
EFEMP2	EGF-containing fibulin-like extracellular matrix protein 2	NM_016938	11	65396852	2.029	0.292	prone	10
ELK2	empty spiracles homeobox 2	NM_004098	10	119291945	7.077	1.310	prone	MCAM
EPHX3	epoxide hydrolase 3	NM_024794	19	15204231	-0.863	1.271	intermediate	10
ERBB4	v-erb-a erythroblastic leukemia viral oncogene homolog 4 (avian)	NM_005235	2	213111597	7.454	1.702	prone	MCAM
ESYT3	extended synaptotagmin-like protein 3	NM_031913	3	139636308	2.109	-1.271	intermediate	10
FAM1042	family with sequence similarity 19 (chemokine [C-C motif]-like), member A2	NM_178539	12	60872818	8.355	1.006	prone	MCAM
GDNF	glial cell derived neurotrophic factor	NM_000514	5	37875539	2.242	2.919	prone	MCAM
GSTP1	glutathione S-transferase pi 1	NM_000852	11	67107861	-1.541	0.471	intermediate	29
HNF1B	HNF1 homeobox B	NM_000458	17	33179209	-0.991	-0.152	resistant	24
HPSE2	heparanase 2	NM_021828	10	100985609	5.000	2.973	prone	MCAM
KCNAB1	potassium voltage-gated channel, shaker-related subfamily, beta member 1	NM_172159	3	157491519	3.711	-2.544	intermediate	MCAM
LHX3	UM homeobox 6	NM_199160	9	124030840	-0.761	1.858	intermediate	Bis-PCR
LHX9	UM homeobox 9	NM_020204	1	196153139	8.240	1.702	prone	MCAM
LHX1A	UM homeobox transcription factor 1, alpha	NM_177399	1	163591641	2.291	0.904	prone	MCAM
LRP2	low density lipoprotein receptor-related protein 2	NM_004525	2	169927368	2.017	1.707	prone	Bis-PCR
MLH1	mutL homolog 1, colon cancer, nonpolyposis type 2 (E. coli)	NM_000249	3	37009982	2.962	-1.882	intermediate	Bis-PCR
MLPH	melanophilin	NM_024101	2	238060616	1.683	0.264	prone	MCAM
MT1G	metallothionein 1G	NM_005950	16	55259478	-0.934	1.702	intermediate	11
MT3	metallothionein 3	NM_005954	16	55180767	-3.033	-2.058	resistant	21
MYO10	myogenic differentiation 1	NM_002478	11	17697685	6.576	-1.652	intermediate	12
NFIX	nuclear factor I/X (CCAAT-binding transcription factor)	NM_002501	19	12967583	1.490	1.702	prone	MCAM
ONECUT2	one cut homeobox 2	NM_004852	18	53253914	9.237	1.702	prone	MCAM
OTP	orthopedia homeobox	NM_032109	5	76970278	9.237	2.299	prone	MCAM
PRDX2	peroxiredoxin 2	NM_181738	19	12773694	-3.191	1.863	intermediate	10
PRRG1	proline rich Gla (G-carboxyglutamic acid) 1	NM_000950	X	37093545	4.674	-0.296	intermediate	MCAM
PTGS2	prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)	NM_000963	1	184916179	2.147	0.992	prone	26
PTPRG	protein tyrosine phosphatase, receptor type, G	NM_002841	3	61522284	-1.424	-0.229	resistant	10
PVRL2	poliovirus receptor-related 2 (herpesvirus entry mediator B)	NM_002856	19	50041232	-4.724	-1.748	resistant	MCAM
RASD1	RAS, dexamethasone-induced 1	NM_016084	17	17340432	3.615	2.434	prone	10
RASGRP2	Ras protein-specific guanine nucleotide-releasing factor 2	NM_006909	5	80292313	4.645	-0.079	intermediate	4
RASSF2	Ras association (RalGDS/AF-6) domain family member 2	NM_014737	20	4752291	1.156	-0.063	intermediate	14
RECSL1	RECB homolog (yeast)	NM_005132	14	23711073	1.261	1.208	prone	10
RECK	reversion-inducing-cysteine-rich protein with kazal motifs	NM_021111	9	36026915	1.046	-5.119	intermediate	7
RPS5	ribosomal protein S5	NM_001009	19	63590447	-1.227	-0.579	resistant	MCAM
SFRP1	secreted frizzled-related protein 1	NM_003012	8	41286137	5.634	2.585	prone	22
SMAD4	SMAD family member 4	NM_005359	18	46810610	0.255	1.258	prone	25
SMARCA2	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 2	NM_003070	9	2005341	5.634	1.702	prone	MCAM
SOC1	suppressor of cytokine signaling 1	NM_003745	16	11257540	-3.038	-1.196	resistant	16
STAT1	signal transducer and activator of transcription 1, 91kDa	NM_139266	2	191587181	4.775	-1.399	intermediate	27
TEX5	T-box 5	NM_080717	12	113330630	5.412	1.841	prone	MCAM
TERT	telomerase reverse transcriptase	NM_198253	5	1348162	2.149	-0.394	intermediate	9
TEX9	testis expressed 9	NM_198524	15	54444935	4.942	-5.345	intermediate	MCAM
THBS1	thrombospondin 1	NM_003246	15	37660571	5.911	1.763	prone	Bis-PCR
UCHL1	ubiquitin carboxyl-terminal esterase L1 (ubiquitin thiolesterase)	NM_004181	4	40953685	-1.778	-1.969	resistant	17
UNC5C	unc-5 homolog C (C. elegans)	NM_003728	4	96689185	5.060	1.889	prone	10
VAX2	ventral anterior homeobox 2	NM_012476	2	70981227	5.211	-0.533	intermediate	MCAM
WIF1	WNT inhibitory factor 1	NM_007191	12	63801383	6.206	1.060	prone	3
WNT10A	wingless-type MMTV integration site family, member 10A	NM_025216	2	219453498	2.172	0.974	prone	18
WNT9	wingless-type MMTV integration site family, member 6	NM_006522	2	219432789	3.192	1.409	prone	18
ZGPAT	zinc finger, CCH-type with G patch domain	NM_032527	20	61809834	0.756	-3.336	intermediate	18
ZHX2	zinc fingers and homeoboxes 2	NM_014943	8	123863081	2.024	0.310	prone	15

Table 4: Methylation-resistant genes in cancer used as test set (reproduced with permission from [38]).

