NORTHWESTERN UNIVERSITY

DNA Methylation in Repetitive Elements and Cancer

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Biomedical Informatics

By

Yinan Zheng

EVANSTON, ILLINOIS

September 2017

ABSTRACT

DNA Methylation in Repetitive Elements and Cancer

Yinan Zheng

DNA methylation in repetitive elements (RE) suppresses their mobility and maintains genomic stability, and decreases in it are frequently observed in tumor and/or surrogate tissues. Averaging methylation across RE in the genome is widely used to quantify global methylation. Methylation of RE in humans is considered a surrogate for global DNA methylation. Previous studies of RE methylation and cancer risk are inconsistent. Methylation may vary in specific RE loci and play diverse roles in disease development. Averaging methylation across RE, though offers a bird's eye view of global methylomic status, may lose significant biological information.

We seek an approach that can profile methylation in locus-specific RE. Ambiguous mapping of short reads by and the high cost of current bisulfite sequencing platforms make them impractical for quantifying locus-specific RE methylation, particularly in population studies and clinical trials. Although microarray-based approaches (e.g., Illumina's Infinium methylation arrays) provide cost-effective and robust genome-wide methylation quantification, the number of interrogated CpGs in RE remains limited. We then developed a random forest-based algorithm (and corresponding R package, `REMP`) that can accurately predict genome-wide locus-specific RE methylation based on Infinium array profiling data. We validated its prediction performance using alternative sequencing and microarray data. Testing its clinical utility with The Cancer Genome Atlas data demonstrated that our algorithm offers more comprehensively extended

locus-specific RE methylation information that can be readily applied to large human studies in a cost-effective manner. Furthermore, regulatory element enrichment and KEGG enrichment analyses revealed that hypomethylated RE in cancer may play a role in transcription activation, potentially impair chromatin stability, and regulate key cancer-related pathways. In addition, we observed a significant positive correlation between intronic locus-specific RE methylation and the host gene expression.

Our work reveals the spatial dynamics of DNA methylation in RE across the human genome and its relations to cancer. Our findings may 1) promote RE methylation research in large human population and clinical studies; and 2) drive further investigations on the biological and pathological effects of RE methylation and improve our understanding of the role of global methylation in human diseases, especially cancer.

ACKNOWLEDGEMENTS

The pursuit of my Ph.D study is fruitless without the involvement of many people who provide mentorship and direction on the multiple aspects of research and of life. I apologize in advance for inadvertently omitting anyone that has helped me during my Ph.D study at Northwestern University.

I would like to express my sincere gratitude to my advisor Dr. Lifang Hou and Dr. Wei Zhang for their scientific support of my PhD study and related research, for their patience, motivation, immense knowledge, and inspiring ideas. Their guidance helped me in all the time of research and writing of my thesis. I am grateful to Dr. Justin Starren, Dr. Neil Jordan, Dr. Lucy Bilaver, and Dr. Suzanne Cox, for their tremendous efforts in creating a diverse and collaborative learning environment for biomedical informatics students at the Health Sciences Integrated Program. I thank my committee members, Dr. Rosemary Braun and Dr. Firas Wehbe, for their support of my training and their valuable feedbacks.

My gratitude is also extended to my fellow labmates, Brian Joyce, Zhou Zhang, and Tao Gao for the stimulating discussion, for the time working together on projects and grants, and for all the fun we have had in the last five years. I am also grateful for all my friends at Northwestern University. Their kindness and support made my experience at Northwestern memorable. On a personal note, I must thank my family, but particularly my wife Zhen Chen and my parents for their constant support and unconditional love.

PREFACE

Chapter 1 includes an overview of the DNA repetitive elements and their impact on human diseases with an emphasis on cancer.

Chapter 2 is derived from an original publication entitled "Prospective changes in global DNA methylation and cancer incidence and mortality" published in British Journal of Cancer with minor modifications (Joyce et al., 2016), distributed under a Creative Commons license (Attribution-Noncommercial). This chapter includes a pilot study using Alu and LINE-1 methylation as surrogates of global methylation to prospectively investigate their associations with cancer risk.

Chapter 3, 4, and 5 are derived from an original publication entitled "Prediction of genome-wide DNA methylation in repetitive elements" published in Nucleic Acids Research with minor modifications (Zheng et al., 2017a). Reuse permission has been obtained from Oxford University Press (License Number: 4156071079513). Chapter 3 introduces a machine learning algorithm for the prediction of DNA methylation in locus-specific RE. Chapter 4 reports the prediction performances. Chapter 5 includes applications of the algorithm in clinical samples and biological implications of locus-specific RE methylation in cancer.

Chapter 6 is derived from an original R package publication entitled "REMP: Repetitive element methylation prediction" published in Bioconductor with minor modifications (Zheng et al., 2017b). This chapter introduces the package structure and demonstrates the usage of the package.

LIST OF ABBREVIATIONS

ac	acetylation
Alu	Alu elements
APC	adenomatosis polyposis coli (tumor suppressor gene)
AUC	area under the ROC curve
BRCA	breast invasive carcinoma
BRCA2	breast cancer 2 (tumor suppressor gene)
cDNA	complementary DNA
CDS	coding DNA sequence
ChIP	chromatin immunoprecipitation
CI	confidence interval
COAD	colon and rectal adenocarcinoma
CpA	5’-cytosine-phosphate-adenine-3’ dinucleotide
CpG	5’-cytosine-phosphate-guanine-3’ dinucleotide
DMR	differentially methylated region
DNase	DNase I hypersensitivity site
ENCODE	The Encyclopedia of DNA Elements
EPIC	Infinium MethylationEPIC BeadChip
FDR	false discovery rate
FWER	family wise error rate
HR	hazard ratio
H2A.Z	histone H2A variant
H3	histone H3
H4	histone H4
HapMap	The International HapMap Project
HM450	Infinium HumanMethylation450 BeadChip
K	lysine
KEGG	Kyoto Encyclopedia of Genes and Genomes
KNN	k nearest neighbor
LCL	lymphoblastoid cell line
LINE	long interspersed elements
LINE-1	long interspersed element-1s
LTR	long terminal repeats
LUSC	lung squamous cell carcinoma
me1	monomethylation
me2	demethylation
me3	trimethylation
NAS	Normative Aging Study

NimbleGen	NimbleGen SeqCap Epi 4M CpGiant
ORF	open reading frame
PRAD	prostate adenocarcinoma
QRF	quantile regression forests
r	Pearson's correlation coefficient
RE	repetitive elements
RE-CpG	CpGs located in RE region
RF	random forest
RF-Trim	RF with prediction reliability control
RMSE	root mean square error
ROC	receiver operating characteristic
RRBS	reduced representation bisulfite sequencing
SD	standard deviation
SEMA3A	semaphorin 3A (tumor suppressor gene)
SINE	short interspersed elements
SVA	SINE-R/VNTR/Alu
SVM	support vector machine
SVR	support vector regression
SW score	Smith-Waterman score
TCGA	The Cancer Genome Atlas
TFBS	transcription factor binding site
TpG	5’-thymine-phosphate-guanine-3’ dinucleotide
TSS	transcription start site
UTR	untranslated region
VNTR	variable number of tandem repeat
WGBS	whole genome bisulfite sequencing

DEDICATION

I dedicate this thesis to my parents and my wife.

TABLE OF CONTENTS

LIST OF FIGURES AND TABLES

CHAPTER 1 INTRODUCTION

I. Prologue

Since completion of the first draft sequence of the human genome in 2001, the number of human protein-coding genes has been revised down from the projected 40,000-100,000 to 30,000-40,000 (Lander et al., 2001). As genome sequence quality and algorithm have improved, the number has been repeatedly revised down to 20,000-25,000 (International Human Genome Sequencing, 2004, Clamp et al., 2007). Recently there are an estimated of only 19,000 human protein-coding genes (Ezkurdia et al., 2014), which is comparable to fly and worn. Human protein-coding genes span 33.45% of the genome, however, exons of protein-coding genes only cover 2.94% of the genome (Encode Project Consortium, 2012). Therefore, the unexpected layer of complexity in human genome lies in the ~97% non-coding sequences and how they potentially regulate the protein-coding genes. However, for a long time, non-coding sequences have received little attention since they were considered non-functional and dismissed with the term ‘junk DNA’. Accumulating studies in the past decade have challenged this notion and it is now well recognized that function of these non-coding sequences could be fundamental yet largely unknown- sometimes referred to as ‘dark matter’ in the human genome. Deciphering the organization and function of the genomic ‘dark matter’ is thus critical and could offer a more comprehensive view of the complexities of the human genome.

II. Overview of DNA repetitive elements

Sequence analysis across different species has demonstrated that biological complexity is not correlated with the abundance of the coding sequence, but is more likely (with a few exceptions) to positively correlate with the increasing abundance of the non-coding sequence of the genome (Taft et al., 2007). DNA repetitive elements (RE) are thought to be a major source of non-coding sequences (Maumus and Quesneville, 2014). RE are certain patterns of DNA that occur in multiple copies throughout the genome. Researchers now have uncovered a diversity of RE families as well as a substantial overall fraction of genomes occupied by these elements. RE usually vastly outnumber gene content and comprise from 30% to over half the whole genomic content across various species (Lander et al., 2001, Goff et al., 2002, Mouse Genome Sequencing et al., 2002, Wicker et al., 2005, Schnable et al., 2009). In the human genome, it is now widely accepted that about 50% of the genome is RE (Lander et al., 2001) (**Figure 1**). Recently this estimate has been raised to about 66-69% (de Koning et al., 2011) using a novel algorithm for RE identification.

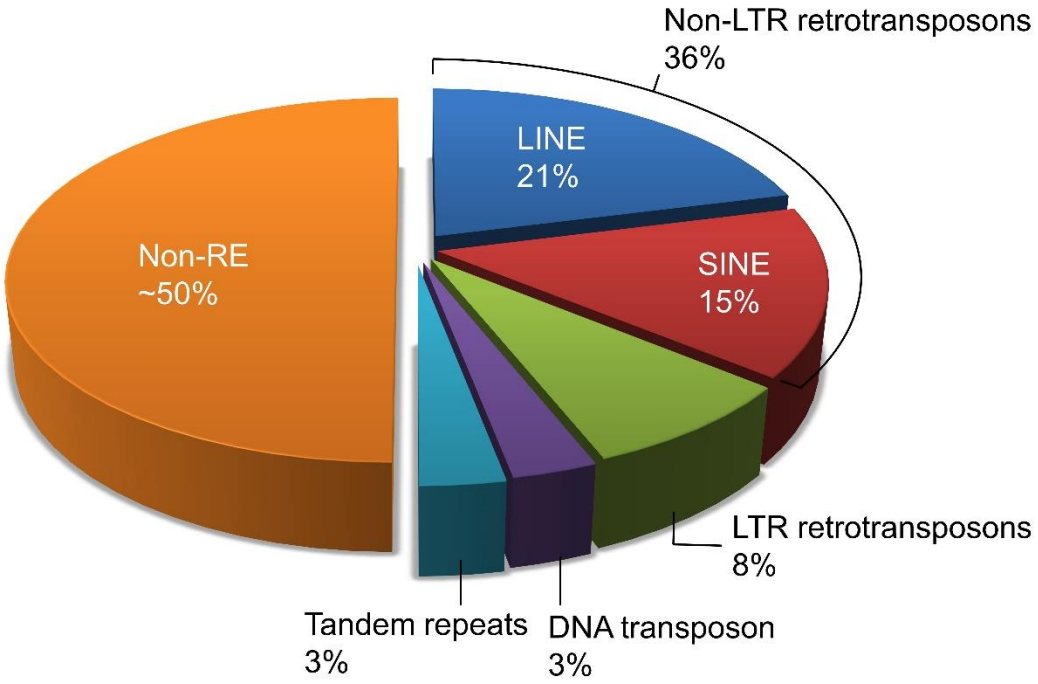

Figure 1 Genomic distribution of major RE types.

The estimates are based on human reference genome build hg19. Note that ~0.3% of human genome are medium sized SVAs (SINE-R/VNTR (variable number of tandem repeat)/Alu) which are a family of non-autonomous retrotransposons within the primate lineage and not shown in this diagram.

Most RE are evolutionary relics of transposons, which are a special type of sequences that can proliferate, mobilize, and populate themselves throughout the genome. RE fall into two major classes based on the mechanism of mobility. One class of RE, named DNA transposons, utilize a 'cut-and-paste' mechanism (**Figure 2A**), where the element sequence is cut from an old genomic location and then inserted into a new genomic location. The other class of RE, termed retrotransposons, employs a 'copy-and-paste' mechanism (**Figure 2B**), where the element sequence is first transcribed into RNA, reverse transcribed in complementary DNA (cDNA), and

the cDNA is then reinserted into a new genomic location with the original element sequence intact.

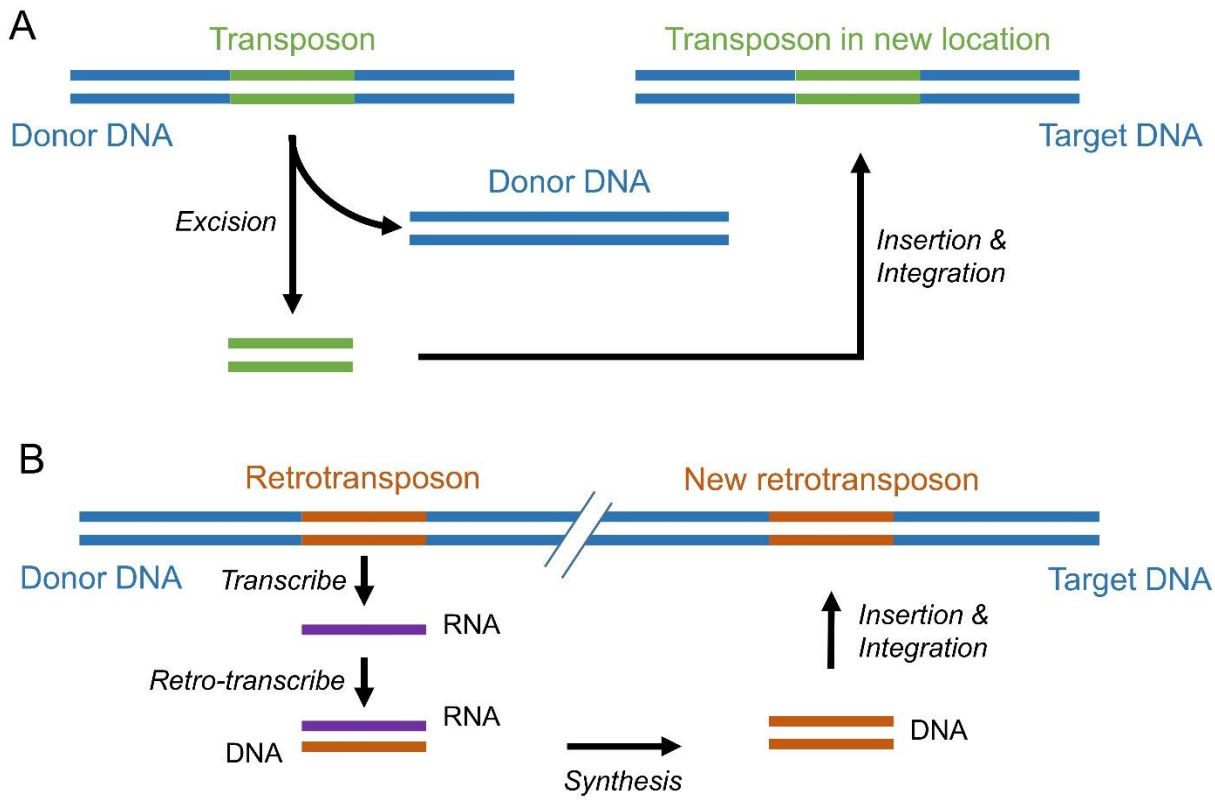

Figure 2 Diagram of mobility mechanism of RE.

A: DNA transposon (cut-and-paste mechanism); B: retrotransposon (copy-and-paste mechanism).

DNA transposons only account for less than 3% of human genome whereas retrotransposons account for ~45% of the human genome (Cordaux and Batzer, 2009, Treangen and Salzberg, 2011). Retrotransposons can be further subdivided into two categories: long terminal repeats (LTR, 8-9% of human genome) and non-LTR retrotransposons (~36% of human genome) (Cordaux and Batzer, 2009, Treangen and Salzberg, 2011). Non-LTR retrotransposons are

ancient genetic elements and long-standing residents in eukaryotic genomes for millions of years and they distinguish themselves from other types of RE for their success multiplying in the human genome. Non-LTR retrotransposons include two sub-families: long interspersed elements (LINE, ~21% of human genome) and short interspersed elements (SINE, ~15% of human genome) (Cordaux and Batzer, 2009, Treangen and Salzberg, 2011). LINE are autonomous element, namely, they are self-sufficient for mobility as they can encode a reverse transcriptase and endonuclease to mobilize themselves. SINE are nonautonomous elements that need to 'hijack' enzymes encoded by autonomous elements (e.g. LINE) and/or the host for their mobility (Ostertag and Kazazian, 2001).

III. Alu element and Long interspersed nuclear elements-1

The most prominent SINE and LINE in human genome are Alu element (Alu) and long interspersed element-1 (LINE-1), respectively, representing the two most abundant types of TE sequences that are currently able to mobilize in modern-day human genomes (Rodic and Burns, 2013, Cordaux and Batzer, 2009) and jointly accounting for approximately 25% of the human genome (Rodic and Burns, 2013, Cordaux and Batzer, 2009).

The RepeatMasker repeat library (Smit et al., 2013) mapped 1,175,329 Alu and 923,315 LINE-1 loci in the UCSC hg19 reference genome assembly, corresponding to 9.9% and 16.4% of the human genome respectively. Most Alu and LINE-1 reside in intergenic (48.3% and 60.5%, respectively) or gene intronic regions (40.0% and 32.0%, respectively) (**Figure 3**).