Gene Symbol	Gene Name	RefSeq	Chrom	Tx Start	S score	L score	Prediction	Source
ABCB10	ATP-binding cassette, sub-family B (MDR/TAP), member 10	NM_012089	1	227761065	-1.740	-0.251	resistant	Bis-PCR
ADAT1	adenosine deaminase, tRNA-specific 1	NM_012091	16	74214655	-0.550	-1.042	resistant	MCAM
ADRB1	adrenergic, beta-1-, receptor	NM_000684	10	115793795	1.304	1.930	prone	Bis-PCR
ALDH7A1	aldehyde dehydrogenase 7 family, member A1	NM_001182	5	125958839	-4.027	0.168	intermediate	MCAM
ALG12	asparagine-linked glycosylation 12, alpha-1,6-mannosyltransferase homolog (S. cerevisiae)	NM_024105	22	48698110	0.421	-1.304	intermediate	Bis-PCR
ATP6V1E1	ATPase, H+ transporting, lysosomal 31kDa, V1 subunit E1	NM_001696	22	16491588	-5.621	-0.730	resistant	Bis-PCR
ATP6V1G1	ATPase, H+ transporting, lysosomal 13kDa, V1 subunit G1	NM_004888	9	116389814	-5.650	-1.336	resistant	MCAM
ATXN10	ataxin 10	NM_013236	22	44446350	-3.514	-6.224	resistant	Bis-PCR
BAG2	BCL2-associated athanogene 2	NM_004282	6	57145292	3.993	-0.241	intermediate	MCAM
BCAT2	branched chain amino-acid transaminase 2, mitochondrial	NM_001190	19	54006113	-1.196	-0.463	resistant	MCAM
BCL2	B-cell CLL/lymphoma 2	NM_000657	18	59137593	4.734	1.163	prone	Bis-PCR
BID	BH3 interacting domain death agonist	NM_197966	22	16637258	0.358	0.980	prone	Bis-PCR
C12orf10	chromosome 12 open reading frame 10	NM_021640	12	51979736	0.841	0.584	prone	MCAM
C14orf21	chromosome 14 open reading frame 21	NM_174913	14	23838937	4.985	-1.985	intermediate	MCAM
COL9A1	collagen, type IX, alpha 1	NM_001851	6	71069494	0.597	-0.290	intermediate	MCAM
CYB5A3C3	cytochrome b, ascorbate dependent 3	NM_153611	11	60886305	-2.257	-0.154	resistant	MCAM
DHRS31	dehydrogenase/reductase (SDR family) member 1	NM_138452	14	23838506	2.535	-2.490	intermediate	MCAM
DUS3L	dihydrouridine synthase 3-like (S. cerevisiae)	NM_020175	19	5742190	-1.711	-0.880	resistant	MCAM
FAM53C	family with sequence similarity 53, member C	NM_016605	5	137701866	-1.716	-0.313	resistant	Bis-PCR
FIBP	fibroblast growth factor (acidic) intracellular binding protein	NM_198897	11	65412586	-3.148	-1.851	resistant	MCAM
FNBP1	formin binding protein 1	NM_015033	9	131845294	-3.395	0.517	intermediate	MCAM
FRAT1	frequently rearranged in advanced T-cell lymphomas	NM_005479	10	99069011	-3.925	2.607	intermediate	MCAM
GABPB1	GA binding protein transcription factor, beta subunit 1	NM_181427	15	48434687	-1.762	-1.130	resistant	Bis-PCR
GNPAT	glyceronephosphate O-acyltransferase	NM_014236	1	229434604	-1.568	-3.546	resistant	MCAM
GPM6B	glycoprotein M6B	NM_005278	X	13866752	4.910	1.898	prone	MCAM
GTSE1	G-2 and S-phase expressed 1	NM_016426	22	45071475	-2.325	-5.270	resistant	Bis-PCR
HNRPA0	heterogeneous nuclear ribonucleoprotein A0	NM_006805	5	137117938	3.743	0.267	prone	Bis-PCR
HTT	huntingtin	NM_002111	4	3046205	-1.297	0.502	intermediate	MCAM
KIAA0404	KIAA0404	NM_014774	1	46957323	-2.563	-2.687	resistant	MCAM
KIF20A	kinesin family member 20A	NM_005733	5	137543247	-0.370	0.670	intermediate	Bis-PCR
KPNA3	karyopherin alpha 3 (importin alpha 4)	NM_002267	13	49265058	-2.690	-0.376	resistant	Bis-PCR
MCEE	methylmalonyl CoA epimerase	NM_032601	2	71210867	-2.576	-5.401	resistant	MCAM
MFAP1	microfibrillar-associated protein 1	NM_005926	15	41904243	-6.891	-1.775	resistant	MCAM
MOBP	myelin-associated oligodendrocyte basic protein	NM_182935	3	39484073	1.464	-2.316	intermediate	MCAM
MPHOSPH10	M-phase phosphoprotein 10 (U3 small nucleolar ribonucleoprotein)	NM_005791	2	71210951	-2.939	-0.845	resistant	MCAM
MTX2	metaxin 2	NM_005554	2	176842382	1.586	-0.246	intermediate	MCAM
NUDT15	nudix (nucleoside diphosphate linked moiety X)-type motif 15	NM_018283	13	47509703	-0.803	-3.181	resistant	Bis-PCR
PARK2	Parkinson disease (autosomal recessive, juvenile) 2, parkin	NM_013988	6	163068790	7.919	1.222	prone	MCAM
PEX3	peroxisomal biogenesis factor 3	NM_009630	6	143813809	-1.002	0.629	intermediate	Bis-PCR
PGD	phosphogluconate dehydrogenase	NM_002631	1	10381671	-5.650	0.555	intermediate	Bis-PCR
POLR2E	polymerase (RNA) II (DNA directed) polypeptide E, 25kDa	NM_002695	19	1046336	-3.333	0.002	intermediate	MCAM
PPAN	peter pan homolog (Drosophila)	NM_020230	19	10077964	-2.877	0.866	intermediate	Bis-PCR
PSMB7	proteasome (prosome, macropain) subunit, beta type, 7	NM_002799	9	126217542	-4.979	-3.440	resistant	MCAM
RBM17	RNA binding motif protein 17	NM_032905	10	6171012	-0.992	-0.437	resistant	Bis-PCR
RCBTB1	regulator of chromosome condensation (RCC1) and BTB (POZ) domain containing protein 1	NM_018191	13	49057720	-0.533	-2.394	resistant	Bis-PCR
SART3	squamous cell carcinoma antigen recognized by T cells 3	NM_014706	12	107479295	0.367	-1.122	intermediate	MCAM
SCFD2	sec1 family domain containing 2	NM_152540	4	53926999	0.882	-3.513	intermediate	MCAM
SELO	selenoprotein O	NM_031454	22	48981534	3.141	-1.841	intermediate	Bis-PCR
SETDB2	SET domain, bifurcated 2	NM_031915	13	48916510	0.037	-2.806	intermediate	Bis-PCR
SIL1	SIL1 homolog, endoplasmic reticulum chaperone (S. cerevisiae)	NM_022464	5	138561964	-2.659	0.866	intermediate	Bis-PCR
SMURF2	SMAD specific E3 ubiquitin protein ligase 2	NM_022739	17	60088848	-2.694	-1.538	resistant	MCAM
SOX7	SRY (sex determining region Y)-box 7	NM_031439	8	10625432	1.781	1.267	prone	MCAM
SPTLC2	serine palmitoyltransferase, long chain base subunit 2	NM_004863	14	77152863	-2.979	-2.403	resistant	MCAM
ST8SIA1	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 1	NM_003034	12	22378915	2.606	-0.710	intermediate	MCAM
STMN1	stathmin 1	NM_203401	1	26105231	-2.580	-0.032	resistant	Bis-PCR
SUCLA2	succinate-CoA ligase, ADP-forming, beta subunit	NM_003850	13	47473463	-2.072	-3.282	resistant	Bis-PCR
TCF21	transcription factor 21	NM_198392	6	134251969	3.646	0.787	prone	MCAM
TMEM33	transmembrane protein 33	NM_018126	4	41631925	1.263	-2.671	intermediate	MCAM
TMEM45A	transmembrane protein 45A	NM_018004	3	101694152	0.954	-1.974	intermediate	MCAM
TRIB2	tribbles homolog 2 (Drosophila)	NM_021643	2	12774658	-1.211	-0.279	resistant	MCAM
TRIP4	thyroid hormone receptor interactor 4	NM_016213	15	62467072	-6.451	0.081	intermediate	MCAM
UBQLN1	ubiquilin 1	NM_013438	9	85512773	3.277	-3.856	intermediate	Bis-PCR
USP18	ubiquitin specific peptidase 18	NM_017414	22	17012757	-1.142	-0.969	resistant	Bis-PCR
ZAK	leucine zipper- and sterile alpha motif-containing kinase	NM_016653	2	173648810	-0.846	-2.494	resistant	MCAM
ZBED4	zinc finger, BED-type containing 4	NM_014838	22	48633500	-0.388	-1.704	resistant	Bis-PCR
ZNF224	zinc finger protein 224	NM_013398	19	49290336	-2.214	-5.826	resistant	MCAM
ZNF418	zinc finger protein 418	NM_133460	19	63138552	-1.253	0.639	intermediate	MCAM
ZNF780	zinc finger protein 786	NM_152411	7	148418720	-4.792	-1.125	resistant	Bis-PCR

As Fig. 6A displays, 92 % of genes predicted to be methylation resistant were not methylated, and 83% of genes predicted to be methylation prone were in fact methylated in cancer. Our observation also showed a 1:1 ratio of unmethylated to methylated genes in our predicted methylation-intermediate group. A natural extension, to validate our predictive method, was to test whether our model would hold in a large-scale analysis. For this, we compared our three classes identified by MCAM analysis, in 32 primary tissue and 28 cancer lines, consisting of more than 26000 probes representing around 6600 CpG island associated gene promoters. MCAM analysis is a sensitive and specific micro method which is based on a selective amplification of methylated DNA after restriction enzyme digestion [21, 23 (explain in more detail MCAM analysis)]. From our MCAM analysis we observed that genes predicted as resistant to methylation had the lowest values of measured promoter methylation. Genes predicted to be methylation prone, however, displayed the highest average values of methylation (Fig. 6B and 6C). This pattern was shown in 59 of 60 (98%) genes of the studied samples (Figure 7).