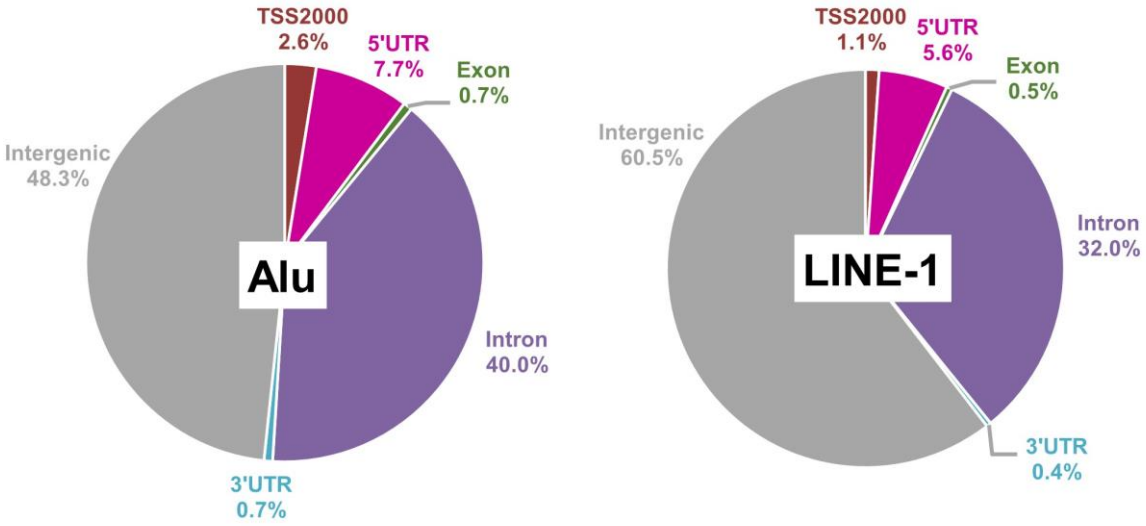

Figure 3 Distribution of Alu and LINE-1 in the human genome (RepeatMasker Library).

Alu arose in mammalian genomes ~65 million years ago (Cordaux and Batzer, 2009). The body of Alu is formed from two monomeric sequences, ancestrally derived from the signal recognition particle RNA (i.e. 7SL RNA) (Batzer and Deininger, 2002, Ullu et al., 1982) and separated by a short adenosine-rich sequence. The left monomer contains two internal RNA polymerase III promoters (Batzer and Deininger, 2002). Active Alu are ~280 base pairs (bp) in length and end in a longer adenosine-rich tail which plays a critical role in its amplification mechanism (Dewannieux and Heidmann, 2005) (**Figure 4**).

LINE-1 have been replicating and evolving in mammals over the past 150 million years (Cordaux and Batzer, 2009). Full length (retrotransposition-competent) LINE-1 are ~6 kb in length and consist of a 5′ untranslated region (UTR), two open reading frames (ORF1 and

ORF2), and a 3′ UTR that is punctuated by an adenosine-rich tail of variable length (Babushok and Kazazian, 2007, Scott et al., 1987) (**Figure 4**). The LINE-1 5′ UTR contains an internal RNA polymerase II promoter that directs transcription of the element (Swergold, 1990); it also contains *cis*-acting binding sites for multiple transcription factors (reviewed in (Beck et al., 2011)). ORF1 is an RNA-binding protein with nucleic chaperone activity (Kolosha and Martin, 1997, Martin and Bushman, 2001) and ORF2 is a multifunctional protein with 1) endonuclease which introduces nicks into the DNA in new location in preparation for LINE-1 insertion, and 2) reverse transcriptase which transcribes RNA intermediates into cDNA copies as a preliminary step in retrotransposition (Feng et al., 1996, Cost et al., 2002, Babushok and Kazazian, 2007). LINE-1 are the only elements able to encode the proteins required for mobilization and they are also responsible for the mobilization of Alu.

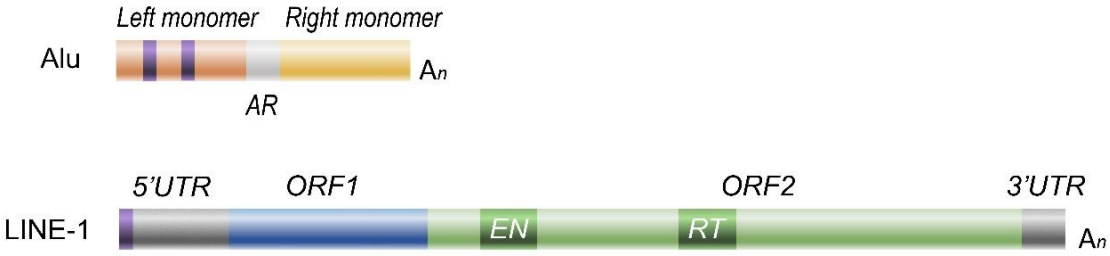

Figure 4 Genetic structure of Alu and LINE-1.

Purple boxes: internal RNA polymerase III (for Alu) or RNA polymerase II (for LINE-1) promoters. AR: the adenosine-rich sequence separating the left and right 7SL monomers; EN: endonulease; RT, reverse transcriptase. A*n*: adenosine-rich (poly-A) tail.

IV. Impact of Alu and LINE-1 upon the human genome

Recent evidence indicates that both Alu and LINE-1 play important roles in genome integrity, gene expression, and disease. *De novo* RE insertions account for about 0.3% of all human mutations (Cordaux and Batzer, 2009). New insertions occurring in approximately one out of 20 births for Alu and out of 200 for LINE-1 (Cordaux and Batzer, 2009, Xing et al., 2009). Alu and/or LINE-1 often target protein-coding genes for insertion (Slotkin and Martienssen, 2007), which may cause genomic instability and contribute to the development of human diseases, particularly cancer (Hancks and Kazazian, 2012, Batzer and Deininger, 2002, Beck et al., 2011, Chen et al., 2006, Criscione et al., 2014). For example, Alu insertion in *BRCA2* gene (tumor suppressor) is associated with breast cancer susceptibility by removing targeted exon from the spliced transcript (Teugels et al., 2005). More interestingly, Alu and LINE-1 insertions in the same insertion location in gene *APC* (tumor suppressor) have been reported to be associated with colon cancer predisposition among two independent populations (Miki et al., 1992, Halling et al., 1999). One possible explanation could be gene enriched in RE sequences may harbor more sequences that resemble endonuclease cleavage sites, possibly increasing the proliferation of RE and creating unexpected alternative splicing events such as exon skipping and introducing alternative splice sites (Konkel and Batzer, 2010). Intronic insertions of LINE-1 have been associated with destabilization of the mRNA resulting in reduced expression (Chen et al., 2006). In addition, insertions of Alu into the 5′ and 3′ prime region of genes can also possibly alter their expression by regulating mRNA stability (Hasler et al., 2007). Recently, RE have been shown to substantially contribute to the origin, evolution, and *cis*-regulation of long non-coding RNA (Kapusta et al., 2013, Kannan et al., 2015) and may also influence gene expression through non-coding RNA-induced degradation of mRNA (Hadjiargyrou and Delihas, 2013). Furthermore,

studies have demonstrated that RE can serve as a supply of regulatory elements and influence the regulation of neighboring gene expression (Feschotte, 2008). For example, some Alu can function as enhancers or contribute the materials for *de novo* birth of enhancers (Su et al., 2014). Therefore, it is fundamental to properly regulate the activity of RE.

CHAPTER 2 GLOBAL DNA METHYLATION AND CANCER

I. DNA methylation in RE: a surrogate of global DNA methylation

DNA methylation in RE is a key regulatory mechanism defending against these transposition activities, and thus maintaining genomic integrity in humans (Morgan et al., 1999, Slotkin and Martienssen, 2007, Bird, 2002, Qu et al., 1999). DNA methylation refers to the addition of a methyl group to DNA, usually the fifth carbon atom of a cytosine ring at the 5'-cytosine-phosphate-guanine-3' dinucleotide sequence, or CpG site. DNA methylation could be directly associated with suppression of RE transcription by blocking the binding of transcription factors to the promoter regions of RE, which is similar to epigenetic regulation of gene expression and chromosome function; or invite mutations as methylated CpGs are prone to be mutated to TpGs (or CpA) dinucleotides by a spontaneous deamination event (Razin and Riggs, 1980).

Decreased DNA methylation in RE, also widely referred as global hypomethylation, plays an important role in tumorigenesis (Ehrlich, 2009, Robertson, 2001, Ehrlich, 2002). Based on the reference genome and the RepeatMasker library, about 35% of all 28 million CpG sites are in Alu (~25%) and LINE-1 (~10%). Over 90% of methylated CpG sites in the human genome occur in RE, particularly Alu and LINE-1 (Beisel and Paro, 2011). Therefore given their genome-wide ubiquity and rich CpG content, bulk estimates of methylation in Alu and/or LINE-1 methylation throughout the genome (Yang et al., 2004) have been widely used as surrogate measures of global DNA methylation content in most human studies (Lisanti et al., 2013, Brennan and Flanagan, 2012).

II. Prospective study of global DNA methylation and cancer risk

a. Inconsistencies in previous studies

Global hypomethylation is predominantly observed in human tumor and surrogate tissues, particularly blood from cancer patients (Esteller, 2007, Lu et al., 2015, Barchitta et al., 2014). Although tissue-specific studies of global DNA methylation have found associations between hypomethylation and increased cancer risk (Bariol et al., 2003, Cravo et al., 1996, Ehrlich et al., 2006, Soares et al., 1999, Van Hoesel et al., 2012), studies involving methylation measured from blood leukocytes have had far more mixed results, with case-control studies finding increased cancer risks associated with hypomethylation (Pufulete et al., 2003, Moore et al., 2008, Lim et al., 2008, Hou et al., 2010, Di et al., 2011) and hypermethylation (Liao et al., 2011, Neale et al., 2014, Walters et al., 2013). Prospective studies of global DNA methylation measured at a single time point in blood leukocytes have also found increased cancer risk associated with both global hypomethylation (Gao et al., 2012) and hypermethylation (Andreotti et al., 2014). Furthermore, to our knowledge, no prospective studies have examined the longitudinal relationship between repeated measures of global DNA methylation and cancer risk over time. There is evidence to suggest a complex biological relationship, with one study suggesting that the timing of blood draws relative to cancer diagnosis matters (Barry et al., 2015) and another finding a 'U-shaped' association (Tajuddin et al., 2014).

Faced with these inconsistencies, we set out to use multiple measures of Alu and LINE-1 DNA methylation and perform an update of our previous investigation of global DNA methylation and cancer risk in the Normative Aging Study (NAS) (Zhu et al., 2011). Our previous study was

subject to a limited sample size which precluded a more thorough analysis of cancer incidence and cancer mortality, as well as less variation in time between sample collection and cancer diagnosis that precluded an in-depth analysis of temporal factors affecting the relationship between global DNA methylation markers and cancer (i.e., the interval between sample collection and cancer diagnosis). Furthermore, this limited sample size forced the inclusion of Alu and LINE-1 methylation measures in blood collected after cancer diagnosis, meaning that it could not rule out potential 'reverse causality' effects of cancer development on Alu and LINE-1 methylation. Our objective is to leverage additional longitudinal data (and new statistical methods for longitudinal analysis developed by our group) to update previous cross-sectional studies regarding Alu and LINE-1 methylation and hazards of cancer incidence and mortality. Our goal is to test the hypothesis that these global DNA methylation surrogates can be temporally dynamic in relation to cancer.

b. Study population and approaches

The NAS was established by the US Department of Veterans Affairs in 1963 with an initial study cohort of 2280 healthy men (age 21-80 years at enrollment) living in the greater Boston area, of primarily (96%) white race. Since then, participants have been recalled periodically for clinical exams every 3-5 years; starting in 1999 these exams included a 7-mL blood sample for genetic and epigenetic analysis. Between January 1st 1999 and December 31st 2012, 802 of 829 active subjects (96.7%) agreed to donate blood for DNA analysis. In total, LINE-1 or Alu methylation was measured at least once in 796 participants. Of these, we excluded 213 participants diagnosed with cancer prior to first blood sample collection, leaving 583 participants cancer-free at the first

blood draw for analysis. This population included our previously published study of LINE-1 and Alu methylation and cancer, with an additional 108 cancer cases and 9 new cancer deaths with NAS data collected from 2007-2012. This study was approved by the Institutional Review Boards of all participating institutions, and written consent forms were obtained from all participants.

In addition to blood samples the NAS consists of anthropometric measurements, standardized medical exams, and questionnaires about medical and smoking history. Our analysis considered the following potential covariates: Race (dichotomized as white or non-white), education (<13 years, 13-16 years, >16 years), two cigarette smoking variables (status of never/former/current, and pack-years), whether the respondent reported taking two or more alcoholic drinks per day on average, body mass index (calculated from weight and height measurements), and age. Our multivariable models also adjusted for white blood cell count and percent neutrophils taken from the blood sample.

Information on cancer diagnoses was obtained from questionnaires and confirmed via review of medical records and histological reports. In contrast to our previous study (Zhu et al., 2011) of Alu and LINE-1 methylation in the NAS, we excluded participants who had been diagnosed with cancer prior to their first blood draw from all analyses. Among the 583 participants free of cancer at first visit, 138 (23.7%) developed cancer during a median of 10.54 (IQR: 6.02-12.10) years of follow up, including 46 prostate cancers and 92 other cancers. For participants who died, death certificates were obtained from the appropriate state health department and reviewed to ensure accurate classification of primary cause of death. These included 37 (6.3%) subjects free

of cancer at the first blood draw who later died of cancer during median of 11.41 (IQR: 8.43-12.34) years of follow up, including 5 subjects who died of prostate cancer.

The full procedure for blood leukocyte DNA extraction and measurement has been reported previously (Hou et al., 2011). To maximize methylation measurement accuracy, we used a pyrosequencing-based assay to measure CpG sites at three positions each for LINE-1 and Alu. All assays used built-in controls. Methylation measurements from each position were averaged via a simple mean, and then standardized by batch number to a mean value of 0 and a standard deviation of 1.

We first examined associations between cancer risk factors at first visit and global DNA methylation using student's t test or fisher's exact test for continuous and categorical variables, respectively. Next, we used multiple Cox proportional hazards regression models to estimate associations between both first visit and time-dependent global DNA methylation, and cancer incidence and mortality. We also examined associations between methylation trajectory and cancer incidence. This was done by comparing the mean difference in methylation between cancer cases and cancer-free individuals each year prior to cancer diagnosis. Due to low sample size, intervals were collapsed into five-year increments (<5 years, 5 to <10 years, and 10+ years). We then analyzed the relationship between methylation rate of change (in units/year) and cancer incidence and mortality using a linear regression model to estimate change in methylation measurement over time for all subjects with multiple measurements, and subsequently treating the beta-coefficient from this model as an independent variable in additional Cox regression models. All models were adjusted for the covariates listed above. All analyses were performed

using SAS (version 9.3, SAS Institute). Two-sided tests were used when comparing means and beta-coefficients, and $p \leq 0.05$ was set as the threshold for statistical significance.

c. Results

Subject characteristics by cancer status have been reported previously (Zhu et al., 2011). Briefly, subjects were male, with a mean age of 72 years at first visit (range 55-100 years), and majority white (95.5%). Characteristics by LINE-1 and Alu methylation at the first visit are listed in **Table 1** .

Table 1 Subject characteristics by global DNA methylation among subjects cancer-free at first visit

	LINE-1			Alu		
	Low	**High**	**p**	**Low**	**High**	**p**
	Mean±SD / n(%)			Mean±SD / n(%)		
Age (years)	71.8±6.6	71.7±6.8	0.97	72.6±6.7	71.0±6.7	0.006*
Body Mass Index (kg/m²)	28.2±3.9	28.3±4.2	0.76	27.9(3.6)	28.7(4.5)	0.031*
Race						
White	271(95.4)	270(95.4)		274(95.8)	281(95.2)	
Non-white	13(4.6)	13(4.6)	0.99	12(4.2)	14(4.8)	0.75
Education (years)						
<13	73(25.7)	89(31.5)		85(29.7)	82(27.8)	
13-16	150(52.8)	130(45.9)		138(48.3)	148(50.2)	
>16	61(21.5)	64(22.6)	0.21	63(22.0)	65(22.0)	0.86
Smoking status						
Never	81(28.5)	75(26.5)		87(30.4)	73(24.8)	
Current	9(3.2)	17(6.0)		14(4.9)	13(4.4)	
Former	194(68.3)	191(67.5)	0.26	185(64.5)	209(70.8)	0.27
Cumulative smoking (pack-years)	19.6±24.6	21.5±23.5	0.38	21.0±24.0	20.6±24.3	0.84
Alcohol Consumption						
0-1 average drinks/day	236(83.1)	232(82.0)		238(83.2)	242(82.0)	
2+ average drinks/day	48(16.9)	51(18.0)	0.73	48(16.8)	53(18.0)	0.71
White blood cell count	6.51±2.76	6.30±1.58	0.28	6.45±2.75	6.40±1.59	0.78
Neutrophils proportion	62.1±8.2	62.1±8.3	0.91	62.5±8.7	61.7±7.9	0.22

*= Statistically significant at $p<0.05$; p-values shown for Student's t-test and Fischer's exact test for continuous and categorical characteristics, respectively

Overall, there were no differences in subject characteristics by LINE-1 methylation. However, subjects with higher Alu methylation tended to be younger (p=0.006) and have a higher Body Mass Index (p=0.031) than subjects with lower Alu methylation.

Table 2 shows the results of our first visit and time-dependent analyses. For cancer incidence, we found a positive association between risk of developing prostate cancer and LINE-1 methylation, which was significant in the time-dependent analysis (Hazard Ratio (HR): 1.38, 95% confidence interval (CI): 1.01-1.88). We found no significant associations between risk of developing cancer and Alu methylation. For cancer mortality, we found a positive association between risk of cancer mortality and LINE-1 methylation in time-dependent analysis (HR: 1.41, 95% CI: 1.03-1.92) as well as Alu methylation at the first visit (HR: 1.39, 95% CI: 1.08-1.79).

Table 2 Association between global DNA methylation with cancer incidence and mortality.

		First Methylation Measurement				Time-dependent Methylation Measurement	
		Cancer-free n	Cancer n	HR (95% CI)	p	HR (95% CI)	p
Incidence							
LINE-1	All Cancer	422	125	1.05(0.88-1.24)	0.60	1.11(0.94-1.31)	0.22
	Prostate Cancer		42	1.32(0.97-1.80)	0.08	1.38(1.01-1.88)	0.045*
Alu	All Cancer	432	129	0.97(0.81-1.17)	0.76	0.95(0.80-1.13)	0.61
	Prostate Cancer		43	0.83(0.58-1.20)	0.33	0.86(0.62-1.19)	0.38
Mortality							
LINE-1	All Cancer	513	34	1.24(0.91-1.68)	0.18	1.41(1.03-1.92)	0.03*
	Prostate Cancer		5	2.52(0.84-7.56)	0.10	1.51(0.68-3.33)	0.31
Alu	All Cancer	524	37	1.39(1.08-1.79)	0.01*	0.91(0.67-1.24)	0.56
	Prostate Cancer		5	1.64(0.62-4.32)	0.32	0.69(0.27-1.78)	0.45

*= Statistically significant at $p<0.05$.