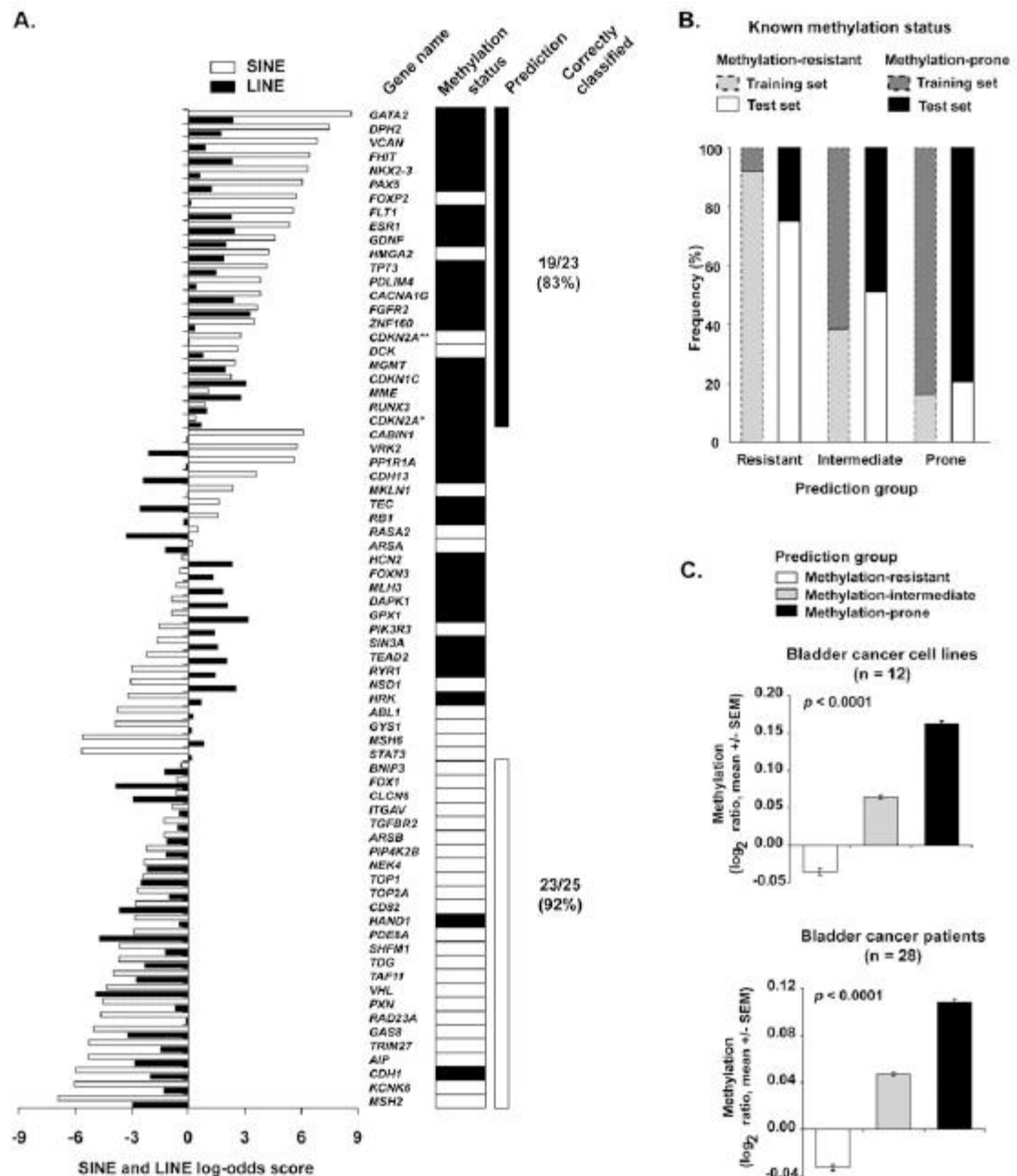


Figure 6. “Prediction of gene predisposition and resistance to hypermethylation in cancer. (A) SINE and LINE scores of the training set genes. The scores were calculated according to the described log-odds ratio method for each gene and are represented as horizontal bars (white bars, SINE score; black bars, LINE score). Methylation status determined by bisulfite PCR methods is shown on the right. Genes with concordant depletion of SINE and LINE retrotransposons (log-odds) were predominantly methylation-prone, with the opposite found for genes with enrichment of both SINE and

LINE repeats. Discordant SINE and LINE scores likely represent a class of genes with intermediate predisposition to methylation. *Transcript variant coding for the P16INK4A protein. **Transcript variant coding for the P14ARF protein. Black rectangles represent methylated genes; white rectangles represent unmethylated genes. (B) The predictive method based on SINE and LINE retrotransposons abundance was applied to a test set composed of 142 genes. The frequency of genes correctly classified according to their DNA methylation status in cancer was 79% for methylation-resistant and 75% for methylation-prone genes. These values were closely related to those found in the training set (gray bars). (C) Validation of the predictive method in a large set of cancer cell lines and primary cancer tissues. Methylation status of more than 6600 autosomal gene promoters was determined by MCAM. X chromosome genes were excluded from this analysis due to their hemimethylated status in female samples. The measured DNA methylation per tissue type was significantly higher in predicted methylation-prone genes than in predicted methylation-resistant and methylation-intermediate genes. Methylation is presented as the log₂ ratio (cancer/control) of all oligonucleotide probes of a predicted methylation status." Figure 6 was reproduced with permission from [38]

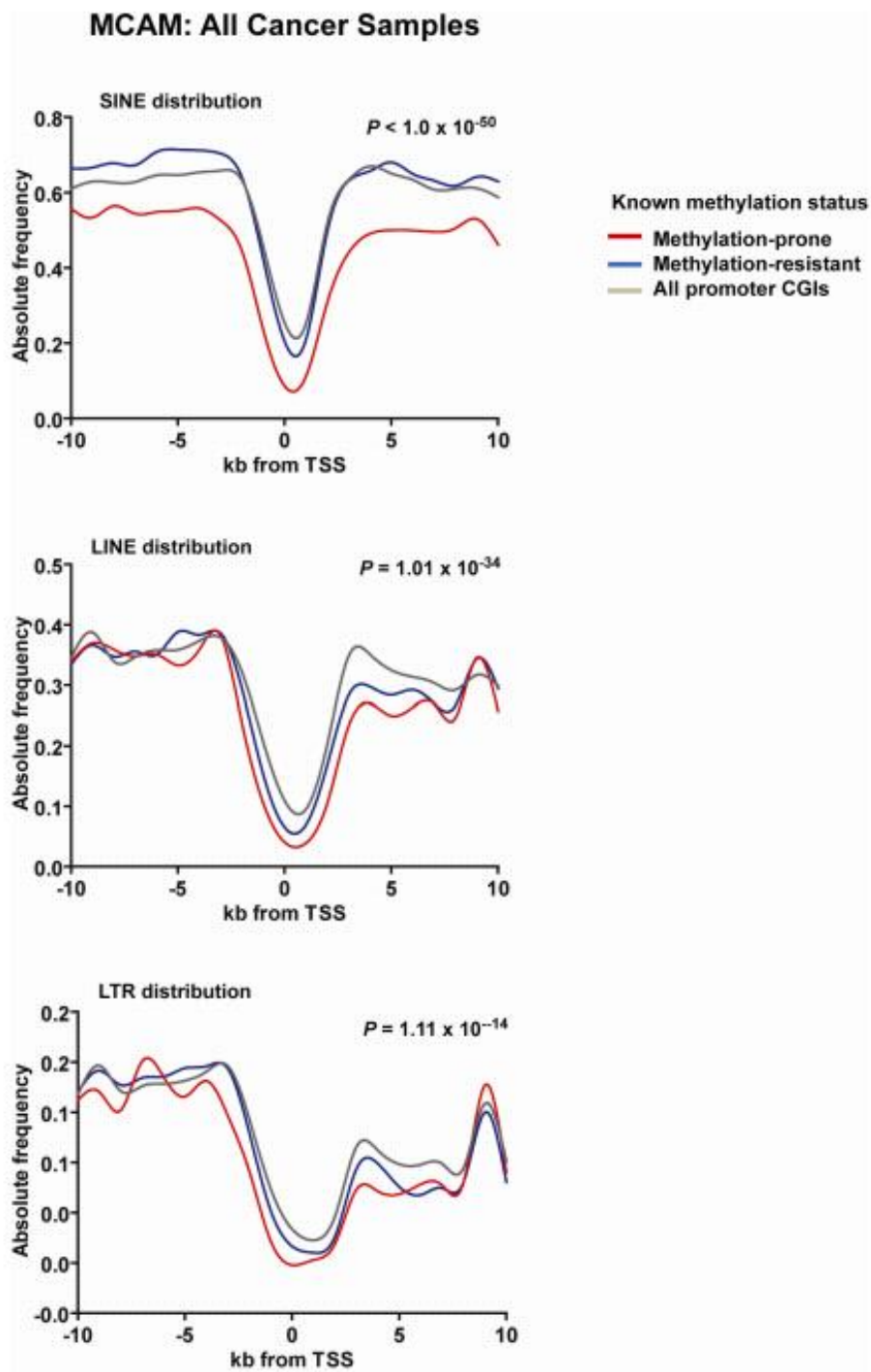


Figure 7. “Distribution of SINE, LINE and LTR repeats around the TSS of frequently hypermethylated and unmethylated genes identified by MCAM analysis of 28 cancer cell lines and 32 primary cancer tissues. The genome-wide distribution of these elements in CpG island promoter genes is shown in gray.” Figure 7 was reproduced with permission from [38].

Lastly, if our model is correct, promoter CpG islands subjected to age-related methylation, which accounts for a large fraction of promoter CGI methylation observed in cancer [24], should also be depleted of SINE and LINE retrotransposons. To test this we compared the distribution of SINE and LINE elements in the 20-kb region surrounding the TSS of more than 6000 promoter CpG islands, that were identified as non-methylated and methylated in young (3 months old) and old (35 months old) mice small intestine tissue by MCAM analysis.

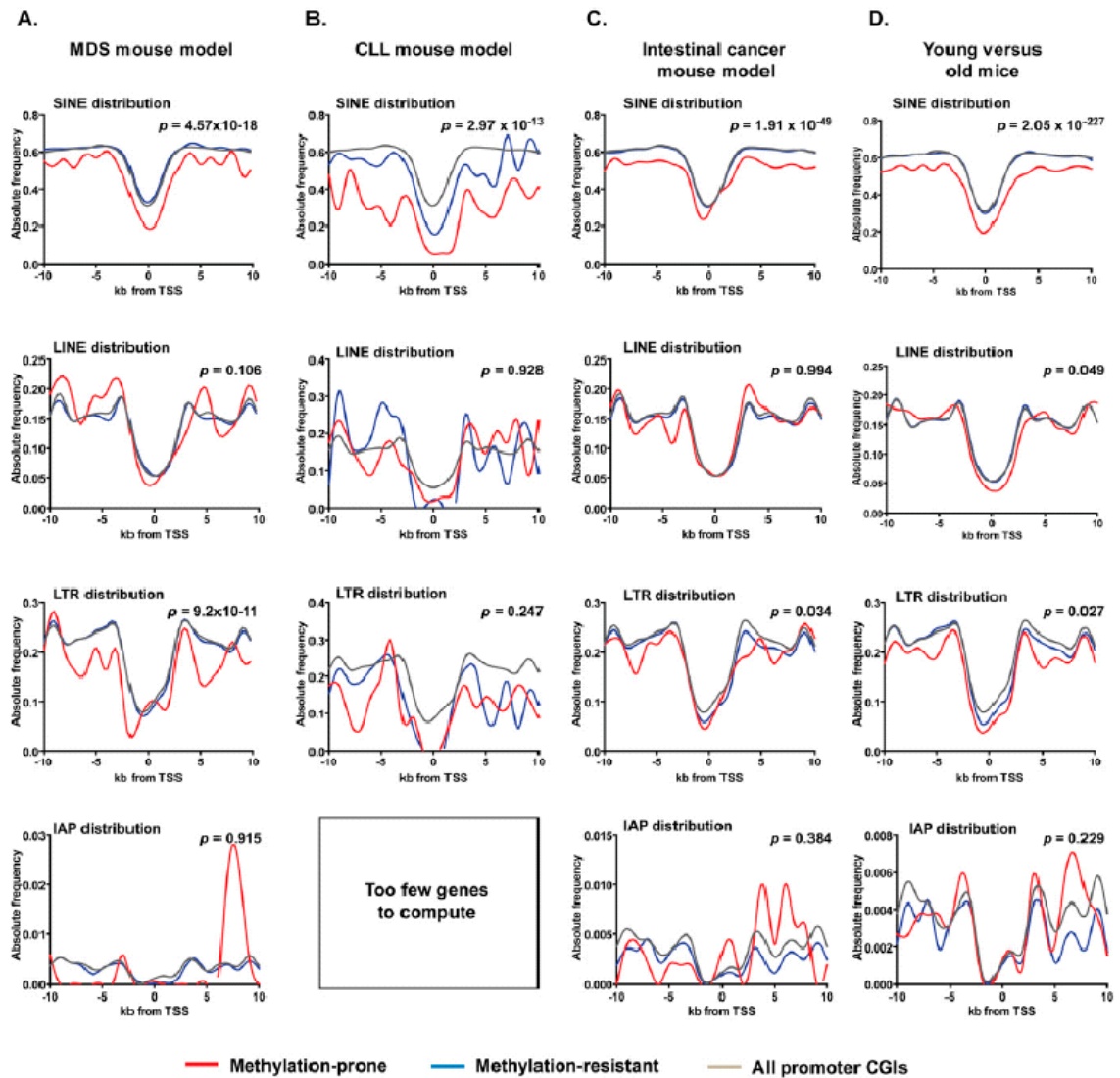


Figure 8. Frequency of retroelements in methylation-prone and methylation-resistant genes identified in mouse cancer models and old mice. (A) Depletion of SINE and LTR but not LINE repeats near TSSmarks methylated promoter CpG islands in a mouse model of myelodysplastic syndrome (MDS). Bone marrow samples of three NUP98-HOXD13 transgene animals that developed MDS (Lin et al. 2005) were studied by MCAM. Bone marrow samples from nontransgene animal of the same mouse strain was used as control, and the methylation status of approximately 6000 CpG island promoter genes was determined in the MCAM experiments. (B,C) The same pattern of retroelements depletion is observed in hypermethylated genes in CLL (Chen et al. 2009) and intestinal cancer mouse models (Hahn et al. 2008). (D) Depletion of SINE, LINE, and LTR repeats near TSS also marks age-related methylation promoter CpG islands. Small intestine tissue harvested from young (3-mo-old) and old (35-mo-old) C57BL/6J mice were used in MCAM experiments to identify age-related methylation. Figure 8 was reproduced with permission from [38].

As displayed in Fig. 8, the results were similar to humans, just as human promoter CpG islands predisposed to methylation in cancer, age-related methylated mouse promoter CpG islands were also depleted of SINE and LINE repeats. Therefore, our data suggest that predisposition to methylation in cancer and aging can be predicted based on the distribution of such elements and that genes that show this predisposition contain a common genome architecture marked by these elements.

The promising results of our predictive model lead us to apply it genome wide. Using the NCBI build 36.1, we had 25,489 unique RefSeq genes, of these genes 16166 or about 63.4% had a promoter CpG island. From these 16166 genes, 3664 or 22.7% were indicated by our predictive model to be prone to methylation (methylation-prone), 7328 or 45.3% of genes were indicated to be moderately predisposed to methylation (methylation-intermediate), and the rest of the 5714 or 32% of genes were predicted as resistant to methylation (methylation-resistant). This description is illustrated by **Fig. 9**.