Figure 5 shows graphs of Alu and LINE-1 methylation by time interval between methylation measurement and cancer diagnosis. On average, subjects who ultimately developed cancer had LINE-1 methylation 0.54 units lower than cancer-free subjects 10 or more years before diagnosis (p=0.0056). The differences in mean LINE-1 methylation between cancer cases and cancer-free subjects at other time points were not significant, nor were the other differences in mean Alu methylation.

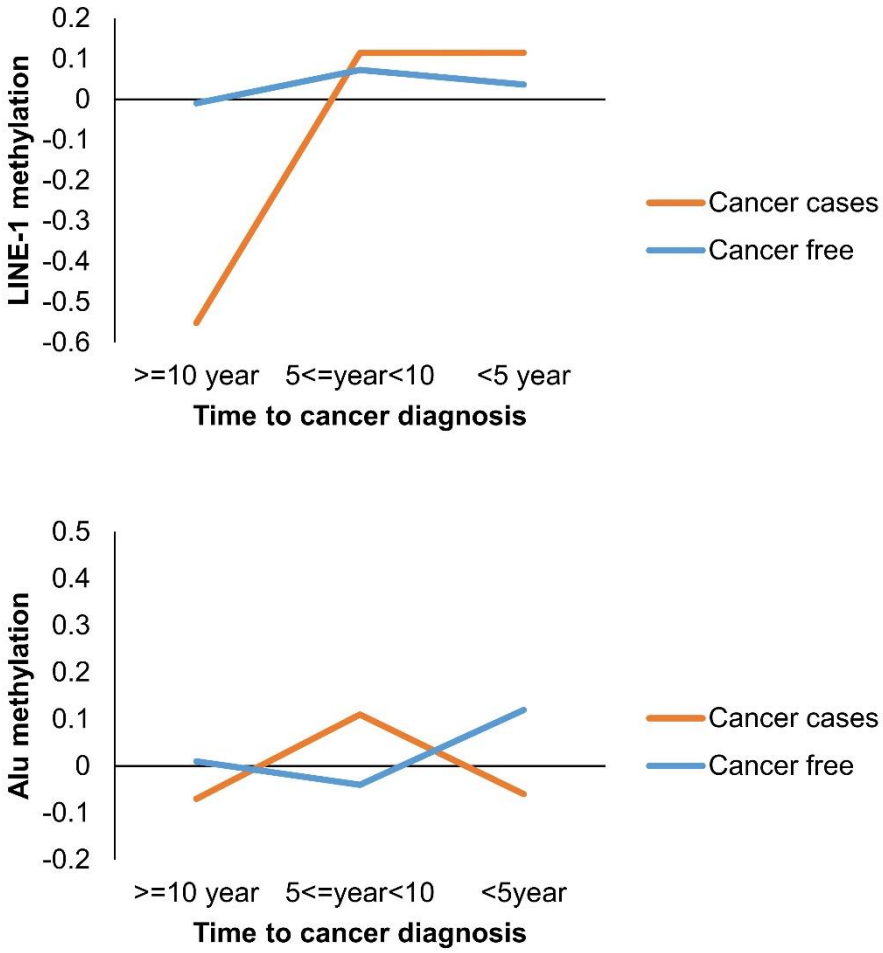

Figure 5 Trajectory of global DNA methylation in cancer and cancer-free subjects by time interval between methylation measurement and cancer diagnosis.

Table 3 shows the results of our rate of change analysis. We found significant associations between all-cancer incidence and the rate of change of Alu methylation (in standardized units/year) (HR: 3.62, 95% CI: 1.09-12.10). We observed no significant associations between rate of change of LINE-1 methylation and cancer incidence. Neither Alu nor LINE-1 rate of change was associated with cancer mortality.

Table 3 Associations between rate of methylation change and cancer.

	Cancer-free	All cancer			Prostate cancer		
	n	n	HR(95% CI)	P	n	HR(95% CI)	P
Incidence							
LINE-1	365	63	1.23(0.29-5.11)	0.77	13	0.75(0.03-19.1)	0.86
Alu			3.62(1.09-12.1)	0.036*		1.58(0.11-22.2)	0.73
Mortality							
LINE-1	389	23	3.88(0.44-34.1)	0.22	4	0.75(0.03-19.1)	0.86
Alu			1.96(0.37-10.2)	0.43		9.97(0.14-700)	0.29

*= Statistically significant at $p<0.05$.

d. Discussion

In this study, we found higher Alu methylation among participants who were younger and had lower Body Mass Index. For cancer incidence time-dependent LINE-1 methylation was positively associated with risk of prostate cancer only. For cancer mortality, Alu methylation at the first blood draw as well as time-dependent LINE-1 methylation were associated with

increased risk of all-cancer mortality. Interestingly, LINE-1 methylation was significantly lower in participants who ultimately developed cancer more than 10 years prior to conventional diagnosis, while a more rapid rate of increase in Alu methylation was significantly associated with risk of developing cancer. Many of these results contradict our previous report of Alu and LINE-1 methylation in the NAS (Zhu et al., 2011). Largely, we believe that this is due to the much larger number of incident cancer cases in our analysis (138 in our analysis vs. 30 in the previous report) as well as our exclusion of prevalent cancers from the mortality analysis (leaving us with 37 deaths due to an incident cancer vs. 11 in the previous report).

While most studies of cancer and Alu and LINE-1 methylation have found inverse relationships between these methylation measures and cancer, the vast majority of these findings have used tumor tissue, or blood leukocytes collected after cancer diagnosis.(Brennan and Flanagan, 2012) Furthermore, two recent meta-analyses found that different measures of global DNA methylation in blood produce different associations with cancer risk, and that the methodological heterogeneity of studies makes drawing conclusions as to its use as a cancer biomarker difficult (Brennan and Flanagan, 2012, Woo and Kim, 2012). The findings of the handful of prospective studies of blood leukocyte Alu or LINE-1 methylation have thus far been less consistent. Three studies of gastric, liver, and breast cancer found no prospective associations between cancer risk and LINE-1 methylation (Balassiano et al., 2011, Brennan et al., 2012, Wu et al., 2012). In another prospective study, LINE-1 hypomethylation was associated with breast cancer risk (Deroo et al., 2014), while two others found marginally significant associations between LINE-1 hypermethylation and kidney (Karami et al., 2015) and bladder cancer (Andreotti et al., 2014). A prospective study of Alu methylation found an inverse association between Alu methylation and

gastric cancer risk, but only with latencies of greater than one year.(Gao et al., 2012) These discrepant results are likely due to high variation in study designs and populations (Brennan and Flanagan, 2012) such as the intervals between blood sample collection and cancer diagnosis, which ranged from 1-16 years in those cited above. A recent prospective analysis of prostate cancer patients found that the relationship between Alu methylation and prostate cancer risk varied by length of this interval, with associations between Alu hypermethylation and increased risk appearing only among subjects diagnosed four or more years after their blood draw (Barry et al., 2015). Although the effects of various intervals between blood sample collection and cancer diagnosis have not been studied in LINE-1 hypermethylation specifically, this effect may also explain our findings of increased prostate cancer risk with time-dependent LINE-1 methylation.

To our knowledge, few studies have examined associations between global DNA methylation of blood leukocytes collected pre-diagnostically and cancer mortality. Two studies found associations between serum LINE-1 hypomethylation and all-cause mortality in populations composed largely of cancer patients (Ramzy et al., 2011, Tangkijvanich et al., 2007), and our prior analysis of NAS data found associations between both LINE-1 and Alu hypomethylation and all-cancer mortality (Zhu et al., 2011). That our results represent a dramatic shift from the prior publication is cause for concern. We feel that this is primarily due to the fact that approximately half (59%) of cancer deaths in the previous study were due to cancers diagnosed prior to the first blood draw, potentially altering the results of the time-dependent analysis. The fact that our mortality analysis was conducted solely on participants who were free of cancer at the first blood draw may explain the discrepancy between our findings and those reported elsewhere in the literature examining cancer mortality among methylation measures obtained

pre- and post-diagnosis. Future studies should further examine Alu and LINE-1 methylation as potential prognostic biomarkers for cancer mortality, and the differences in these relationships between methylation measures pre- and post-diagnosis.

Recent research suggests a dynamic role of Alu methylation during normal development and aging (Luo et al., 2014). Our analysis found a significant association between faster Alu methylation and higher cancer incidence- coupled with what is already known about the involvement of Alu methylation in tumorigenesis, this raises the possibility that Alu may play a dynamic role in cancer development as well. Similarly, we found that on average participants who would ultimately develop cancer had much lower LINE-1 methylation compared to cancer-free participants, but only at approximately 10 years before diagnosis. While both these findings are not unprecedented in this dataset (Joyce et al., 2015), the low sample size of the NAS means that these results should be interpreted with caution. Further validation of these longitudinal biomarkers of cancer is required, but if successful could yield potent new tools for cancer early detection.

Although methylation of LINE-1 and Alu elements have both been widely used as surrogates for global DNA methylation, growing evidence shows that they each have a distinct functional role potentially affecting cancer development and progression via different mechanisms in methylation regulation, responses to cellular stressors and environmental exposures, and methylation levels (Biemont and Vieira, 2006, Rusiecki et al., 2008, Li and Schmid, 2001, Jones and Baylin, 2002). Even specific LINE-1 loci may exert different effects on disease risk (Nusgen et al., 2015) (though there was no evidence for this in our particular data set). Thus, our finding

different relationships between Alu and LINE-1 methylation and cancer is unsurprising despite both being surrogates for global DNA methylation.

Our use of samples collected and stored before cancer diagnosis is a particular strength of this study as it avoids a number of biases inherent in case-control designs, particularly disease- or treatment-induced epigenetic changes. Nonetheless, this study has several limitations. The trade off to collecting a large quantity of data across multiple follow-up measurements is a relatively low sample size. These factors limited our ability to examine specific subtypes of cancer beyond prostate cancer. In addition, as noted above our sample was not representative of the general population. Being older, male, white, educated, and/or having a history of military service may all influence both global DNA methylation and cancer. Therefore, our results will need to be confirmed in other, more representative populations.

CHAPTER 3 Prediction of DNA methylation in locus-specific RE

I. Variability of methylation in locus-specific RE

RE are recently identified as a major source of epigenetic variations in the mammalian genome (Ekram et al., 2012). Accumulating evidence shows that Alu/LINE-1 methylation at specific genomic loci vary and exert distinct biological and/or pathological effects in cancer (Pobsook et al., 2011, Phokaew et al., 2008, Xie et al., 2009, Xie et al., 2011, Szpakowski et al., 2009, Nusgen et al., 2015, Luo et al., 2014) (**Figure 6**)

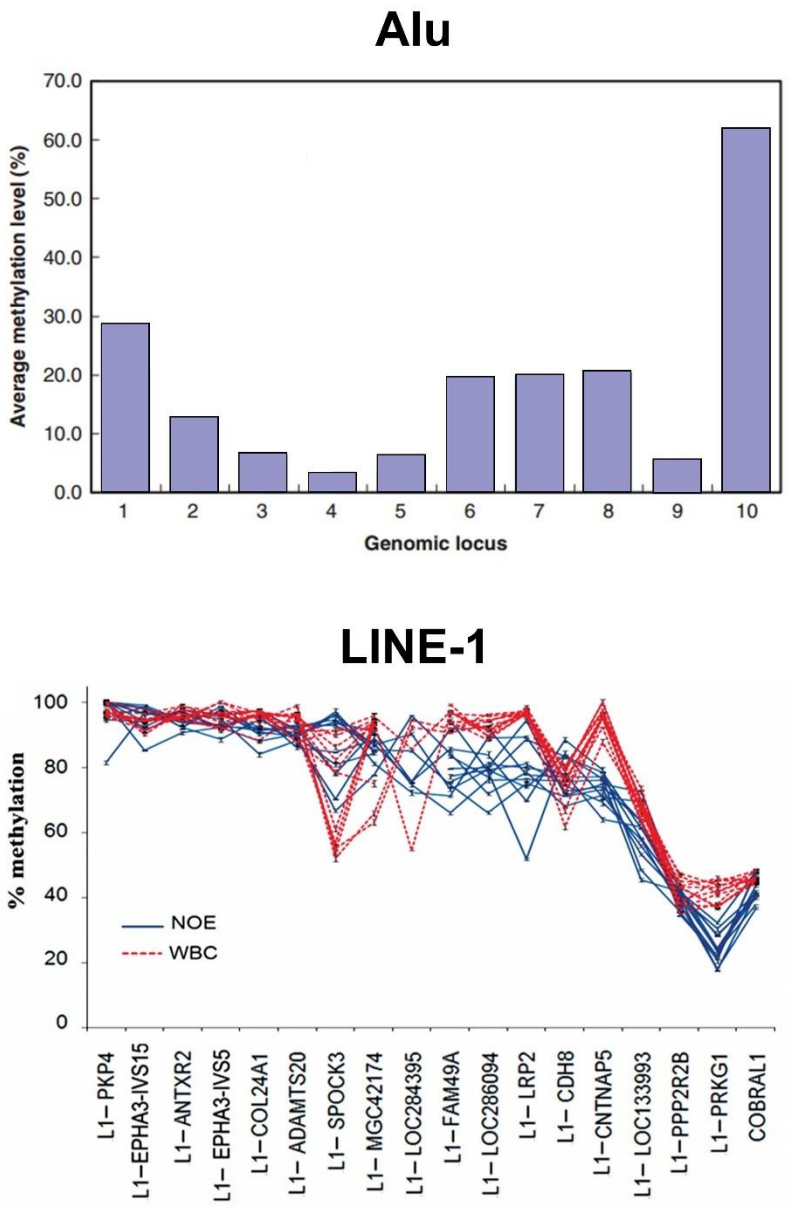

Figure 6 Variability of methylation in different Alu and LINE-1 loci.

The figure is reproduced based on the work in Xie, et al. (Nucleic Acids Research, 2009) and Phokaew, et al. (Nucleic Acids Research, 2008) for Alu and LINE-1, respectively, by permission of Oxford University Press (License Number: 4156080426407 and 4156080588923). Methylation in selected Alu and LINE-1 loci can be as low as 10-20% and as high as 70-100%. NOE: normal oral epithelial; WBC: white blood cell.

These studies suggest that using mean values of methylation in RE as surrogates of global methylation may lead to biological information loss and hindering scientists from elucidating the distinct biological roles of DNA methylation in locus-specific RE. Besides the temporal dynamic role of global RE methylation in tumorigenesis and cancer progression illustrated in **Chapter 2**, we therefore suspect that the previous inconsistent investigations into the roles of RE methylation in cancer for both tissue (Weisenberger et al., 2005) and blood (Brennan and Flanagan, 2012), at least partially, could be also due to the inability to assess RE methylation levels at specific loci.

II. Challenges of DNA methylation profiling in RE

Whole-genome sequencing may plausibly allow us to study locus-specific RE methylation. However single-base resolution sequencing of locus-specific RE is not optimal as the repeats create ambiguities in alignment and assembly, which produce biases and errors when interpreting results (Treangen and Salzberg, 2012). Furthermore, profiling methylation in RE using sequencing is even more challenging as it produces higher mapping errors due to the reduced complexity reads from bisulfite conversion (Stevens et al., 2013, Hansen et al., 2011). Finally, sequencing genome-wide methylation remains prohibitively expensive.

In recent years, microarrays with optimized probes, such as the Infinium HumanMethylation450 BeadChip (HM450) and the upgraded Infinium MethylationEPIC BeadChip (EPIC) (Bibikova et al., 2011), have been widely used for robust genome-wide DNA methylation investigations in

human studies. These array-based DNA methylation data may provide a cost-effective opportunity to study the role of locus-specific RE methylation in relation to cancers and other chronic diseases. However, RE coverage of Infinium methylation arrays are still limited and the profiled CpGs in RE are generally sparse.

This issue can be further elaborated using the HapMap (The International HapMap Project) lymphoblastoid cell line (LCL) GM12878, a Tier-1 sample from a female Utah resident with ancestry from Northern and Western Europe (International HapMap, 2003, International HapMap, 2005). CpG coverage in Alu and LINE-1 were investigated among four single-base methylation profiling approaches, i.e., HM450/EPIC, NimbleGen SeqCap Epi 4M CpGiant (NimbleGen) (Duhaime-Ross, 2014), Reduced Representation Bisulfite Sequencing (RRBS) (Meissner et al., 2005), and Whole Genome Bisulfite Sequencing (WGBS) (Lister et al., 2009). While all approaches except WGBS suffered from depleted coverage in Alu and LINE-1, all platforms cover a variety of Alu/LINE-1 subfamilies (**Table 4**).

Table 4 Alu/LINE-1 coverage with single-base profiling platforms.

	# of RE	# of RE CpGs	# of genes covered [a]	# of RE subfamilies covered
		Alu		
HM450	12255	13155	14276	37
EPIC	21300	23784	19856	40
NimbleGen	1289	2463	2178	31
RRBS	2985	5902	3266	34
WGBS	929874	3652457	40870	41
		LINE-1		
HM450	8309	9797	7399	115
EPIC	24713	29404	15558	116
NimbleGen	4667	13376	4617	115
RRBS	753	2023	663	94
WGBS	586345	2141737	31928	117

[a] RefSeq genes, including gene proximal promoter region (i.e., 2000 bp upstream of the transcription start site).

HM450/EPIC achieved the second highest coverage, significantly higher than NimbleGen and RRBS. To evaluate the reliability of profiled CpGs in Alu/LINE-1, inter-platform correlation and error were calculated to compare concordance between Alu/LINE-1 CpGs vs non-Alu/LINE-1 CpGs (with high concordance indicating robust methylation profiling). HM450/EPIC achieved high concordance with correlations of 0.93 vs 0.96 and errors of 0.094 vs 0.090 for Alu/LINE-1 vs non-Alu/LINE-1 CpGs (**Figure 7A**), respectively. Hence with HM450/EPIC as the benchmark, concordance of NimbleGen was the highest, whereas in RRBS and WGBS correlations decreased and errors increased among Alu/LINE-1 CpGs (**Figure 7B**), suggesting potential measurement bias due to the ambiguous mapping of reads.