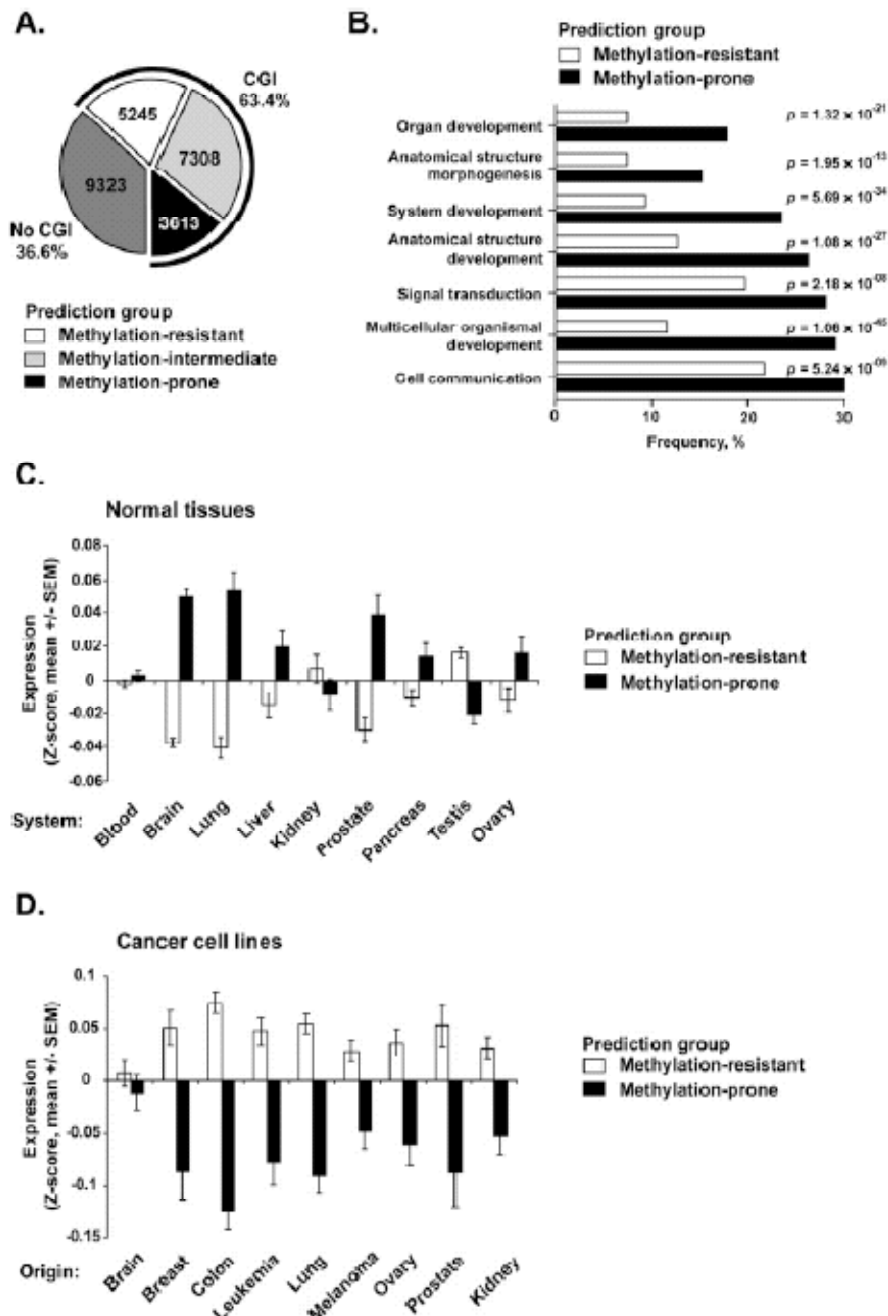


Figure 9: “Genome-wide prediction of predisposition to DNA methylation in cancer. (A) The pie chart shows the number of RefSeq genes with no CpG islands (dark gray) and the number of predicted methylation-resistant (white), methylation-intermediate (light gray), and methylation-prone (black) genes in promoter CpG island genes. (B) Gene Ontology (GO) analysis of 1952 predicted methylation-prone and 2583 predicted methylation-resistant genes for which functional information was available. Horizontal bars represent the frequency of significant GO terms. (C) Gene expression analysis for 2822 promoter CpG island associated genes predicted methylation-prone and 3651 predicted methylation-resistant genes in normal tissues. Expression values were retrieved from the GNF database (Su et al. 2004) and Z-score normalized per

tissue. Each bar represents the mean \pm SEM expression values in each tissue according to their predicted methylation predisposition. (D) Gene expression analysis for 599 promoter CpG island associated genes predicted methylation-prone and 996 predicted methylation-resistant genes in 52 cancer cell lines. Expression values were retrieved from a published work (Ross et al. 2000) and were analyzed as described in C. Only genes present in the studied array platforms could be evaluated, resulting in a different number of analyzed genes in each experiment.” Figure 9 was reproduced with permission from [38].

Tables 5 and Table 6 also show the top 50 genes predicted by our model as prone to methylation (methylation-prone) and resistant to methylation (methylation-resistant). In order to check the biological implications of our model, we compared the mRNA expression of predicted methylation-resistant and methylation-prone promoter CpG island genes in 52 human cancer cell lines and 28 normal differential human tissue using public microarray databases [25, 26].

Table 5: Top 50 predicted Methylation-prone genes (reproduced with permission from [38]).

Gene symbol	Gene name	RefSeq name	Chromosome	Transcription start	S score	L score
ZFPM2	Zinc finger protein, multitype 2	NM_012082	chr 8	106400322	10.0794	2.5599
TOX3	TOX high mobility group box family member 3	NM_001080430	chr 16	51138307	9.7349	2.8928
TPPP	Tubulin polymerization promoting protein	NM_007030	chr 5	746510	10.0794	2.4442
FOXA1	Forkhead box A1	NM_004496	chr 14	37134240	10.3338	2.1021
NKX2-2	NK2 homeobox 2	NM_002509	chr 20	21442664	9.4206	2.9432
TCERG1L	Transcription elongation regulator 1-like	NM_174937	chr 10	132999974	9.0299	3.2956
POU3F3	POU class 3 homeobox 3	NM_006236	chr 2	104838400	10.0794	2.2061
DUX4	Double homeobox, 4	NM_033178	chr 4	191229360	10.0794	1.978
LHX9	LIM homeobox 9	NM_001014434	chr 1	196148257	10.0794	1.9779
EBF3	Early B-cell factor 3	NM_001005463	chr 10	131652081	10.0794	1.9779
GATA3	GATA binding protein 3	NM_002051	chr 10	8136672	10.0794	1.9779
GPR123	G protein-coupled receptor 123	NM_001083909	chr 10	134751398	10.0794	1.9779
IGF2	Insulin-like growth factor 2 (somatomedin A)	NM_000612	chr 11	2116780	10.0794	1.9779
PAX6	Paired box 6	NM_001604	chr 11	31789434	10.0794	1.9779
HOXC4	Homeobox C4	NM_014620	chr 12	52696908	10.0794	1.9779
HOXC8	Homeobox C8	NM_022658	chr 12	52689156	10.0794	1.9779
HOXC9	Homeobox C9	NM_006897	chr 12	52680143	10.0794	1.9779
ZIC2	Zic family member 2 (odd-paired homolog, <i>Drosophila</i>)	NM_007129	chr 13	99432319	10.0794	1.9779
ZIC5	Zic family member 5 (odd-paired homolog, <i>Drosophila</i>)	NM_033132	chr 13	99422179	10.0794	1.9779
CRIP2	Cysteine-rich protein 2	NM_001312	chr 14	105012175	10.0794	1.9779
SIX1	SIX homeobox 1	NM_005982	chr 14	60185933	10.0794	1.9779
NR2F2	Nuclear receptor subfamily 2, group F, member 2	NM_021005	chr 15	94674949	10.0794	1.9779
HOXB4	Homeobox B4	NM_024015	chr 17	44010742	10.0794	1.9779
HOXB5	Homeobox B5	NM_002147	chr 17	44026102	10.0794	1.9779
ZADH2	Zinc binding alcohol dehydrogenase domain containing 2	NM_175907	chr 18	71050105	10.0794	1.9779
TSHZ3	Teashirt zinc finger homeobox 3	NM_020856	chr 19	36532030	10.0794	1.9779
DLX1	Distal-less homeobox 1	NM_178120	chr 2	172658453	10.0794	1.9779
HOXD10	Homeobox D10	NM_002148	chr 2	176689737	10.0794	1.9779
HOXD11	Homeobox D11	NM_021192	chr 2	176680329	10.0794	1.9779
HOXD12	Homeobox D12	NM_021193	chr 2	176672775	10.0794	1.9779
HOXD8	Homeobox D8	NM_019558	chr 2	176702722	10.0794	1.9779
HOXD9	Homeobox D9	NM_014213	chr 2	176695333	10.0794	1.9779
MEIS1	Meis homeobox 1	NM_002398	chr 2	66516035	10.0794	1.9779
NR4A2	Nuclear receptor subfamily 4, group A, member 2	NM_006186	chr 2	156897446	10.0794	1.9779
SATB2	SATB homeobox 2	NM_015265	chr 2	200033446	10.0794	1.9779
POU4F2	POU class 4 homeobox 2	NM_004575	chr 4	147779494	10.0794	1.9779
IRX1	Iroquois homeobox 1	NM_024337	chr 5	3649167	10.0794	1.9779
POU3F2	POU class 3 homeobox 2	NM_005604	chr 6	99389300	10.0794	1.9779
DLX6	Distal-less homeobox 6	NM_005222	chr 7	96473225	10.0794	1.9779
HOXA10	Homeobox A10 (isoform a)	NM_018951	chr 7	27186368	10.0794	1.9779
HOXA10	Homeobox A10 (isoform b)	NM_153715	chr 7	27180480	10.0794	1.9779
HOXA5	Homeobox A5	NM_019102	chr 7	27149812	10.0794	1.9779
HOXA6	Homeobox A6	NM_024014	chr 7	27153893	10.0794	1.9779
HOXA7	Homeobox A7	NM_006896	chr 7	27162821	10.0794	1.9779
HOXA9	Homeobox A9	NM_152739	chr 7	27171674	10.0794	1.9779
SCRIB	Scribbled homolog (<i>Drosophila</i>)	NM_182706	chr 8	144969537	10.0794	1.9779
SCX8	Scleraxis homolog B (mouse)	NM_001080514	chr 8	145461410	10.0794	1.9779
NFIB	Nuclear factor I/B	NM_005596	chr 9	14303945	10.0794	1.9779
METRNL	Meteorin, glial cell differentiation regulator-like	NM_001004431	chr 17	78630855	9.1408	2.8277
OTP	Orthopedia homeobox	NM_032109	chr 5	76970278	9.4805	2.4357

Table 6: Top 50 predicted Methylation-resistant genes (reproduced with permission from [38]).

Gene symbol	Gene name	RefSeq name	Chrom	Transcription start	S score	L score
SMYD4	SET and MYND domain containing 4	NM_052928	chr 17	1679925	-7.6213	-7.5946
SMN2	Survival of motor neuron 2, centromeric	NM_022877	chr 5	70256523	-7.5439	-6.2266
SMN1	Survival of motor neuron 1, telomeric	NM_000344	chr 5	70256523	-7.5439	-6.2266
PPIL2	Peptidylprolyl isomerase (cyclophilin)-like 2	NM_148176	chr 22	20350272	-6.4342	-6.9256
RBM44	RNA binding motif protein 44	NM_001080504	chr 2	238372126	-5.5751	-7.7633
NOSIP	Nitric oxide synthase interacting protein	NM_015953	chr 19	54775615	-6.8053	-6.3264
ZFP1	Zinc finger protein 1 homolog (mouse)	NM_153688	chr 16	73739921	-7.6213	-5.5062
PXMP4	Peroxisomal membrane protein 4, 24 kDa	NM_007238	chr 20	31771797	-6.8522	-6.1378
NHP2L1	NHP2 non-histone chromosome protein 2-like 1 (<i>S. cerevisiae</i>)	NM_005008	chr 22	40408502	-7.6213	-5.3096
DRG1	Developmentally regulated GTP binding protein 1	NM_004147	chr 22	30125538	-7.2939	-5.3636
RPA1	Replication protein A1, 70 kDa	NM_002945	chr 17	1680094	-7.2939	-5.3193
JAGN1	Jagunal homolog 1 (<i>Drosophila</i>)	NM_032492	chr 3	9907271	-6.8225	-5.5777
EP400	E1A binding protein p400	NM_015409	chr 12	131000460	-6.7769	-5.4792
PAAF1	Proteasomal ATPase-associated factor 1	NM_025155	chr 11	73265680	-7.6213	-4.5654
TRPV4	Transient receptor potential cation channel, subfamily V, member 4	NM_147204	chr 12	108755595	-7.6213	-4.51
CDK5RAP2	CDK5 regulatory subunit associated protein 2	NM_018249	chr 9	122382258	-7.6213	-4.4664
IQCD	IQ motif containing D	NM_138451	chr 12	112143263	-6.2945	-5.7063
C12orf32	Chromosome 12 open reading frame 32	NM_031465	chr 12	2856649	-6.2921	-5.683
CYB5RL	Cytochrome b5 reductase-like	NM_001031672	chr 1	54438334	-6.5718	-5.3888
NPRL3	Nitrogen permease regulator-like 3 (<i>S. cerevisiae</i>)	NM_001039476	chr 16	128672	-7.6213	-4.2671
CDK5RAP1	CDK5 regulatory subunit associated protein 1	NM_016408	chr 20	31452998	-7.1813	-4.6612
CCDC101	Coiled-coil domain containing 101	NM_138414	chr 16	28472757	-6.9085	-4.9031
CHCHD8	Coiled-coil-helix-coiled-coil-helix domain containing 8	NM_016565	chr 11	73265538	-7.2895	-4.4949
SLC24A6	Solute carrier family 24 (sodium/potassium/calcium exchanger), member 6	NM_024959	chr 12	112257308	-7.6213	-4.1074
DHX37	DEAH (Asp-Glu-Ala-His) box polypeptide 37	NM_032656	chr 12	124039620	-7.6213	-4.0842
DNAJC8	DnaJ (Hsp40) homolog, subfamily C, member 8	NM_014280	chr 1	28432129	-7.6213	-4.0769
ZNF562	Zinc finger protein 562	NM_017656	chr 19	9646734	-7.8757	-3.8041
DRG2	Developmentally regulated GTP binding protein 2	NM_001388	chr 17	17932007	-5.7396	-5.8857
C16orf45	Chromosome 16 open reading frame 45	NM_033201	chr 16	15435825	-5.9831	-5.6375
PLA2G16	Phospholipase A2, group XVI	NM_007069	chr 11	63138469	-7.8757	-3.6817
FOXR1	Forkhead box R1	NM_181721	chr 11	118347626	-5.4078	-6.0503
KIF3A	Kinesin family member 3A	NM_007054	chr 5	132101164	-6.4861	-4.9674
RNF185	Ring finger protein 185	NM_152267	chr 22	29886178	-7.6213	-3.8222
MRPL37	Mitochondrial ribosomal protein L37	NM_016491	chr 1	54438427	-6.2901	-5.1153
YIPF1	Yip1 domain family, member 1	NM_018982	chr 1	54128041	-4.4365	-6.8572
RAD51L3	RAD51-like 3 (<i>S. cerevisiae</i>)	NM_002878	chr 17	30471001	-7.1327	-4.1404
DNAL1	Dynein, axonemal, light chain 1	NM_031427	chr 14	73181454	-5.9831	-5.265
HLC5	Holocarboxylase synthetase [biotin-(propionyl-CoA-carboxylase [ATP-hydrolyzing]) ligase]	NM_000411	chr 21	37284373	-5.6994	-5.5373
MMP24	Matrix metalloproteinase 24 (membrane-inserted)	NM_006690	chr 20	33278116	-6.7495	-4.4684
MRPL1	Mitochondrial ribosomal protein L1	NM_020236	chr 4	79002828	-5.4106	-5.7917
SETDB1	SET domain, bifurcated 1	NM_012432	chr 1	14916511	-6.9085	-4.291
CTNNA1	Catenin (cadherin-associated protein), alpha 1, 102 kDa	NM_001903	chr 5	138117005	-5.5787	-5.6172
SPNS1	Spinster homolog 1 (<i>Drosophila</i>)	NM_032038	chr 16	28893649	-5.9223	-5.233
C6orf203	Chromosome 6 open reading frame 203	NM_016487	chr 6	107456109	-7.6213	-3.4953
KIF18B	Kinesin family member 18B	NM_001080443	chr 17	40380608	-6.0268	-5.0785
C19orf50	Chromosome 19 open reading frame 50	NM_024069	chr 19	18529603	-7.6213	-3.4712
TMEM219	Transmembrane protein 219	NM_001083613	chr 16	29880851	-6.9625	-4.1224
SLC29A2	Solute carrier family 29 (nucleoside transporters), member 2	NM_001532	chr 11	65895867	-6.3653	-4.5099
ENG	Endoglin	NM_000118	chr 9	129656805	-7.6213	-3.2082
MDM4	Mdm4 p53 binding protein homolog (mouse)	NM_002393	chr 1	202752133	-7.6213	-3.1811

As illustrated in Fig. 10-B and 10-D, genes predicted as methylation-resistant typically showed a lower expression in normal tissue than methylation-prone predicted genes, as would be expected if the model was correct due to the fact that silencing is correlated with hypermethylation of the promoter region. We also observed that genes predicted as prone to methylation were down regulated in cancer cell lines when compared to genes predicted as resistant to methylation (Fig. 10-C and 10-D). Therefore,

the depletion of these elements near the TSS is identified as an independent predictor of down regulation in cancer. The cause of down regulation is likely due to the promoter CpG island methylation rather than the cell culture or even the cell type specific differences in gene expression, since cancer cell lines previously shown to have the highest degree of promoter CpG island methylation showed the lowest average expression of predicted methylation-prone genes (Fig. 10-E).

In an interesting observation, we noted that non-CpG island promoter genes depleted of SINE and LINE retrotransposons were found to be moderately down regulated in cancer (Fig. 10-F).