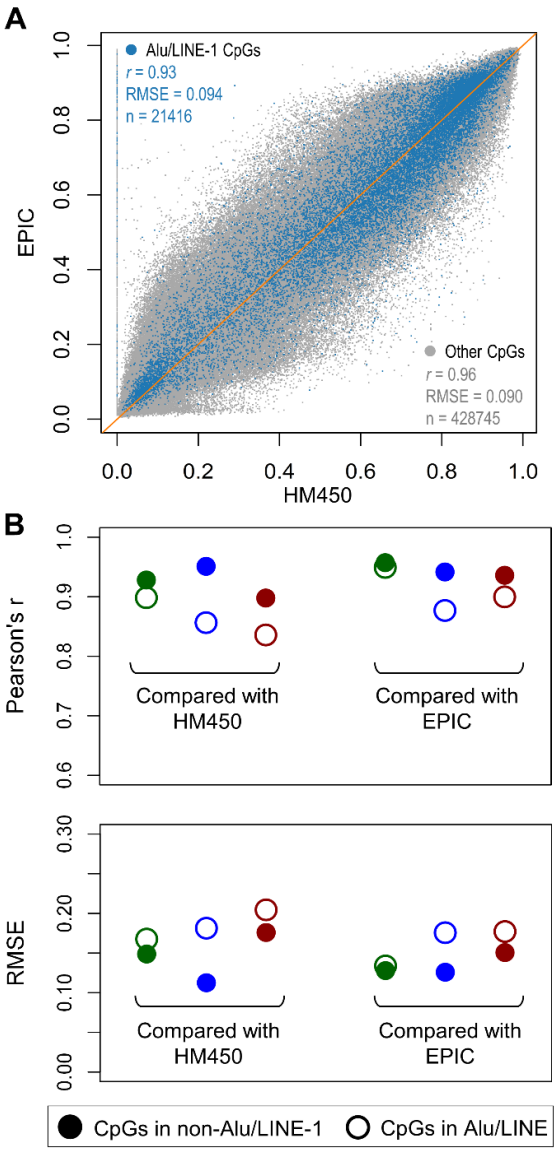

Figure 7 Reliability of the profiling platforms interrogating CpG sites in Alu and LINE-1.

If probes or reads targeting RE regions such as Alu and LINE-1 are affected by ambiguous mapping, methylation readings on these CpGs are more likely to yield different values for the same sample across different platforms. A: Plot showing high correlation between CpGs profiled using both HM450 and EPIC, with CpGs in Alu/LINE-1 showing slightly smaller *r* and larger RMSE (root mean square error). B: Evaluation of the reliability of the three sequencing-based platforms (using Infinium methylation arrays as the benchmark): NimbleGen (green), RRBS (blue), and WGBS (red). NimbleGen shows the highest concordance between both Alu/LINE-1 and non-Alu/LINE-1 CpGs.

Previous studies have shown that the methylation levels of two nearby CpG sites are more likely to be co-methylated (Bell et al., 2011, Eckhardt et al., 2006, Zhang et al., 2015, Li et al., 2010). Given the superior methylation profiling quality with the HM450/EPIC and NimbleGen, we opted to use the HM450/EPIC as the input data source for prediction and NimbleGen as the validation data source. We therefore proposed to develop a predictive algorithm to computationally extend RE methylation based on the Infinium methylation array data. We further evaluated the prediction performance of our algorithm and demonstrated the algorithm's clinical utilities by exploring the biological implications of locus-specific Alu/LINE-1 methylation in cancer. To facilitate calculations, we developed an R package, `REMP` (*Repetitive Element Methylation Prediction*), available in Bioconductor repository.

III. Development of prediction framework

a. Database

For RE identification and annotation, we used the RepeatMasker (Smit et al., 2013) and NCBI RefSeq Gene databases (Pruitt et al., 2014) to identify and annotate candidate RE loci for methylation prediction. We obtained the RepeatMasker Library (build hg19) and RefSeq Gene annotation database (build hg19) through the R package `AnnotationHub` (Morgan et al., 2016) (record number AH5122 and AH5040, respectively).

b. Model structure

We proposed to predict the methylation levels of the target CpGs in RE using neighboring profiled CpGs within a flanking window (**Figure 8**).

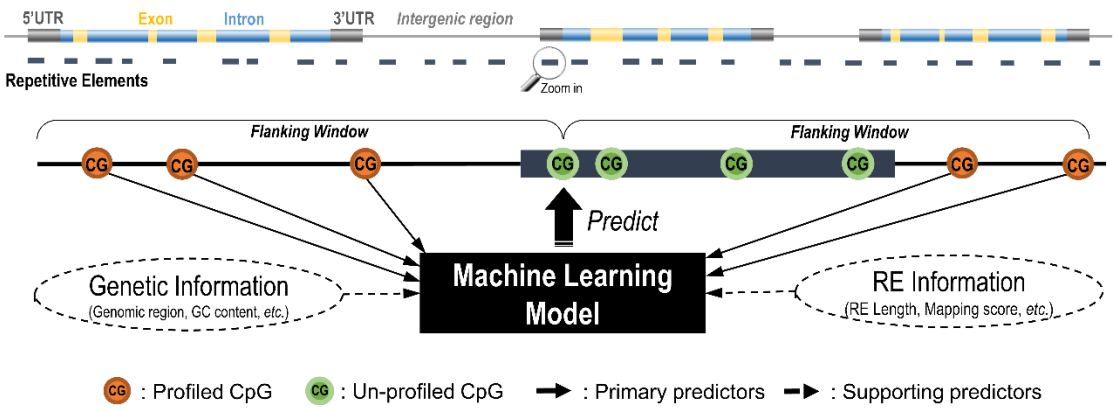

Figure 8 Diagram of the RE methylation prediction algorithm.

For each un-profiled CpGs identified within a RE sequence, the neighboring profiled CpGs are identified within a given flanking window, where the primary and supporting predictors are collected. Those profiled CpGs in RE with sufficient neighboring information are included as a set for model training whereas CpGs not profiled in RE will be predicted using the trained model.

Within the flanking window of target RE CpGs with at least two neighboring profiled CpGs were considered to improve prediction reliability. Based on previous work in predictor prioritization (Zhang et al., 2015) and our extensive experiments in selecting contributive predictors, we constructed the following primary predictors for each target CpG:

- Methylation level of the closest and second-closest profiled CpGs in the flanking region of the target CpG.

- Genomic distance in base pair (bp) from the closest and second-closest profiled CpGs to the target CpG.
- Mean and variance of methylation levels at all neighboring profiled CpGs.
- Mean and variance of genomic distance between all neighboring profiled CpGs and the target CpG.

We also constructed the following supporting predictors to better model local genomic characteristics of the target CpGs and their relationships with RE methylation:

- RE CpG density: CpG density is correlated with DNA methylation across various tissues (Meissner et al., 2008, Eckhardt et al., 2006). For CpGs in RE, methylation level showed a reverse U-shaped relationship with increasing CpG density (Edwards et al., 2010). We defined RE CpG density as the number of CpGs within RE divided by the length of RE.
- RE length: Full-length RE sequences tend to be more active, usually representing more recently-evolved elements (particularly for LINE-1) (Rangwala et al., 2009). Increasing DNA methylation has been shown to correlate with younger evolutionary age of RE (Price et al., 2012).
- Smith-Waterman (SW) score: The RepeatMasker database employed a SW alignment algorithm (Smith and Waterman, 1981) to computationally identify Alu and LINE-1 sequences in the reference genome. A higher score indicates fewer insertions and

deletions in query RE sequences compared to consensus RE sequences. We included this factor to account for potential bias induced by SW alignment.

- Number of neighboring profiled CpGs: More neighboring CpG profiles results in more reliable and informative primary predictors. We included this predictor to account for potential bias due to profiling platform design.

- Genomic region of the target CpG: It is well-known that methylation levels differ by genomic regions. Our algorithm included a set of seven indicator variables for genomic region (as annotated by RefSeq Gene) including: 2000 bp upstream of transcript start site (TSS2000), 5'UTR (untranslated region), coding DNA sequence (CDS), exon, 3'UTR, protein-coding gene, and noncoding RNA gene. Note that intron and intergenic regions can be inferred by the combinations of these indicator variables.

c. Random Forest vs Support Vector Machine

For a given flanking window size, we generated these predictors and trained a model to predict methylation levels of the target CpGs. We considered the following approaches:

- Naïve method: This approach takes the methylation level of the closest neighboring CpG profiled by HM450 or EPIC as that of the target CpG. We treated this method as our 'control'.

- Support Vector Machine (SVM) (Cortes and Vapnik, 1995): SVM is supervised learning technique and best known for its application in classification problem. It has been extensively used for predicting methylation status (methylated vs. unmethylated) (Zheng

et al., 2013, James et al., 2013, Fan et al., 2008, Bock et al., 2007, Fang et al., 2006, Bock et al., 2006). The basic idea for SVM is to determine a hyperplane that can classify data into two categories and achieve minimal classification error. The power of SVM is greatly boosted and applicable to non-linear classification with "kernel function", which is to project the original data to a higher dimension such that a hyperplane is available to classify the projected data. SVM can be also extended to regression (Support Vector Regression (SVR)) to solve prediction problems. SVR shares the same principles as SVM for classification with only minor modifications. In this study, since the methylation β value is continuous, we employed SVR and considered two different kernel functions to determine the underlying SVM architecture: the linear kernel and the radial basis function (RBF) kernel (Vert et al., 2004), as there is no definite way to determine a "best" kernel function.

- Random Forest (RF) (Breiman, 2001): RF is a supervised learning technique. It is an ensemble classifier consisting of classification and regression tree-structured classifiers. Briefly, each tree is trained on two-thirds of the random samples in the original dataset, leaving about the remaining one-third as the "out-of-bag" testing samples. When growing a tree, a randomly selected subset of predictors is used to split at each node. The number of predictors randomly selected at each split was one-third of the total number of predictors. Since the methylation β value is a continuous variable, the predictions were made by averaging the prediction of regression trees. RF is a competitor of SVM. RF recently demonstrated superior performance over other machine learning models in predicting methylation levels (Zhang et al., 2015).

A 3-time repeated 5-fold cross validation was performed to determine the best model parameters for SVM and RF using the R package `caret` (Kuhn, 2008). The search grid was Cost = (2^{-15}, 2^{-13}, 2^{-11}, …, 2^{3}) for the parameter in linear SVM, Cost = (2^{-7}, 2^{-5}, 2^{-3}, …, 2^{7}) and γ = (2^{-9}, 2^{-7}, 2^{-5}, …, 2^{1}) for the parameters in RBF SVM, and the number of predictors sampled for splitting at each node (3, 6, and 12) for the parameter in RF.

d. Prediction quality control

We also evaluated and controlled the prediction reliability when performing model extrapolation out of training data. Quantifying prediction reliability in SVM is challenging and computationally intensive (Jiang et al., 2008). In contrast, prediction reliability can be readily inferred by Quantile Regression Forests (QRF) (Meinshausen, 2006) (available in the R package `quantregForest` (Meinshausen, 2016)). Briefly, by taking advantage of the established random trees, QRF estimates the full conditional distribution for each of the predicted values. We therefore defined prediction error using the standard deviation (SD) of this conditional distribution to reflect variation in the predicted values. Less reliable RF predictions (results with greater prediction error) can be trimmed off (RF-Trim).

CHAPTER 4 PREDICTION PERFORMANCE

I. Materials and Methods

The performance of prediction model was tested using GM12878 sample. There are extensive publicly-accessible methylation data on GM12878, making it an ideal sample for model development and validation. The HM450 data, RRBS, WGBS data on GM12878 were downloaded from the ENCODE (The Encyclopedia of DNA Elements) (Encode Project Consortium, 2012); the EPIC data were the means of three technical replicates of GM12878 obtained from R package `minfiDataEPIC` (Fortin and Hansen, 2016). For HM450 and EPIC data, methylation level is expressed as β value, which is the proportion of methylated intensity out of the total intensity (methylated + unmethylated intensity + a constant offset (by default, offset = 100)). β value has significant heteroscedasticity in the low and high methylation range, which can be effectively resolved by logit transformation of β value to M value, i.e., $\log_2\left(\frac{\beta}{1-\beta}\right)$ (Du et al., 2010). We used M value for RE methylation prediction and transformed the predicted methylation values to β value for external validation. The NimbleGen SeqCap Epi 4M CpGiant (NimbleGen) (Duhaime-Ross, 2014) profiling data are courtesy of Roche Sequencing. Raw NimbleGen sequencing data processing followed the manufacturer's recommended workflow (Roche Diagnostics, 2014). For NimleGen, RRBS, and WGBS the processed BAM files of two replicates were united into a single dataset using R package `methylKit` (Akalin et al., 2012). The ratio of methylated read counts (i.e., count of cytosine) to sequencing depth (i.e., count of cytosine + thymine) was calculated to represent methylation level, which shares the similar

definition of β value in HM450 and EPIC data. CpG sites with greater than 30× sequencing depth were retained.

To evaluate and compare the predictive performance of different models, an external validation study was conducted. Alu and LINE-1 were prioritized for demonstration due to their high abundance throughout the genome as well as their biological relevance. HM450 was selected as the primary platform for evaluation. Model performance was traced using incremental window sizes from 200 to 2,000 bp for Alu and LINE-1 and employed two evaluation metrics: Pearson's correlation coefficient (*r*) and root mean square error (RMSE) between predicted and profiled CpG methylation levels. Predicted RE methylation using the HM450 and EPIC were validated by NimbleGen. To account for evaluation bias (caused by the inherent variation between the HM450/EPIC and the sequencing platforms), "benchmark" evaluation metrics (*r* and RMSE) between both types of platforms were calculated using the common CpGs profiled in Alu/LINE-1 as the best theoretically possible performance the algorithm could achieve. Since the EPIC covers twice as many CpGs in Alu/LINE-1 as the HM450 (**Table 4**), EPIC was also used to validate the HM450 prediction results.

II. Model performance of predicting methylation in Alu and LINE-1

Validation results showed that RF had the best prediction performances. After trimming off less reliable predictions (RF-Trim, error ≤1.7), it achieved higher correlations and lower errors that approached the best theoretically possible performance. As window size increased above 1000

bp, prediction performances for Alu declined (**Figure 9A**) and the number of reliable predictions for LINE-1 leveled off (**Figure 9B**). These observations were consistent with the previous findings that two nearby CpG sites within 1000 bp are more likely to be co-methylated (Bell et al., 2011, Eckhardt et al., 2006, Zhang et al., 2015, Edwards et al., 2010, Moen et al., 2013).

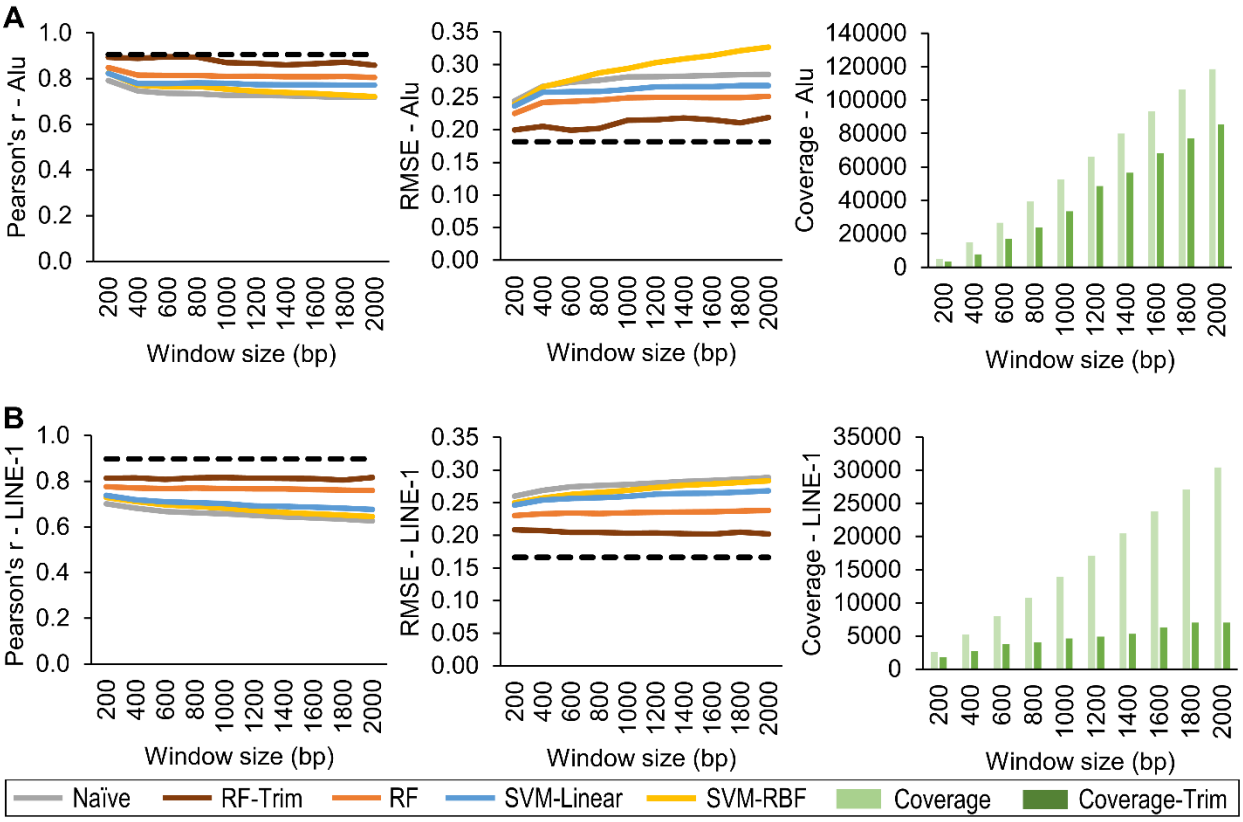

Figure 9 Performance of RE methylation prediction algorithm in different prediction models.

Comparison of correlation and RMSE between measured (NimbleGen) and predicted (based on HM450) values for five prediction models (Naïve, RF, RF-Trim, SVM-Linear, and SVM-RBF) relative to the best theoretically possible performance (dashed line). RF-Trim achieved the best performance for both Alu (A) and LINE-1 (B) and approach to the best theoretical level. Compared with RF, RF-Trim removed more unreliable predictions, leading to less coverage but superior performance. RF: random forest; SVM-Linear: support vector machine with linear kernel; SVM-RBF: support vector machine with radial basis function kernel.

We observed similar prediction performance using the EPIC (**Figure 10**).