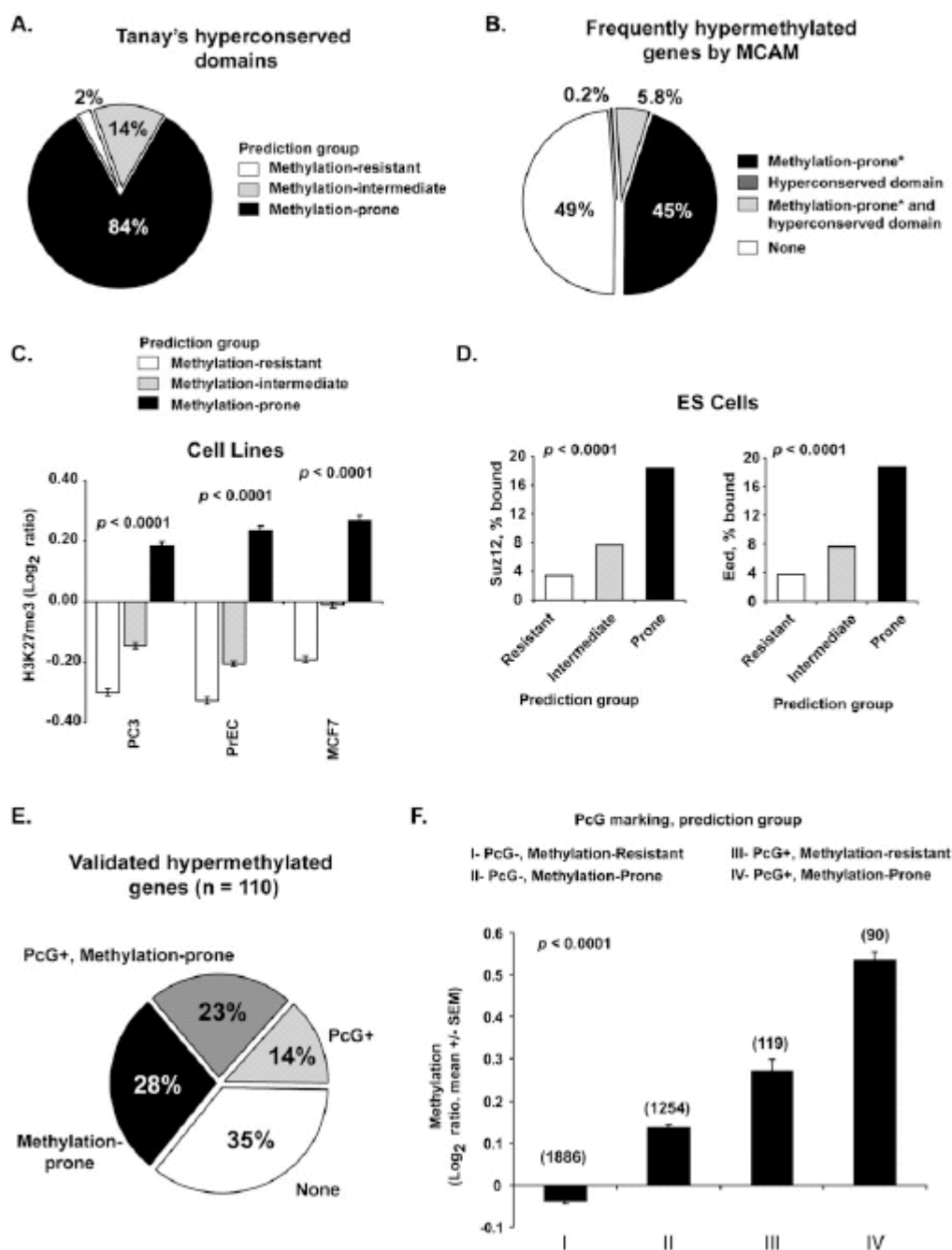


Figure 10. Genome architecture influences on PcG protein binding in embryonic and differentiated cells. (A) Frequency of predicted methylation groups among hyperconserved domains. (B) Relative contribution of hyperconserved domains and retrotransposon depletion in marking frequently methylated genes in cancer. MCAM data from 32 primary tissues and 28 cancer cell lines were averaged to identify frequently methylated genes. *Predicted status. (C) Enrichment of H3K27me3 mark in predicted methylation-prone genes in cancer (PC3, prostate; MCF7, breast) and normal mortalized (PrEC, prostate epithelium) cell lines. H3K27me3 marking was measured by ChIP with microarray hybridization (ChIP-chip) and is quantified as log₂ ratio of pull-down signal

over no antibody signal (Kondo et al. 2008). (D) Frequency of binding of SUZ12 and EED (PcG proteins) in human embryonic stem cells to 2583 methylation-resistant, 3655 methylation-intermediate, and 1690 methylation-prone genes based on our predictive model. Note that genes predicted methylation-prone (thus depleted for SINE and LINE retrotransposons) are preferential targets of PcG proteins. (E) Comparison of PcG marking and our predictive model in identifying methylation-prone genes from our training and first testing set. (F) Average measuredmethylation of predicted methylation-prone andmethylation-resistant genes in PcG marked genes. MCAM data from 32 primary tumors and 28 cancer cell lines were averaged per comparison group, and methylation is presented as log₂ ratio (cancer/control). The number of genes per category is presented above each column.” Figure 10 was reproduced with permission from [38].

Since the influence of non-CpG island promoter methylation on gene expression is limited, this result prompted us to ask the question does the genome architecture defined by SINE and LINE retrotransposons influence other epigenetic events besides DNA methylation? To address this question, we did a comparison of the presence of H3K27me3 between methylation resistant and methylation prone genes according to chromatin imprinted microarray results for 8727 gene promoters in 3 cell lines: PC3, a prostate cancer cell ; MCF7, a breast cancer cell line; and PrEC, an immortalized normal prostate epithelial cell line [27].

Our work reveals a role for selected repetitive elements in determining polycomb group proteins targeting, aberrant DNA methylation, and gene expression in cancer. A paradoxical finding we noticed, was the depletion of the element SINE in genes prone to methylation, finding was paradoxical due to the fact that the main family of these elements (Alu family) were shown as nucleation centers in both plants and animals, and these repeats have been hypothesized to spread DNA methylation in Cancer [32]. However, we must take into account, that these repetitive elements have shown the opposite effect, such as working as insulator function [33, 34]. Despite this observation, the precise mechanism by which these repeat elements contribute to protection is unknown, it is likely that protection from de novo methylation is not directly mediated by these elements, but instead by transcription factors such as CTCF [35] and SP1 [36]

which contain euchromatin/heterochromatin boundary activity. In such cases, genes lacking the binding sites for such boundary proteins would be adversely affected by the insertion of these elements near their promoters and therefore during evolution their presence was counter selected. In fact, it has been displayed that these elements when methylated are preferentially retained or inserted in gene-poor areas; this feature likely came about through negative selection [37]. Aside of the actual mechanism by which retrotransposons participate and are associated to the protection of de novo DNA methylation in cancer; our research suggests that the genomic architecture has a greater influence on disease and physiology than previously suspected.

Conclusion and Future Direction

In recent years epigenetic modifications have been established as a key molecular signature in the progression of tumorigenesis. Discoveries, such as hypermethylation of the CpG island of certain tumor suppressor genes like BRAC1, link DNA methylation to the established genetic understanding for disruption of critical pathways in tumorigenesis. There range from apoptosis to DNA repair, cell cycle regulation and cellular adhesion. Therefore, hypermethylation of the promoter is now viewed as an established mechanism for gene inactivation. A framework for understanding the possible relationship between mutational events and altered DNA methylation is presented. A case is made here to show how retrotransposable elements, specifically LINEs and SINEs, are likely to function as key players a variety of toxicities, including but not limited to carcinogenesis. As it is known methylation plays an important role in suppressing the gene activity of the inactive X chromosome in female mammals. Methylation also plays a role in suppression of transposable elements, for example Alu sequences within the gene. Speak on tumor suppressor genes hypermethylation and how they disrupt and contribute to the disruption of many cell pathways. The genes p15, p14, and p16 on chromosome 9p21 are methylated in several cancers.

Future direction upon this project is to use other algorithm and methods such as R-scan which is not restricted or is confined to a certain window. Development of a threshold which would allow for better separation of methylation-prone genes and methylation-resistant may create a more accurate using elements such as the retrotransposons mentioned in this report. Although this study had some flaws, such as threshold selection, predetermined windows, exclusion of other element (information)

that may strengthen its predictive power, the Positional Weight Matrix showed evidence of prediction with a high percentage value. Therefore, we found evidence that confirms our belief that the genomic architecture marked by retrotransposon elements does in fact play a part in the regulation of epigenetics. As for limitations of our research, to cite a few, are that we can not model selection (some genes are tumor suppressor genes and may be found methylated more frequently due their function in cancer, while oncogenes, once silenced, may lead to cell elimination) and that we can only predict methylation across multiple tissue, but not tissue-specific methylation. Therefore, we can better predict genes that will be methylated in colon, breast or liver, but we cannot pick genes that are exclusively methylated in one specific tissue. In short, our model is a generalization of the microenvironment that allows for the de-novo methylation, but we do not identify the triggers of the process. In conclusion, our findings help the field as a whole by showing that not only selection and gene function are at work in gene silencing, but also the gene microenvironment, and sheds light in the importance of genome organization in diseases. Also, our data supports previous work from Vertino, where some degree of concordance between repetitive elements and methylation was mentioned. It refutes the idea that histone marking is the most crucial marker of methylation-predisposition, as we show when comparing our method against H3K27me3.

References

1. JABLONKA, E. and LAMB, M. J. (2002), The Changing Concept of Epigenetics. *Annals of the New York Academy of Sciences*, 981: 82–96. doi: 10.1111/j.1749-6632.2002.tb04913.x
2. Russo, V.E.A., Martienssen, R.A., and Riggs, A.D. 1996. *Epigenetic mechanisms of gene regulation*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
3. Esteller M. The necessity of a human epigenome project. *Carcinogenesis* 2006;27:1121-1125.
4. Ross, S. A, Milner, J. A (2007). Epigenetic modulation and cancer: effect of metabolic syndrome?. *Am J Clin Nutr* 86: 872S-877S.
5. Dolinoy DC, Weidman JR and Jirtle RL (2007). Epigenetic gene regulation: linking early developmental environment to adult disease. *Reprod. Toxicol.* 23: 297-307.
6. Dolinoy DC, Das R, Weidman JR, Jirtle RL. Metastable epialleles, imprinting, and the fetal origins of adult diseases. *Pediatr Res* 2007; 61:30R–7R
7. Chorderet DF, Gartler SM: Analysis of CpG suppression in methylated and nonmethylated species. *Proc Natl Acad Sci USA* 1992, 89:957-961.
8. Baylin, S.B. and Herman, J.G. (2000) DNA hypermethylation in tumorigenesis: epigenetics joins genetics. *Trends Genet.*, 16, 168–174.
9. JONES, P. A. (2003), Epigenetics in Carcinogenesis and Cancer Prevention. *Annals of the New York Academy of Sciences*, 983: 213–219. doi: 10.1111/j.1749-6632.2003.tb05976.x
10. Lander, E. S, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J. International Human Genome Sequencing Consortium (2001). Initial sequencing and analysis of the human genome. *Nature* 409, 860-921.

11. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG. (2001). The sequence of the human genome. *Science* 291, 1304-1351.
12. Antequera, F. & Bird, A. CpG islands. *Exs.*, 64:169-85, 1993, Bird Adrian. DNA methylation patterns and epigenetic memory. *Genes Dev.* 2002 Jan 1;16(1):6–21.
13. Baylin, S.B. and Herman, J.G. (2000) DNA hypermethylation in tumorigenesis: epigenetics joins genetics. *Trends Genet.*, 16, 168–174.
14. Maze, I. and Nestler, E. J. (2011), The epigenetic landscape of addiction. *Annals of the New York Academy of Sciences*, 1216: 99–113. doi: 10.1111/j.1749-6632.2010.05893.x
15. Scarpa S, Cavallaro RA, D'Anselmi F, Fusco A., Gene silencing through methylation: an epigenetic intervention on Alzheimer disease. *J Alzheimers Dis*, 2006. 9(4): p. 407 - 14.
16. Jones, Peter A., and Peter W. Laird. "Cancer Epigenetics Comes of Age." *Nature Genetics* 21 (1999): 163-167.
17. Knudson AG. Chasing the cancer demon. *Annu Rev Genet* 2000;34:1–19.
18. Zuckerkandl E, Cavalli G. *Combinatorial epigenetics, "junk DNA", and the evolution of complex organisms*. *Gene*. 2007 Apr 1;390(1-2):232-42.
19. Arnaud P, Goubely C, Pélissier T, Deragon JM. SINE retrotransposons can be used in vivo as nucleation centers for de novo methylation. *Mol. Cell. Biol.* 20 (2000), pp. 3434–3441.
20. Shen, L., Kondo, Y., Ahmed, S., Boumber, Y., Konishi, K., Guo, Y., Chen, X., Vilaythong, J. N., and Issa, J. P. (2007a). Drug sensitivity prediction by CpG island methylation profile in the NCI-60 cancer cell line panel. *Cancer research* 67, 11335-11343.
21. Shen, L., Kondo, Y., Guo, Y., Zhang, J., Zhang, L., Ahmed, S., Shu, J., Chen, X., Waterland, R. A., and Issa, J. P. (2007b). Genome-wide profiling of DNA

- methylation reveals a class of normally methylated CpG island promoters. *PLoS genetics* 3, 2023-2036.
22. Walsh, C. P., Chaillet, J. R., and Bestor, T. H. (1998). Transcription of IAP endogenous retroviruses is constrained by cytosine methylation. *Nature genetics* 20, 116-117.
 23. Estecio, M. R., Yan, P. S., Ibrahim, A. E., Tellez, C. S., Shen, L., Huang, T. H., and Issa, J. P. (2007). High-throughput methylation profiling by MCA coupled to CpG island microarray. *Genome research* 17, 1529-1536.
 24. Toyota, M., and Issa, J. P. (1999). CpG island methylator phenotypes in aging and cancer. *Seminars in cancer biology* 9, 349-357.
 25. Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S. S., Van de Rijn, M., Waltham, M. (2000). Systematic variation in gene expression patterns in human cancer cell lines. *Nature genetics* 24, 227-235.
 26. Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G.. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* 101, 6062-6067.
 27. Kondo, Y., Shen, L., Cheng, A. S., Ahmed, S., Bumber, Y., Charo, C., Yamochi, T., Urano, T., Furukawa, K., Kwabi-Addo, B.. (2008). Gene silencing in cancer by histone H3 lysine 27 trimethylation independent of promoter DNA methylation. *Nature genetics* 40, 741-750.
 28. Ohm, J. E., McGarvey, K. M., Yu, X., Cheng, L., Schuebel, K. E., Cope, L., Mohammad, H. P., Chen, W., Daniel, V. C., Yu, W.. (2007). A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nature genetics* 39, 237-242.

29. Schlesinger, Y., Straussman, R., Keshet, I., Farkash, S., Hecht, M., Zimmerman, J., Eden, E., Yakhini, Z., Ben-Shushan, E., Reubinoff, B. E.. (2007). Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nature genetics* 39,
30. Widschwendter, M., Fiegl, H., Egle, D., Mueller-Holzner, E., Spizzo, G., Marth, C., Weisenberger, D. J., Campan, M., Young, J., Jacobs, I., and Laird, P. W. (2007). Epigenetic stem cell signature in cancer. *Nature genetics* 39, 157-158.
31. Lee, T. I., Jenner, R. G., Boyer, L. A., Guenther, M. G., Levine, S. S., Kumar, R. M., Chevalier, B., Johnstone, S. E., Cole, M. F., Isono, K.. (2006). Control of developmental regulators by Polycomb in human embryonic stem cells. *Cell* 125, 301-313.
32. Jones, P. A., and Baylin, S. B. (2002). The fundamental role of epigenetic events in cancer. *Nature reviews* 3, 415-428.
33. Gdula, D. A., Gerasimova, T. I., and Corces, V. G. (1996). Genetic and molecular analysis of the gypsy chromatin insulator of *Drosophila*. *Proceedings of the National Academy of Sciences of the United States of America* 93, 9378-9383.
34. Lunyak, V. V., Prefontaine, G. G., Nunez, E., Cramer, T., Ju, B. G., Ohgi, K. A., Hutt, K., Roy, R., Garcia-Diaz, A., Zhu, X.. (2007). Developmentally regulated activation of a SINE B2 repeat as a domain boundary in organogenesis. *Science* (New York, NY 317, 248-251.
35. Bell, A. C., West, A. G., and Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. *Cell* 98, 387-396
36. Mummaneni, P., Yates, P., Simpson, J., Rose, J., and Turker, M. S. (1998). The primary function of a redundant Sp1 binding site in the mouse *aprt* gene promoter is to block epigenetic gene inactivation. *Nucleic acids research* 26, 5163-5169.

37. Hollister, J. D., and Gaut, B. S. (2009). Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome research*.
38. Estécio MR, Gallegos J, Vallot C, Castoro RJ, Chung W, Maegawa S, Oki Y, Kondo Y, Jelinek J, Shen L, Hartung H, Aplan PD, Czerniak BA, Liang S, Issa JP. Genome architecture marked by retrotransposons modulates predisposition to DNA methylation in cancer. *Genome Res* 2010;20: 1369–82.
39. Ammie N. Carnell and Jay I. Goodman. The Long (LINEs) and the Short (SINEs) of It: Altered Methylation as a Precursor to Toxicity *Toxicol. Sci.* (2003) 75(2): 229-235 first published online May 28, 2003 doi:10.1093/toxsci/kfg138
40. Stormo, G.D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

Vita

Juan Gallegos was born on July 1st, 1980 in Matamoros, Tamps, Mexico to Celso and Maria Gallegos. He was raised in Houston, TX and attended H.I.S.D. throughout his youth. While having a difficult environment and very few chances to excel, he did not let his surrounding hinder him from doing his best. Juan graduated from Sam Houston High School and then continued his education at University of Houston-Downtown where he received his B.S. in Applied Mathematics. An incredible accomplishment, since his first few semesters at University of Houston-Downtown was solely dedicated to catch up by taking remedial courses. Although, he was not well prepared he was not discourage and decided to stay the course and complete his education.

In August of 2005 he was fortunately accepted to become part of the University Of Texas Graduate School Of Biomedical Sciences in pursuit of a graduate degree in Biostatistics. He continues to gain confidence and develop his academic career. This M.S. thesis is evidence to those that did not believe and a thank you to those who have helped him along the way. He continues to seek knowledge and is still in love with education. He plans to one day to return and ultimately finish his Ph.D. in Biostatistics.

Permanent address:

Juan Gallegos

10604 Park LN.

Houston, TX 77093