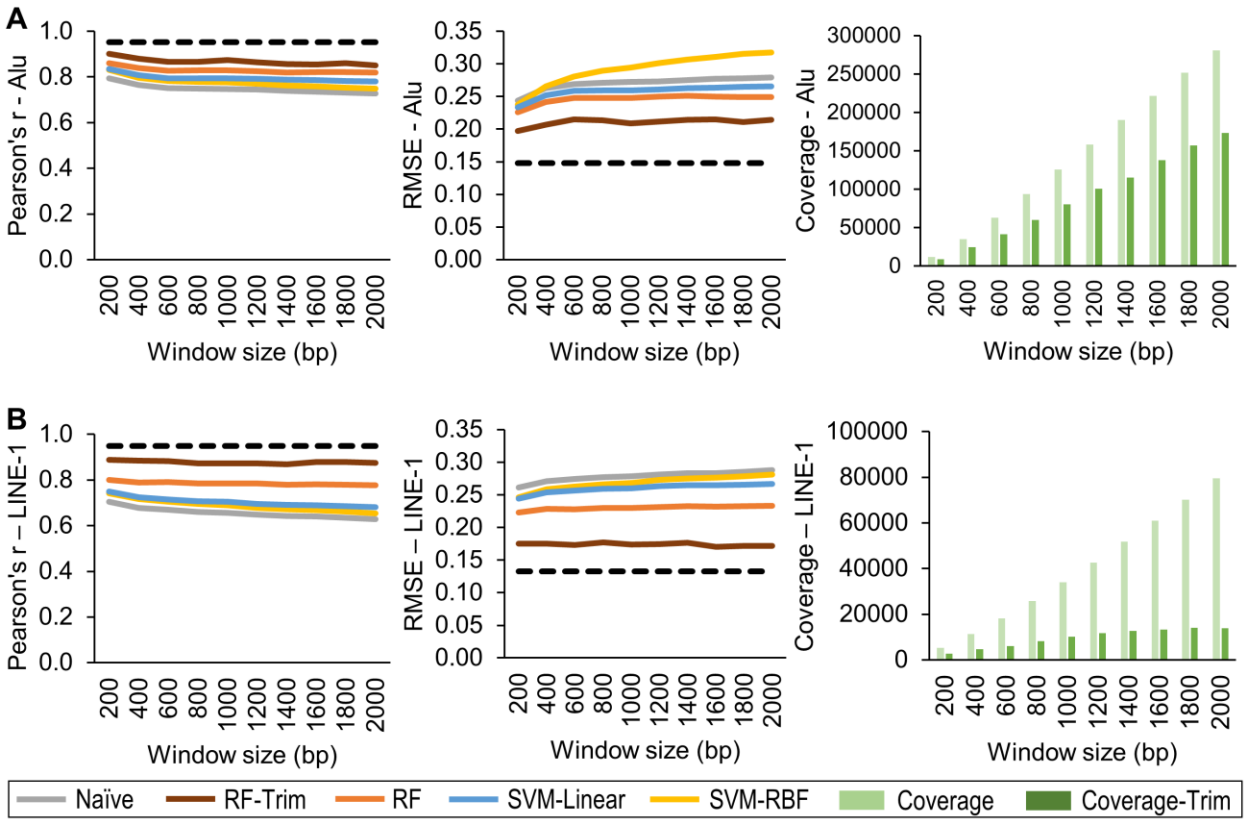

Figure 10 Performance of RE methylation prediction algorithm using EPIC data across different prediction models.

A: Alu. B: LINE-1. We conducted the predictions using EPIC data and validated using NimbleGen.

We further validated the HM450 predicted results using the EPIC. RF-Trim (error ≤1.7) achieved the highest accuracy with Person's correlation coefficient (r) = 0.86 and 0.89 and root mean square error (RMSE) = 0.12 and 0.12 for Alu (**Figure 11**) and LINE-1 (**Figure 12**), respectively. The cutoff of 1.7 for prediction error in RF-Trim is empirical, to balance the

tradeoff between coverage and accuracy (i.e., more stringent prediction error threshold led to higher accuracy but lower Alu/LINE-1 coverage).

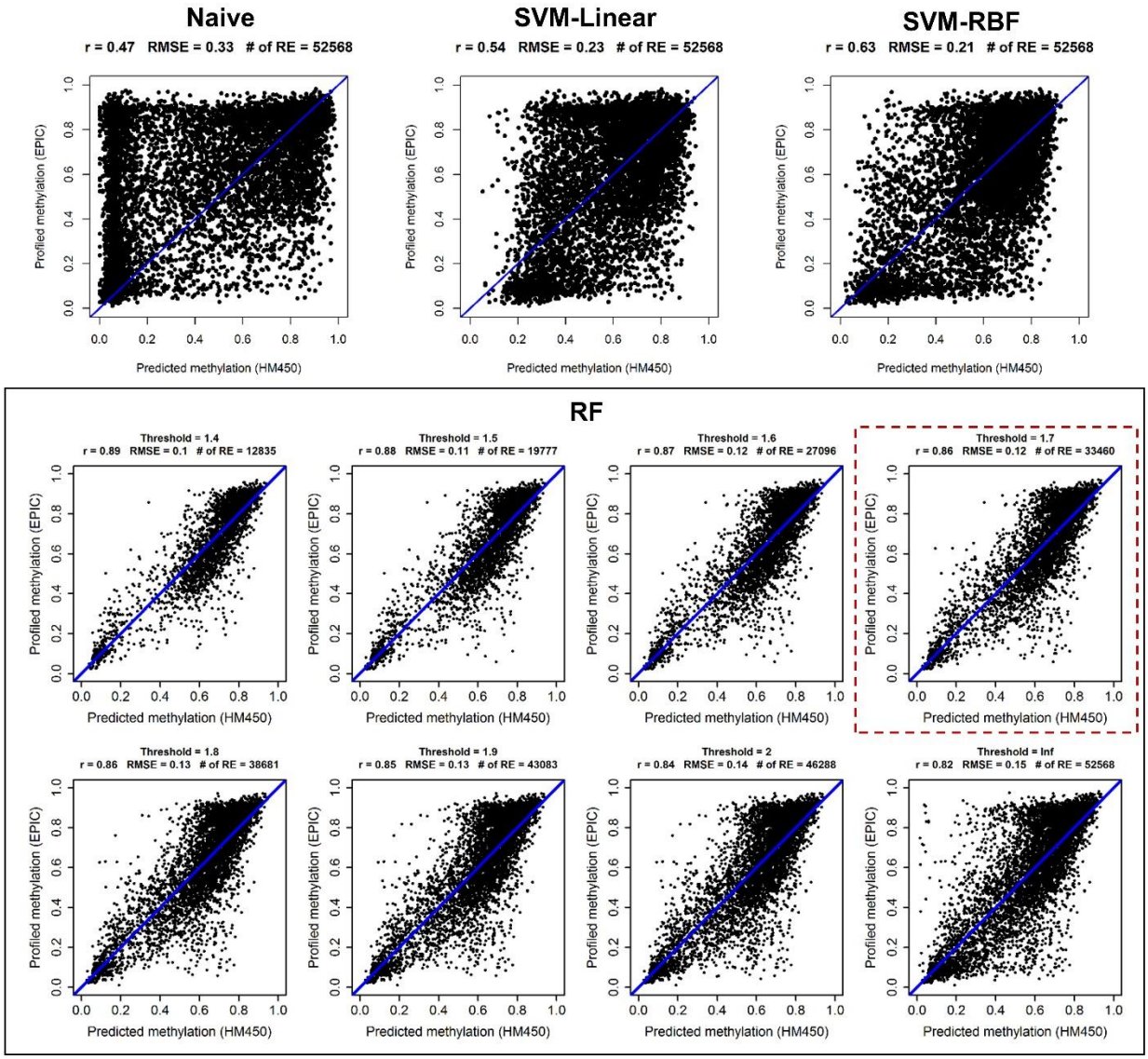

Figure 11 Performance of HM450-based prediction validated using EPIC data on GM12878 (Alu).

For RF we applied different cutoffs, ranging from 1.4 to 2.0 as well as no cutoff (infinity), to control prediction quality. More stringent (i.e., smaller) prediction error thresholds led to higher accuracy but lower Alu coverage.

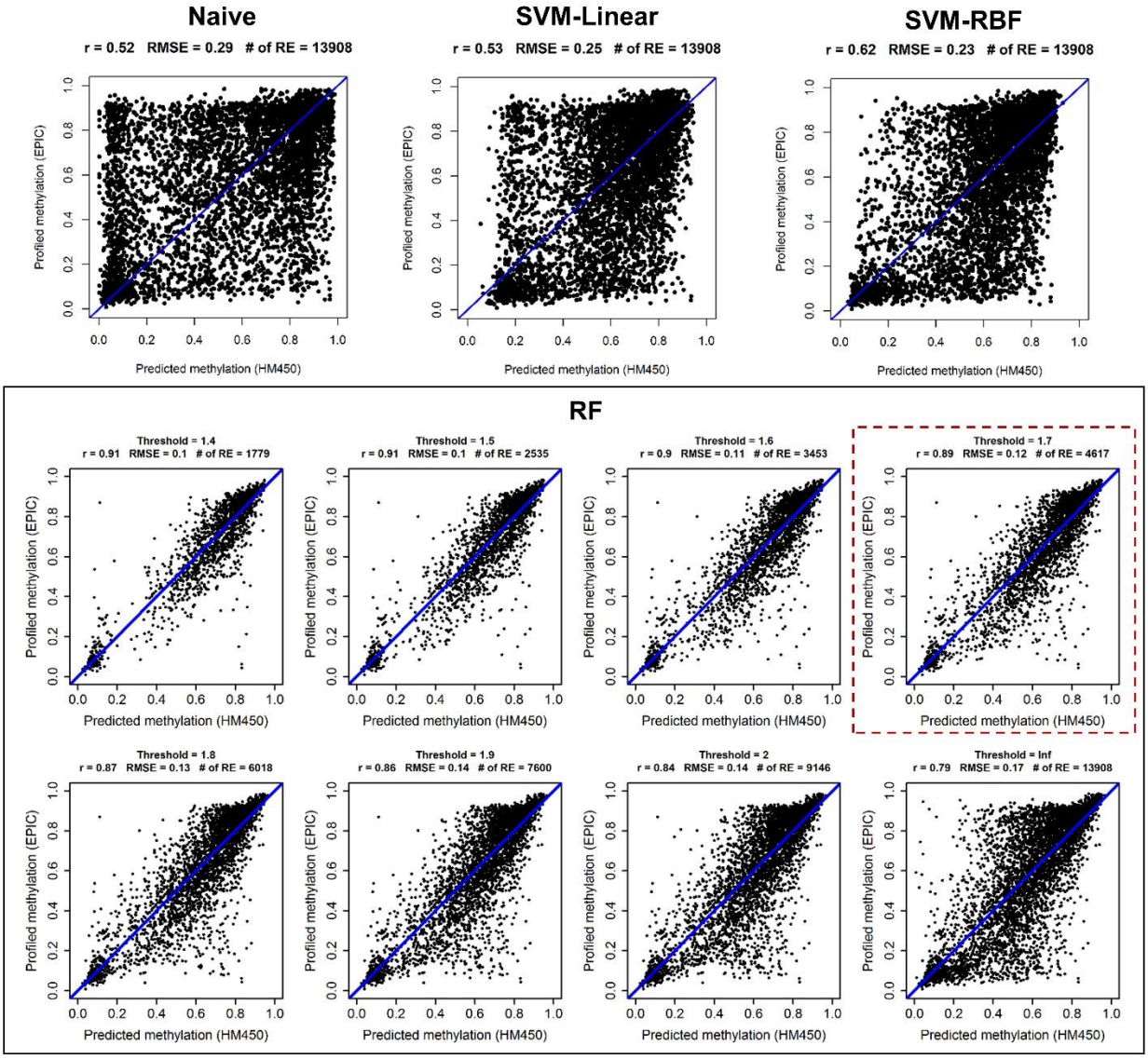

Figure 12 Performance of HM450-based prediction validated using EPIC data on GM12878 (LINE-1).

For RF we applied different cutoffs, ranging from 1.4 to 2.0 as well as no cutoff (infinity), to control prediction quality. More stringent (i.e., smaller) prediction error thresholds led to higher accuracy but lower LINE-1 coverage.

Taken altogether, RF-Trim with a 1000 bp window is our preferred method as it offers more accurate prediction and enables prediction quality control.

Compared with the profiled Alu/LINE-1 methylation using the HM450/EPIC, our algorithm predicted 2.7-3.7 times as many Alu and about 20% more LINE-1; predictions based on the EPIC yielded nearly 2-3 times as many Alu/LINE-1 coverage than those based on the HM450 (**Figure 13A**). Moreover, our algorithm improved the CpG density in Alu/LINE-1. For example, using the HM450, each Alu contained 6.1 reliable predicted CpGs and each LINE-1 contained 5.0 reliable CpGs predicted, both 5-6 times higher than the HM450 and comparable with the average CpG density calculated in the full RE database (**Figure 13B**).

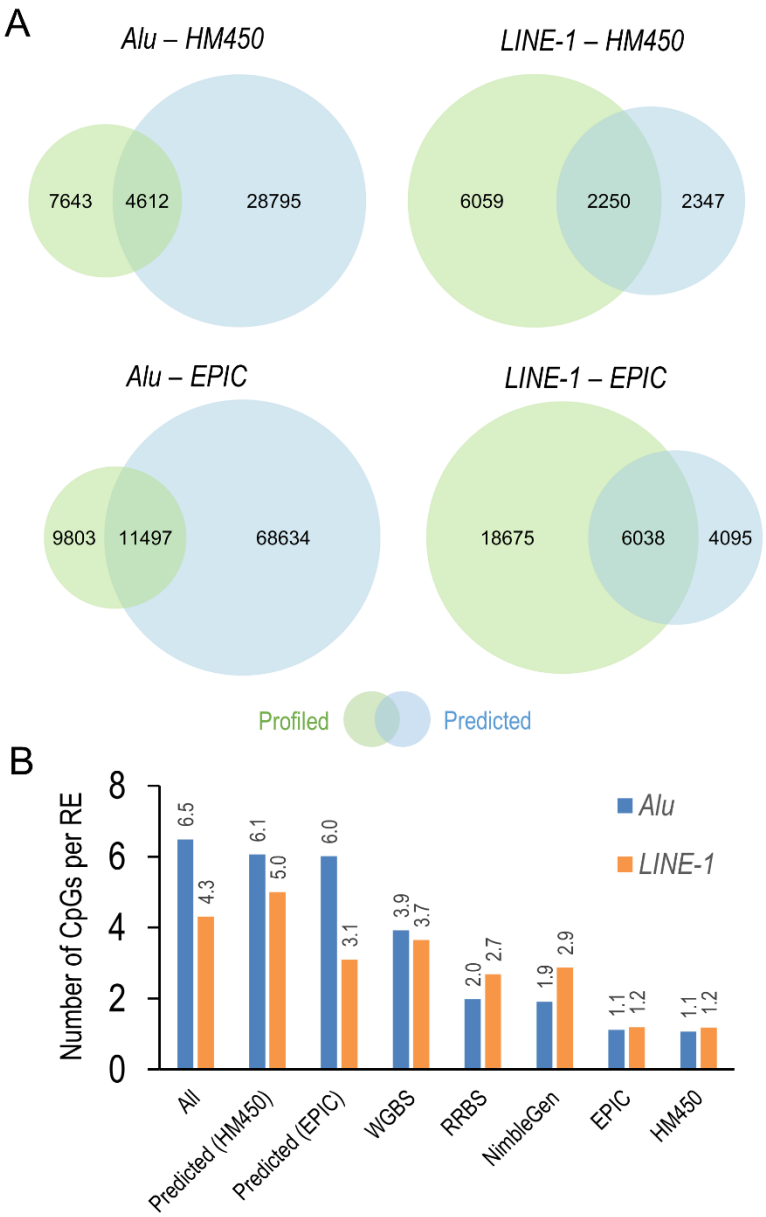

Figure 13 Comparisons of Alu and LINE-1 coverage and CpG density using the prediction algorithm vs. profiling platforms.

A: Alu and LINE-1 actual vs. predicted coverage based on HM450 and EPIC. B: Density of CpGs interrogated per Alu and LINE-1 locus of predicted vs. profiled values. The prediction algorithm enhanced CpG density by 5-6 fold, more comparable with the natural level of Alu/LINE-1 methylation.

III. Proof of concept: methylation and evolutionary age of Alu and LINE-1

A proof-of-concept study was designed to test whether predicted Alu/LINE-1 methylation can correlate with the evolutionary ages of Alu/LINE-1 from the HapMap LCL GM12878 sample. The evolutionary age of Alu/LINE-1 is inferred from the divergence of copies from the consensus sequence as new base substitutions, insertions, or deletions accumulate in Alu/LINE-1 through the "copy and paste" retrotransposition activity (**Figure 2B**). Older Alu/LINE-1 copies are in general inactive since more mutations were induced (partially by CpG methylation). Younger Alu/LINE-1, especially currently active ones, have fewer mutations and thus CpG methylation is a more important defense mechanism for suppressing retrotransposition activity. Therefore, it is expected that DNA methylation level to be lower in older Alu/LINE-1 than in younger Alu/LINE-1. We calculated and compared the average methylation level across three evolutionary subfamilies in Alu (ranked from young to old): AluY, AluS, and AluJ, and five evolutionary subfamilies in LINE-1 (ranked from young to old): L1Hs, L1P1, L1P2, L1P3, and L1P4. We tested trends in average methylation level across evolutionary age groups using linear regression models.

Using GM12878 data, we observed that HM450 predicted methylation level as associated with an inverse dose-response relationship with evolutionary age, indicating the defensive role of DNA methylation in RE (**Figure 14**).

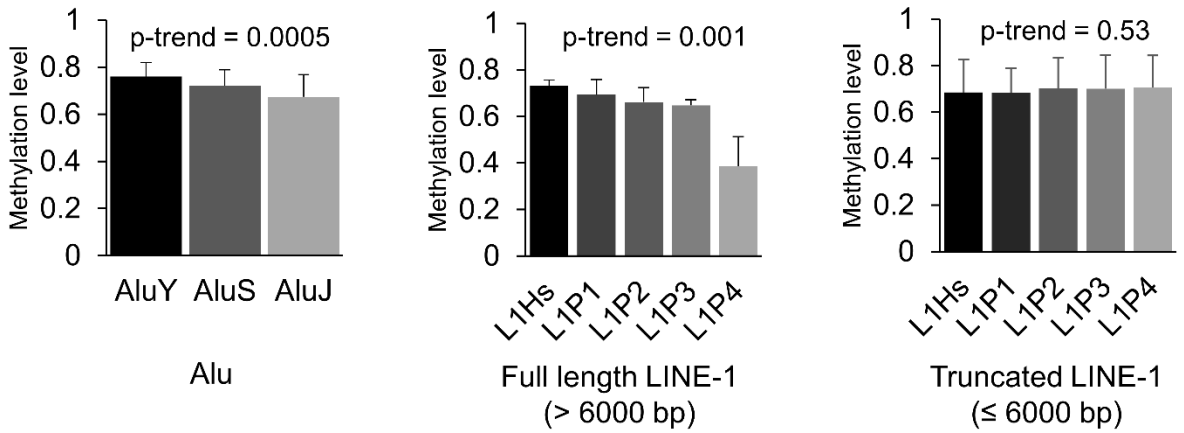

Figure 14 Inverse relationship between evolutionary ages of Alu and LINE-1 and mean methylation level based on predicted values (HM450-based prediction).

The analysis includes three evolutionary subfamilies in Alu, from young to old: AluY, AluS, and AluJ, and five evolutionary subfamilies in LINE-1, from young to old: L1Hs, L1P1, L1P2, L1P3, and L1P4. The histograms and error bars represent the average and standard deviation of methylation level, respectively.

A similar relationship was evident among Alu and full-length (>6,000 bp) LINE-1 but not truncated LINE-1; we found similar relationships using EPIC predicted values as well (**Figure 15**).

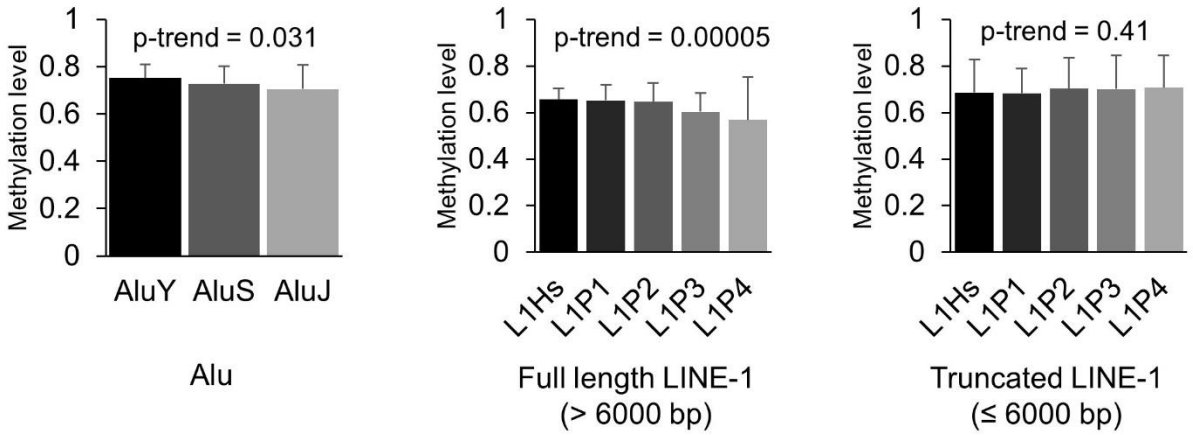

Figure 15 Inverse relationship between evolutionary ages of Alu and LINE-1 and mean methylation level (EPIC-based prediction).

The analysis includes three evolutionary subfamilies in Alu, from young to old: AluY, AluS, and AluJ, and five evolutionary subfamilies in LINE-1, from young to old: L1Hs, L1P1, L1P2, L1P3, and L1P4. The histograms and error bars represent the average and standard deviation of each methylation subfamily, respectively.

CHAPTER 5 APPLICATION TO CLINICAL SAMPLES IN TCGA

I. Materials and Methods

For algorithm application to clinical samples, we used The Cancer Genome Atlas (TCGA) database. Four common types of cancer in the US (American Cancer Society, 2015) were selected: Breast invasive carcinoma (BRCA, 90 tumor samples), Prostate adenocarcinoma (PRAD, 50 tumor samples), Lung squamous cell carcinoma (LUSC, 40 tumor samples), and Colon and rectal adenocarcinoma (COAD, 38 tumor samples). Sample inclusion criteria include 1) primary tumor tissue and 2) paired normal solid tissue collected from the same individual. Processed and normalized (level 3) HM450 methylation data and RNA-Seq gene expression data were downloaded from the TCGA open-access database using the R package `TCGAbiolinks` (Colaprico et al., 2016). We used M value for RE methylation prediction and differential methylation analysis for better statistical property.

To demonstrate our algorithm's utility, the following analysis is to investigate:

a) Differentially methylated RE in tumor vs normal tissue and their biological implications; and

b) Tumor discrimination ability using global methylation surrogates (i.e., mean Alu and LINE-1) vs the predicted locus-specific RE methylation.

To best utilize data, these analyses were conducted using the union set of the HM450 profiled and predicted CpGs in Alu/LINE-1, defined here as the extended CpGs.

For a), differentially methylated CpGs in Alu and LINE-1 between tumor and paired normal tissues were identified via paired t-tests (R package `limma` (Ritchie et al., 2015)). Tested CpGs were grouped and identified as differentially methylated regions (DMR) using R package `Bumphunter` (Jaffe et al., 2012) and family wise error rates (FWER) estimated from bootstraps to account for multiple comparisons. Regulatory element enrichment analyses were conducted to test for functional enrichment of significant DMR. The analysis includes DNase I hypersensitivity sites (DNase), transcription factor binding sites (TFBS), and annotations of histone modification ChIP (chromatin immunoprecipitation) peaks pooled across cell lines (data available in the ENCODE Analysis Hub at the European Bioinformatics Institute). For each regulatory element, the number of overlapping regions amongst the significant DMR (observed) and 10,000 permuted sets of DMR markers (expected) were calculated. The ratio of observed to mean expected was used as the enrichment fold and obtained an empirical p-value from the distribution of expected. Focusing on gene regions, KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway enrichment analysis were conducted using hypergeometric tests via the R package `clusterProfiler` (Yu et al., 2012). To minimize bias in our enrichment test, the background genes were set as genes targeted by all bumps tested. False discovery rate (FDR) <0.05 was considered significant in both enrichment analyses.

For b), conditional logistic regression was employed with elastic net penalties (R package `clogitL1`) (Reid and Tibshirani, 2014) to select locus-specific Alu and LINE-1 methylation for discriminating tumor and normal tissue. Missing methylation data due to insufficient data quality were imputed using K Nearest Neighbor (KNN) imputation (Troyanskaya et al., 2001). Tuning parameter was set to $\alpha = 0.5$ and λ was tuned via 10-fold cross validation. To account for overfitting, 50% of the data were randomly selected to serve as the training dataset with the remaining 50% as the testing dataset. A classifier was constructed by using the selected Alu and LINE-1 to refit the conditional logistic regression model, and another using the mean of all Alu and LINE-1 methylation as a surrogate of global methylation. Finally, using R package `pROC` (Robin et al., 2011), receiver operating characteristic (ROC) analysis was performed and area under the ROC curves (AUC) was used to compare the performance of each discrimination method in the testing dataset via DeLong tests (DeLong et al., 1988).

II. Differential methylation analyses of locus-specific Alu and LINE-1

Using RF-Trim, we predicted about 37,000 Alu and 8,000 LINE-1 across the genome in TCGA samples (**Table 5**).

Table 5 Coverage of predicted Alu/LINE-1 using TCGA data.

	# of RE	# of RE CpGs	# of genes covered [a]	# of RE subfamilies covered
		Alu		
Breast cancer (BRCA)	38848	235533 [b]	22924	41
Lung cancer (LUSC)	37647	225367 [b]	22786	41
Colon cancer (COAD)	34605	209313 [b]	21046	41
Prostate cancer (PRAD)	36224	219007 [b]	21533	41
		LINE-1		
Breast cancer (BRCA)	9174	44185 [b]	6897	116
Lung cancer (LUSC)	8928	43308 [b]	6824	115
Colon cancer (COAD)	6768	34595 [b]	5388	113
Prostate cancer (PRAD)	7715	37096 [b]	6021	115

[a] RefSeq genes, including gene proximal promoter region (i.e., 2000 bp upstream of the transcription start site).
[b] CpG sites with reliable prediction across more than 80% of the samples were retained.

Most Alu and LINE-1 loci showed a unimodal distribution centered at a high methylation level (β ~0.9) in both tumor and paired normal tissues, but was relatively lower and more widely dispersed in tumors (**Figure 16**).

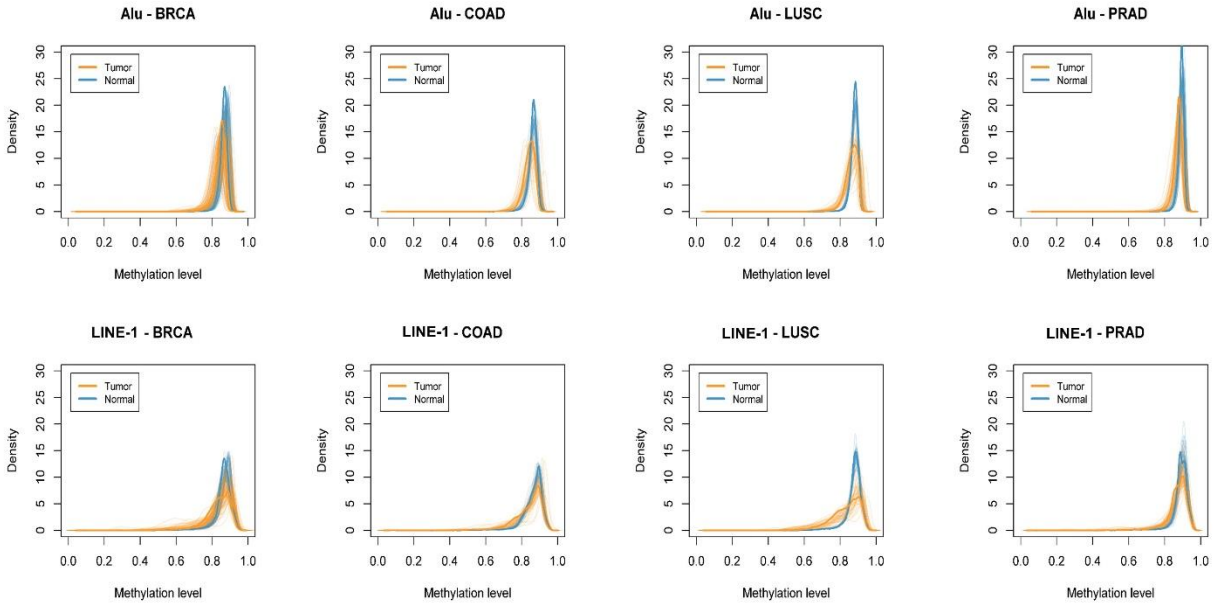

Figure 16 Density plots of predicted Alu and LINE-1 using TCGA data.

Methylation among four tumor types in TCGA Alu and LINE-1 loci showed a unimodal distribution centered at a high methylation level (β ~0.9) in both tumor and paired normal tissues, but was relatively lower and more widely dispersed in tumors. Light-colored lines denote the density curves of individual samples and bolded lines the density curves of mean methylation levels across all samples.

On average, around 77,000 extended (i.e., union set of profiled and predicted) CpGs (98%) in Alu and 15,000 (90%) in LINE-1 were hypomethylated across all four types of tumor tissues, with a general overall trend towards global hypomethylation (exemplified by breast cancer, **Figure 17A**). In contrast, using only the profiled CpGs we found that roughly 2,500 (~88% of profiled CpGs) in Alu or LINE-1 were hypomethylated. Regional analysis was conducted to summarize significant DMR (FWER <0.05) in Alu/LINE-1 using extended CpGs and compared the results using profiled CpGs. The genomic distribution of all Alu/LINE-1 CpGs, all identified

bumps, and significant DMR had similar proportions observed using both profiled and extended CpGs (exemplified by breast cancer, **Figure 17B**).

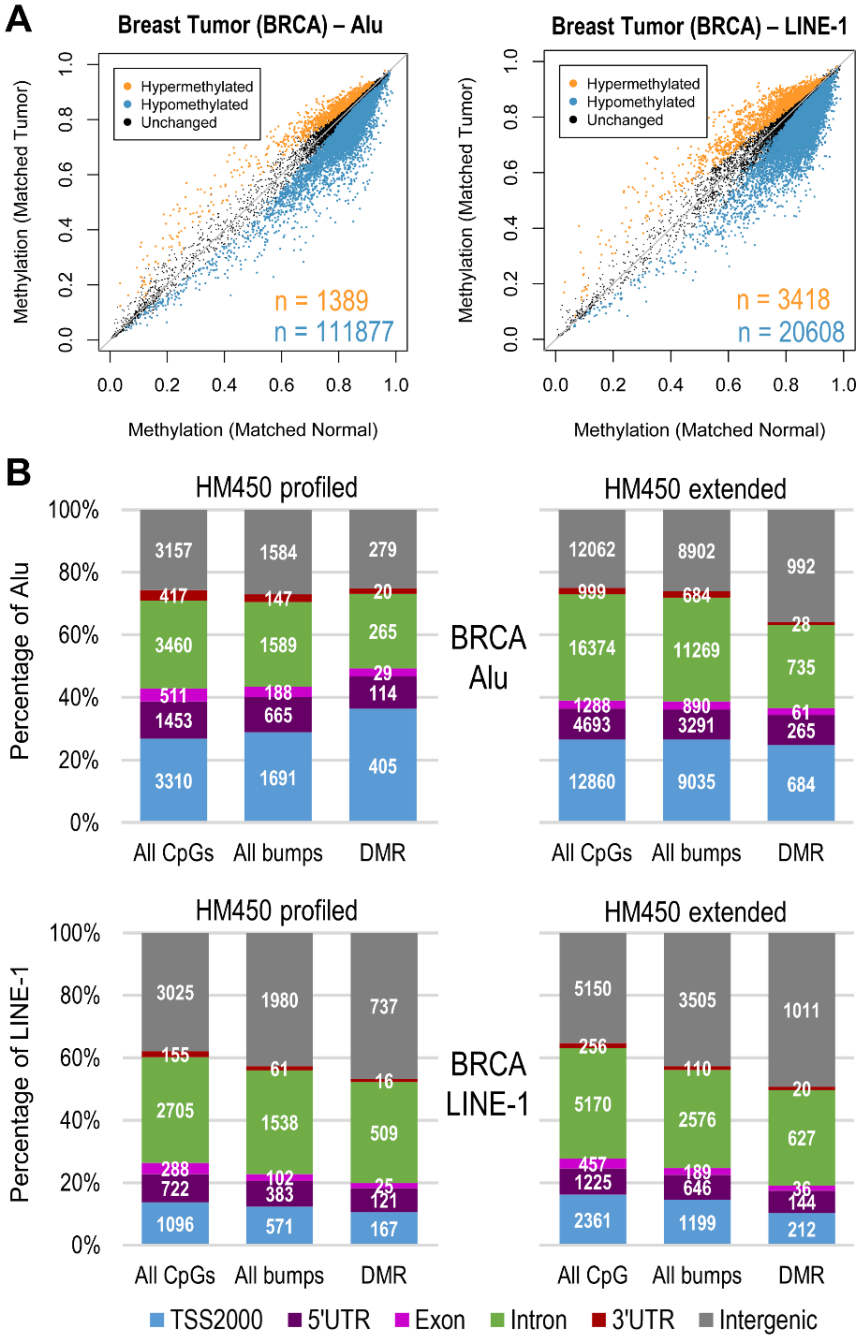

Figure 17 Differentially methylated CpGs/regions in Alu and LINE-1 (Breast tumor).

A: Scatter plot comparing extended CpGs in Alu and LINE-1 between breast tumor and matched normal tissue; significant differences at Bonferroni p <0.05 are colored and n is the number of CpGs (orange: hypermethylated; blue: hypomethylated). B: Genome-wide break down of all CpGs tested, bumps formed using bumphunter, and significant DMR (FWER <0.05) identified in breast cancer. Genomic distribution of extended CpGs was similar to profiled and identified more DMR of interest, especially in the intron and intergenic regions.

Therefore, it is unlikely that the prediction introduces any artificial bias towards specific genomic regions. Furthermore, due to the higher density of the predicted CpGs in Alu/LINE-1 there were more bumps detected using the extended CpGs compared to the profiled CpGs, particularly in Alu. Similarly compared to the profiled CpGs, the extended CpGs yielded nearly twice as many Alu/LINE-1 with significant DMR, especially in the intron and intergenic regions.

The remaining three cancer types yielded similar results as breast cancer (Scatter plot: **Figure 18**; Genomic distribution: **Figure 19**).

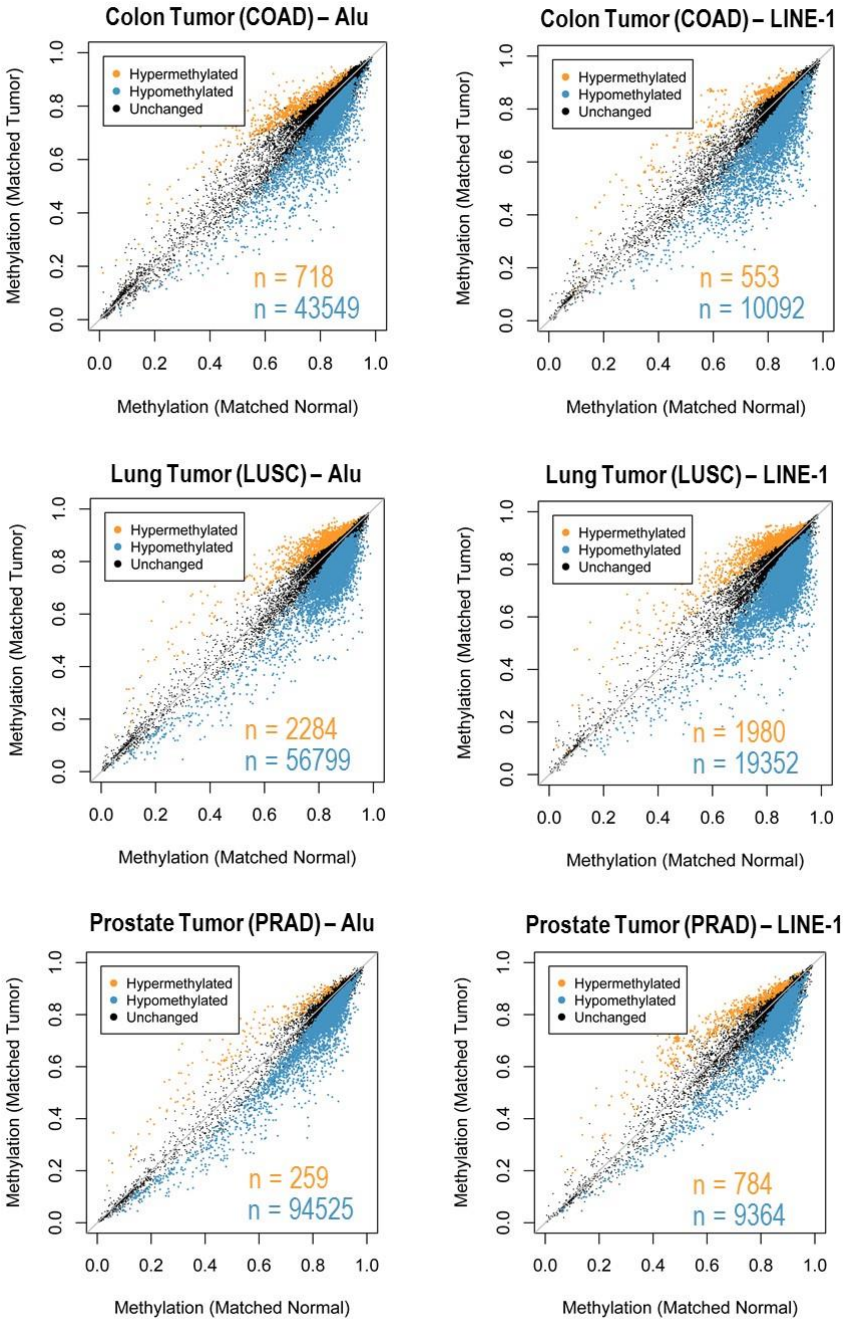

Figure 18 Differentially methylated CpGs predicted in Alu and LINE-1 (Colon, Lung, and Prostate tumor).

Scatter plot comparing extended CpGs in Alu and LINE-1 between tumor and matched normal tissue. Significant differences at Bonferroni $p < 0.05$ are colored and n is the number of CpGs (orange: hypermethylated; blue: hypomethylated).

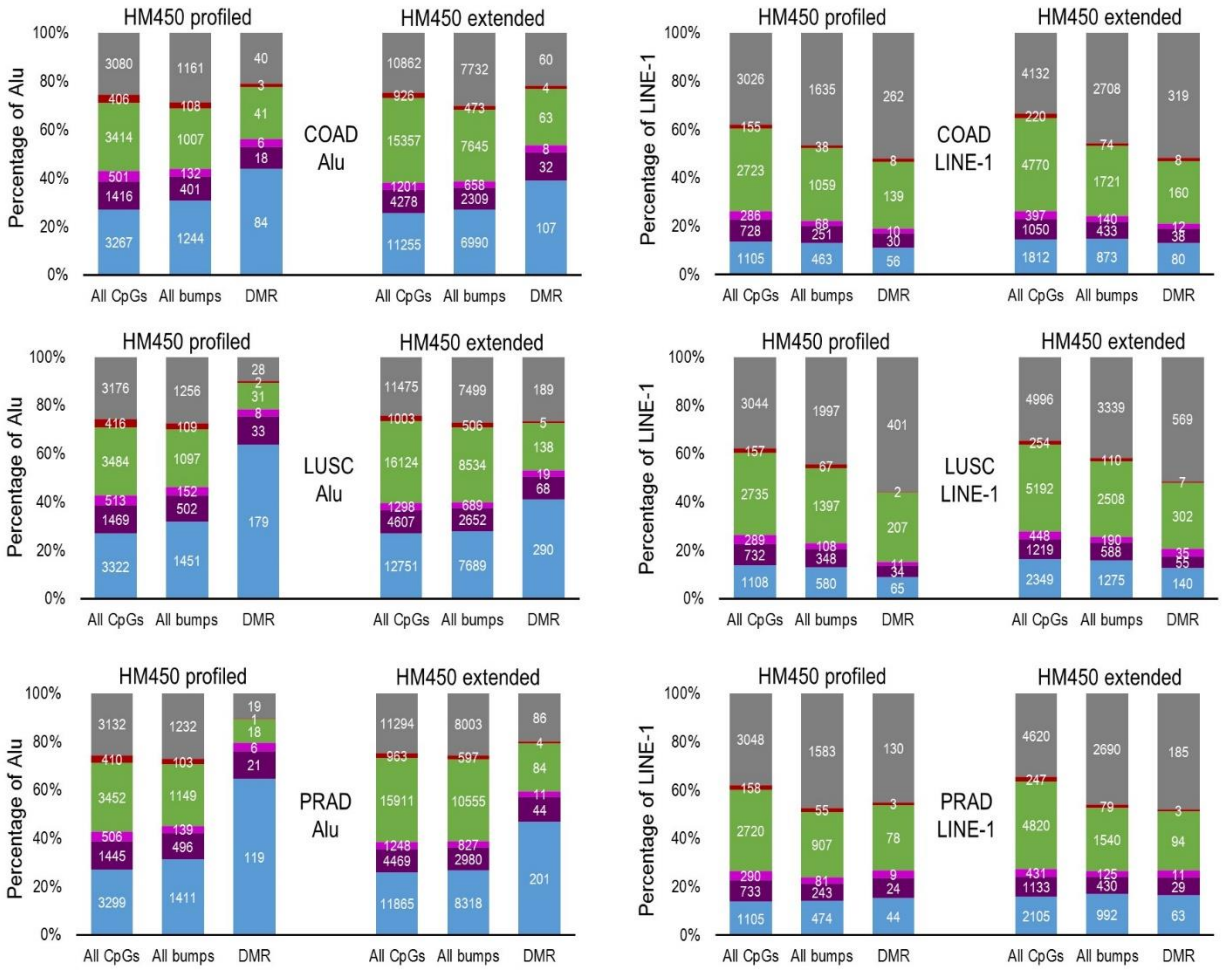

Figure 19 Genome-wide breakdown of all CpGs tested, bumps formed using bumphunter, and significant DMR (Colon, Lung, and Prostate tumor).

Genomic distribution of extended was similar to profiled CpGs and identified more DMR of interest (FWER <0.05), especially in the intron and intergenic regions.

To explore the functional insights of locus-specific RE methylation in tumor tissue, we conducted the regulatory elements and KEGG enrichment analyses based on the significant hypo- and hyper-methylated Alu/LINE-1 DMR from the extended CpGs. Due to the limited

number of hypermethylated DMR, only hypomethylated DMR yielded significant results. The enrichment can be found in regulatory elements including TFBS; active chromatin markers including DNase, H2A.Z, and H3K4me3; and repressive chromatin markers such as H3K9me3 and H3K27me3. We found no enrichment found in the remaining active chromatin marks (H3K4me1, H3K9ac, H3K27ac, H3K36me3, H3K79me2, and H4K20me1) (**Figure 20A**). Common pathways across the four cancers were identified including olfactory transduction and axon guidance (**Figure 20B**).

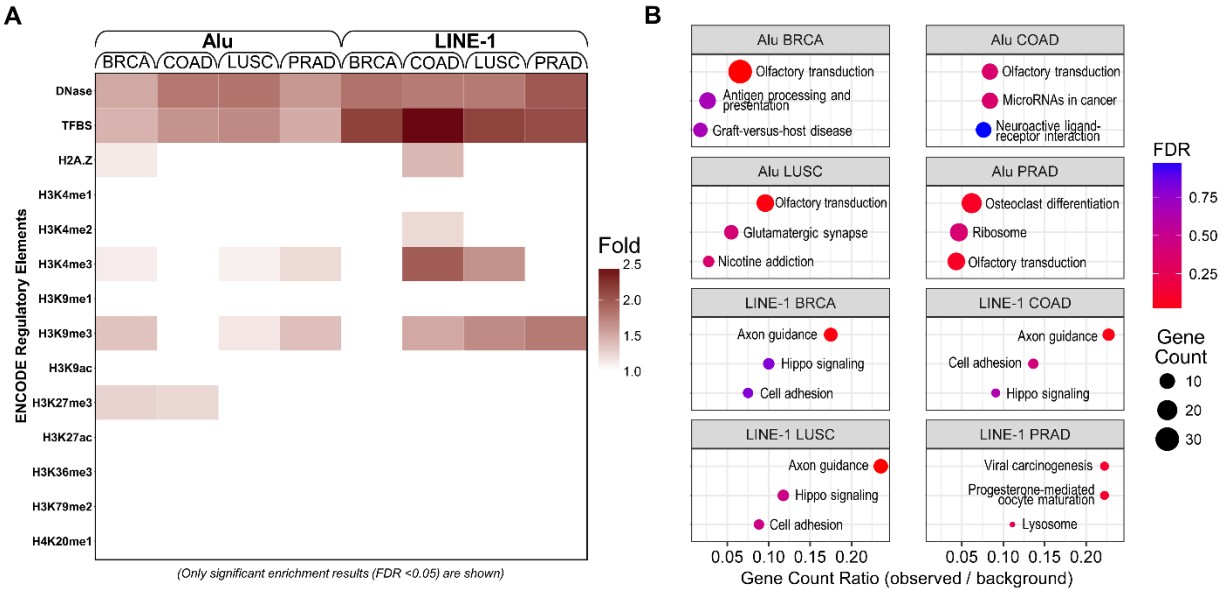

Figure 20 Regulatory element enrichment analysis and KEGG pathway enrichment analysis using significant hypomethylated Alu/LINE-1 DMR.

Significant enrichment indicates Alu and LINE-1 DMR that are more likely to appear in regulatory elements (a) or KEGG pathways (b). Hypomethylation of Alu and LINE-1 may involve in *cis*-regulatory changes and potential transcription activation events in cancer-related pathways in tumor tissues. DNase: DNase I hypersensitivity sites; TFBS: transcription factor binding sites; H2A.Z: histone H2A variant; H3: histone H3; H4: histone H4; K: lysine; me1: monomethylation; me2: demethylation; me3: trimethylation; ac: acetylation.

Higher enrichment fold in regulatory elements analysis and gene count ratio in KEGG analysis were observed in LINE-1 than in Alu, indicating a more active functional role for LINE-1 hypomethylation. Two full-length LINE-1 loci in the introns of *SEMA3A* (Semaphorin 3A) (a gene in the axon guidance pathway) were hypomethylated in breast, colon, and lung tumor tissues (**Figure 21**).

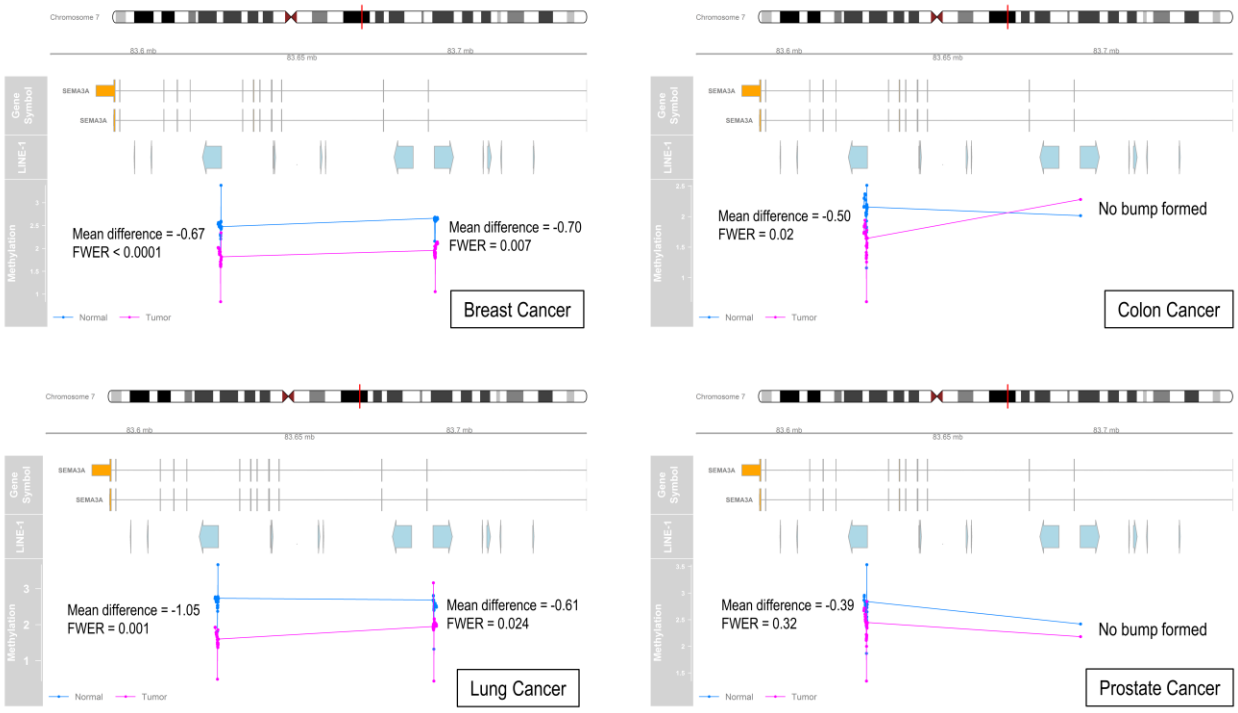

Figure 21 Two differentially methylated regions (DMR) of two intronic LINE-1 loci in gene *SEMA3A*.

The 5'UTR regions of two LINE-1 loci were hypomethylated in tumor tissues compared to matched normal tissues. In the "LINE-1" track, arrows indicate the strand direction. In the "Methylation" track, points represent CpG sites and lines represent mean methylation of the DMR.

SEMA3A can inhibit angiogenesis and endothelial cell migration and its downregulation has been identified in breast cancer development (Mishra et al., 2015). Using TCGA gene expression data,

we confirmed that *SEMA3A* gene expression was significantly downregulated in breast tumor tissues in relative to the matched normal tissues. This could be attributed to the hypomethylated LINE-1 loci as we observed significant positive correlation between methylation at each of the LINE-1 loci and *SEMA3A* gene expression in the normal tissues, but substantially attenuated and non-significant correlation in the tumor tissues (**Figure 22**).

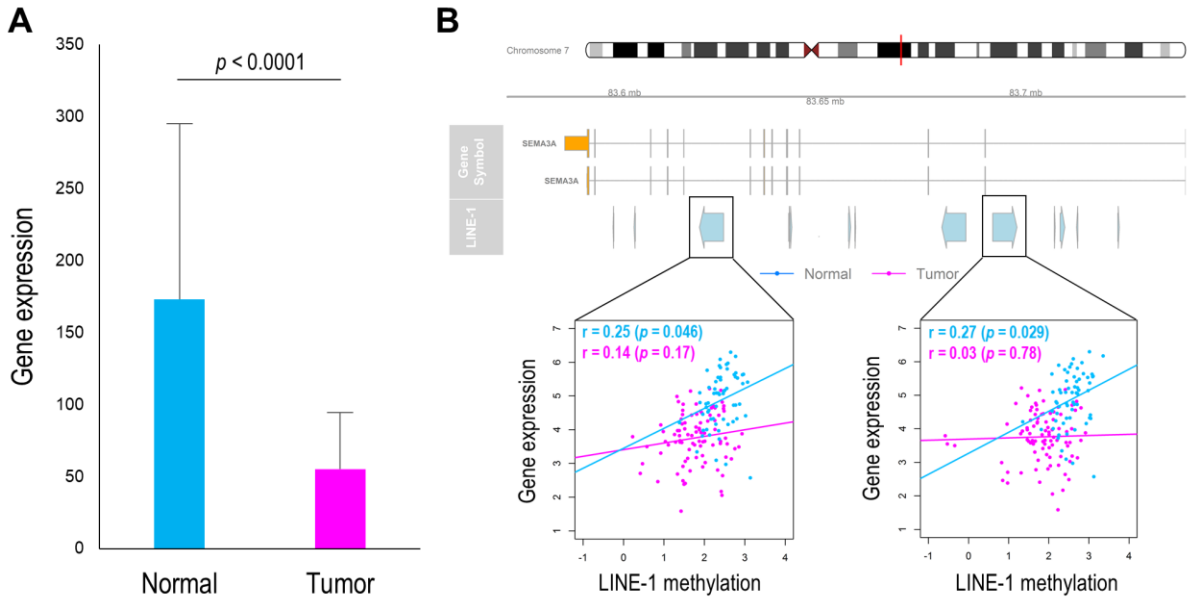

Figure 22 *SEMA3A* gene expression and LINE-1 methylation in breast tissue.

A: *SEMA3A* gene expression was significantly downregulated in breast tumor relative to matched normal tissues. B: Significant positive correlation between methylation of LINE-1 loci (same as shown in **Figure 21**) and *SEMA3A* gene expression in normal tissues, with greatly attenuated correlation in tumor tissues.

III. Discriminating tumor from normal tissue using locus-specific Alu and LINE-1 methylation

Finally, we implemented an ROC plot to compare the power of locus-specific Alu and LINE-1 methylation versus mean global methylation to discriminate between tumor and the paired normal samples. Mean methylation of CpGs in each Alu and LINE-1 locus were calculated to represent locus-specific methylation level. We demonstrated the discrimination power using extended or profiled Alu and LINE-1 in breast tumors, as other three tumors failed to yield convergent results due to limited sample sizes. The surrogate global methylation was computed by averaging all extended or profiled CpG methylation in Alu and LINE-1. We observed that locus-specific methylation achieved AUC of 98.3 (95% CI: 96.1 – 100.0), which was higher than that using the surrogate global methylation (74.1; 95% CI: 64.1 – 84.2; $p < 0.001$) in the extended Alu and LINE-1 (**Figure 23A**). For the profiled Alu and LINE-1 methylation, we observed lower AUC of 87.6 (95% CI: 80.6 – 94.6) for locus-specific methylation, which was again higher than the AUC using surrogate global methylation (76.9; 95% CI: 67.4 – 86.5), but not significantly so (**Figure 23B**).

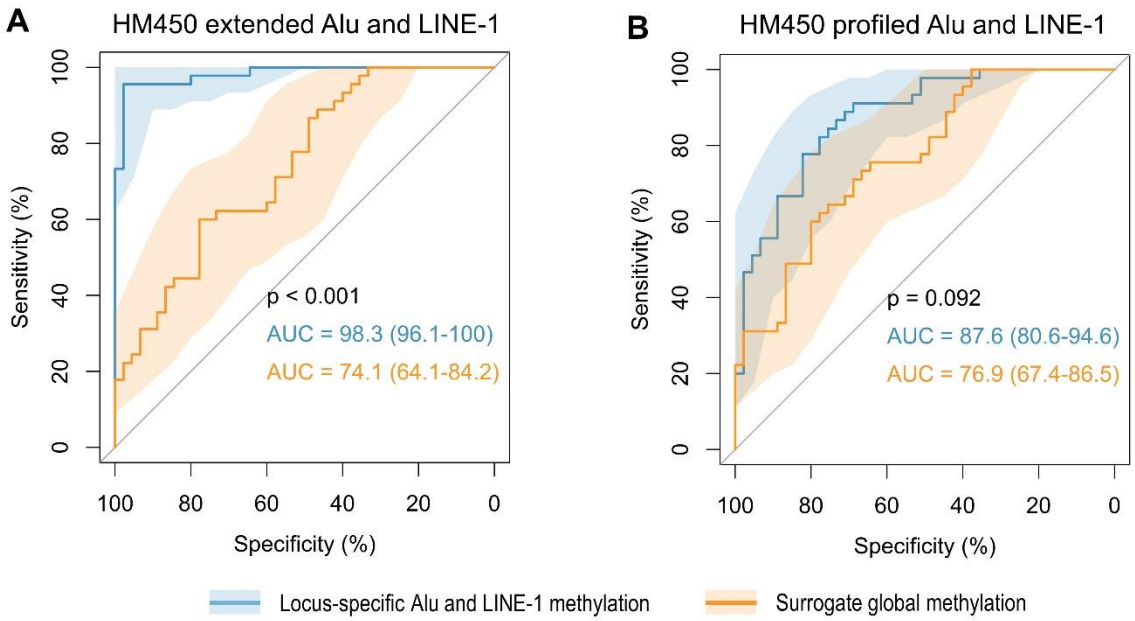

Figure 23 Discrimination power of locus-specific Alu/LINE-1 methylation vs surrogate global methylation.

A: extended Alu and LINE-1 methylation. B: profiled only. Shaded regions represent 95% confidence intervals of ROC curves. Locus-specific Alu and LINE-1 methylation achieved higher AUC than that using surrogate global methylation. Our predicted methylation achieved higher AUC than that using HM450-profiled methylation.

IV. Availability

`REMP` is available for download at Bioconductor: http://bioconductor.org/packages/REMP.

RepeatMasker Library (build hg19) and RefSeq Gene annotation database (build hg19) are available through the R package `AnnotationHub`, record number = AH5122 and AH5040, respectively.

GM12878 HM450 data are available at ENCODE:

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibMethyl450/

GM12878 EPIC data are available in R package `minfiDataEPIC`.

GM12878 RRBS data are available at ENCODE:

http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibMethylRrbs/

GM12878 WGBS data are available at ENCODE:

https://www.encodeproject.org/experiments/ENCSR000AJI/

GM12878 NimbleGen data are available upon reasonable request and with permission of Roche Sequencing.

HM450 data of TCGA tumor tissue and paired normal tissue are available at GDC Data Portal:

https://portal.gdc.cancer.gov/

Regulatory element data available in the ENCODE Analysis Hub at the European Bioinformatics Institute: http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/

CHAPTER 6 AN INTRODUCTION TO THE REMP PACKAGE

I. Introduction

`REMP` predicts DNA methylation of locus-specific repetitive elements (RE) by learning surrounding genetic and epigenetic information. `REMP` provides genomewide single-base resolution of DNA methylation on RE that are difficult to measure using array-based or sequencing-based platforms, which enables epigenome-wide association study and DMR analysis on RE (**Figure 24**). Please refer to **Appendix** for detailed information on functions and objects.

Standard procedure:

Step 1

Start out generating required datasets for prediction using `initREMP`. The datasets include RE information, RE-CpG (i.e. CpGs located in RE region) information, and gene annotation, which are maintained in a `REMParcel` object. It is recommended to save these generated data to working directory so they can be used in the future.

Step 2

Clean Illumina methylation dataset using `grooMethy`. This function can help identify and fix abnormal values and automatically impute missing values, which are essential for downstream prediction.

Step 3

Run `remp` to predict genome-wide locus specific RE methylation.

Step 4

Use the built-in accessors and utilities in `REMProduct` object to get or refine the prediction results.

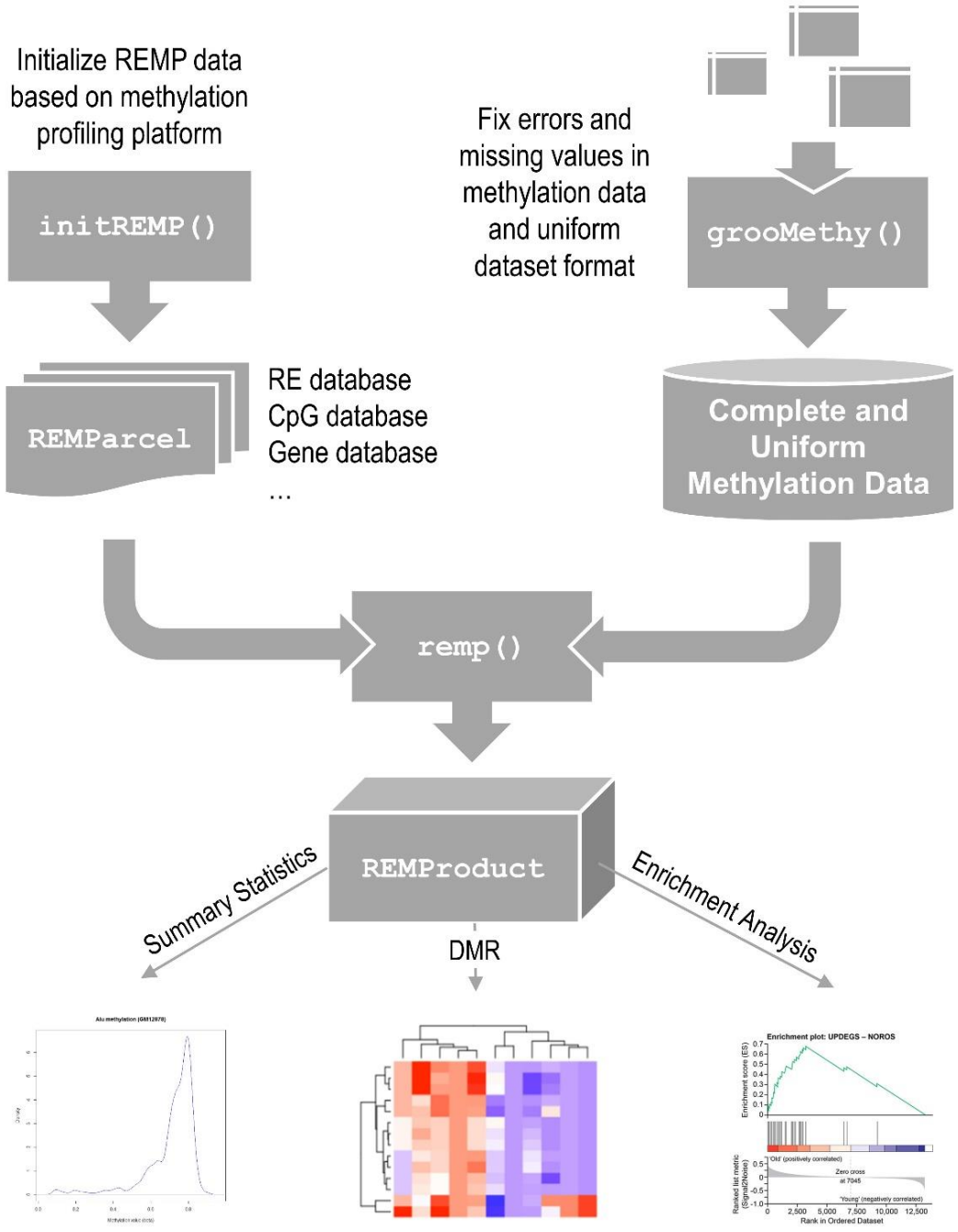

Figure 24 Standard analytical procedure using package `REMP`.

II. Installation

Install REMP (release version):

```
> source("http://bioconductor.org/biocLite.R")
> biocLite("REMP")
```

To install devel version:

```
> library(devtools)
> install_github("YinanZheng/REMP")
```

Load REMP into the workspace:

```
> library(REMP)
```

III. Usage

Currently REMP supports Human (hg19, build 37) Alu and LINE-1 (L1) repetitive element (RE) methylation prediction using Illumina HM450 or EPIC array.

a. Groom methylation data

Appropriate data preprocessing including quality control and normalization of methylation data are recommended before running REMP. Many packages are available to carry out these data preprocessing steps, for example, minfi, wateRmelon, and methylumi.

REMP is trying to minimize the requirement of the methylation data format. User can maintain the methylation data in RatioSet or GenomicRatioSet object offered by minfi,

`data.table`, `data.frame`, `DataFrame`, or `matrix` . User can input either β value or M value. There are only two basic requirements of the methylation data:

1. Each row should represent CpG probe and each column should represent sample.
2. The row names should indicate Illumina probe ID (i.e. cg00000029).

However, there are some other common data issues that may prevent `REMP` from running correctly. For example, if the methylation data are in β value and contain zero methylation values, logit transformation (to create M value) will create negative infinite value; or the methylation data contain `NA`, `Inf`, or `NaN` data. To tackle these potential issues, `REMP` includes a handy function grooMethy which can help detect and fix these issues. It is highly recommended to take advantage of this function:

```
> # Get GM12878 methylation data (450k array)
> GM12878_450k <- getGM12878('450k')
> GM12878_450k <- grooMethy(GM12878_450k, verbose = TRUE)
Illumina 450k Methylation data in beta value detected.
A total of 1715 zero beta values are found.
Fixing zero beta values with the smallest beta value = 0.001 ...
Converting beta value to M value ...
Packaging data to [Genomic]RatioSet object ...

> GM12878_450k
class: RatioSet
dim: 482421 1
metadata(0):
assays(2): Beta M
rownames(482421): cg00000029 cg00000108 ... cg27666046
cg27666123
```

```
rowData names(0):
colnames(1): GM12878
colData names(0):
Annotation
array: IlluminaHumanMethylation450k
annotation: ilmn12.hg19
Preprocessing
Method: NA
minfi version: NA
Manifest version: NA
```

For zero β values, `grooMethy` will replace them with smallest non-zero β value. For `NA/NaN/Inf` values, `grooMethy` will treat them as missing values and then apply KNN-imputation to complete the dataset. It is noteworthy that KNN-imputation can sometimes generate a few imputed values that are out of the original data range across samples. For these 'bad' imputed value, `grooMethy` will instead use the mean value across samples to do the imputation.

b. Prepare annotation data

To run `REMP` for RE methylation prediction, user first needs to prepare some annotation datasets. The function `initREMP` is designed to do the job.

Suppose user will predict Alu methylation using Illumina 450k array data:

```
> data(Alu.demo)
> remparcel <- initREMP(arrayType = "450k", REtype = "Alu", + RE
= Alu.demo, ncore = 1)
Start Alu annotation data initialization ... (0 sec.)
Illumina platform: 450k Done. (28 sec.)
```

```
> remparcel
REMParcel object
RE type: Alu
Illumina platform: 450k
Valid (max) RE-CpG flanking window size: 1200
Number of RE: 500
Number of RE-CpG: 5039
```

For demonstration, we only use 500 selected Alu sequence dataset which comes along with the `package (Alu.demo)`. We specify `RE = Alu.demo`, so that the annotation dataset will be generated for the 500 selected Alu sequences. In real-world prediction, specifying RE is not necessary, as the function will pull up the complete RE sequence dataset from package `AnnotationHub`.

All data are stored in the `REMParcel` object. It is recommended to specify a working directory so that the data generated can be preserved for later use:

```
> saveParcel(remparcel)
```

Without specifying working directory using option `work.dir`, the annotation dataset will be created under the temporal directory `tempdir()` by default. User can also turn on the export parameter in `initREMP` to save the data automatically.

c. Run prediction

Once the annotation data are ready, user can pass the annotation data parcel to `remp` for prediction:

```
> remp.res <- remp(GM12878_450k, REtype = 'Alu', parcel =
remparcel, autoTune = FALSE, param = 6, ncore = 1)
Start RE methylation prediction with 1 core(s) ... (0 sec.)
Preparing neighboring CpG information ... (0 sec.)
Predicting sample GM12878 ... (2 sec.)
     GM12878 completed! 0 sample(s) left ... (7 sec.)
Done. (7 sec.)
```

If parcel is missing, `remp` will then try to search the `REMParcel` data file in the directory indicated by `work.dir` (use `tempdir()` if not specified).

By default, `remp` uses Random Forest (method = '`rf`') model (package `randomForest`) for prediction. Random Forest model is recommended because it offers more accurate prediction results and it automatically enables Quantile Regression Forest (package `quantregForest`) for prediction reliability evaluation. `remp` constructs 19 predictors to carry out the prediction. For Random Forest model, the tuning parameter `param` = 6 (i.e. `mtry` in `randomForest`) indicates how many predictors will be randomly selected for building the individual trees. To speed up the prediction process, user can turn off the tuning functionality by `autoTune = FALSE` and specify the tuning parameter by `param = 6` - this is about one third of the total number of predictors, which is also default in package `randomForest`). The performance of Random Forest model is often relatively insensitive to the choice of `mtry`.

`remp` will return a `REMPset` object, which inherits Bioconductor's

`RangedSummarizedExperiment` class:

```
> remp.res
class: REMProduct
dim: 4808 1
metadata(7): REannotation RECpG ... REStats GeneStats
assays(3): rempB rempM rempQC
rownames: NULL
rowData names(1): RE.Index
colnames(1): GM12878
colData names(1): mtry

> # Display more detailed information
> details(remp.res)

RE type: Alu
Methylation profiling platform: 450k
Flanking window size: 1000
Prediction model: Random Forest
QC model: Quantile Regression Forest
Predicted 4808 CpG sites in 500 Alu

Number of predicted CpGs by chromosome:
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8
 449  276  293  131  179  397  292  102

chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16
  98   148   254   310   66   127   133   333

chr17 chr18 chr19 chr20 chr21 chr22
  295    81   674    66    37    67

Coverage information:
```

```
     There are 500 profiled Alu by Illumina array.
     There are 481 RE-CpGs that have neighboring profiled CpGs
are used for model training.
     In total, REMP predicts 500 Alu (4808 RE-CpG).
     Gene coverage by predicted Alu (out of total refSeq Gene):
          492 (1.96%) total genes;
          413 (2.15%) protein-coding genes;
          117 (1.59%) non-coding RNA genes.

Distribution of predicted methylation value (beta value):
   Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
0.03436    0.47619    0.66368    0.59703    0.75284    0.92060

Distribution of prediction reliability score:
   Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
0.7339     1.4700     1.7715     1.7947     2.0427     5.5871
```

Prediction results can be obtained by accessors:

```
> # Predicted RE-CpG methylation value (Beta value)
> head(rempB(remp.res))

DataFrame with 6 rows and 1 column
    GM12878
  <numeric>
1 0.7720857
2 0.7907542
3 0.7922319
4 0.7954692
5 0.7990411
6 0.7955649

> # Predicted RE-CpG methylation value (M value)
> head(rempM(remp.res))

DataFrame with 6 rows and 1 column
   GM12878
```

```
 <numeric>
1 1.760269
2 1.918030
3 1.930948
4 1.959488
5 1.991370
6 1.960336

> # Genomic location information of the predicted RE-CpG
> head(rowRanges(remp.res))

GRanges object with 6 ranges and 1 metadata column:
     seqnames          ranges  strand |    RE.Index
        <Rle>       <IRanges>   <Rle> |       <Rle>
[1]      chr1 [943927, 943928]      - | Alu_0000527
[2]      chr1 [943935, 943936]      - | Alu_0000527
[3]      chr1 [943968, 943969]      - | Alu_0000527
[4]      chr1 [943974, 943975]      - | Alu_0000527
[5]      chr1 [943991, 943992]      - | Alu_0000527
[6]      chr1 [944062, 944063]      - | Alu_0000527
-------
seqinfo: 93 sequences from an unspecified genome; no seqlengths
```

Trim off less reliable predicted results:

```
> # Any predicted CpG values with quality score < threshold
(default = 1.7) will be replaced with NA. CpGs contain more than
missingRate * 100% (default = 20%) missing rate across samples
will be discarded.
> # For mechanism study, more stringent cutoff is recommended.
> remp.res <- trim(remp.res)
> details(remp.res)

RE type: Alu
Methylation profiling platform: 450k
Flanking window size: 1000
Prediction model: Random Forest - trimmed (1.7)
QC model: Quantile Regression Forest
```

```
Predicted 2125 CpG sites in 388 Alu

Number of predicted CpGs by chromosome:
chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8
 210  131  143   77   67  204  135   45

chr9 chr10 chr11 chr12 chr13 chr14 chr15 chr16
  45    73    70   107    16    96    26   173

chr17 chr18 chr19 chr20 chr21 chr22
  136    35   263    42     4    27

Coverage information:
     There are 388 profiled Alu by Illumina array.
     There are 356 RE-CpGs that have neighboring profiled CpGs
are used for model training.
     In total, REMP predicts 388 Alu (2125 RE-CpG).
     Gene coverage by predicted Alu (out of total refSeq Gene):
          373 (1.49%) total genes;
          307 (1.6%) protein-coding genes;
          87 (1.18%) non-coding RNA genes.

Distribution of predicted methylation value (beta value):
  Min.     1st Qu.     Median     Mean     3rd Qu.     Max.
0.0720      0.6771     0.7479   0.7102      0.7922   0.9120

Distribution of prediction reliability score:
  Min.     1st Qu.     Median     Mean     3rd Qu.     Max.
0.7339      1.3167     1.4433   1.4278      1.5619   1.6998
```

To add genomic regions annotation of the predicted REs:

```
> # By default gene symbol annotation will be added
> remp.res <- decodeAnnot(remp.res)
You have successfully set parallel mode with 8 workers
(SnowParam).
Decoding Alu annotation to symbol ...
```

```
> annotation(remp.res)
```

Seven genomic region indicators will be added to the annotation data in the input `REMProduct` object:

- `InNM`: in protein-coding genes (overlap with refSeq gene's "NM" transcripts + 2000 bp upstream of the transcription start site (TSS))
- `InNR`: in noncoding RNA genes (overlap with refSeq gene's "NR" transcripts + 2000 bp upstream of the TSS)
- `InTSS`: in flanking region of 2000 bp upstream of the TSS. Default upstream limit is 2000 bp, which can be modified globally using remp_options
- `In5UTR`: in 5'untranslated regions (UTRs)
- `InCDS`: in coding DNA sequence regions
- `InExon`: in exon regions
- `In3UTR`: in 3'UTRs

Note that intron region and intergenic region information can be derived from the above genomic region indicators: if "`InNM`" and/or "`InNR`" is not missing but "`InTSS`", "`In5UTR`", "`InExon`", and "`In3UTR`" are missing, then the RE is strictly located within intron region; if all indicators are missing, then the RE is strictly located in intergenic region.

d. Plot prediction

Make a density plot of the predicted methylation (β values):

```
> plot(remp.res, main = "Alu methylation (GM12878)", col =
+ "blue")
```

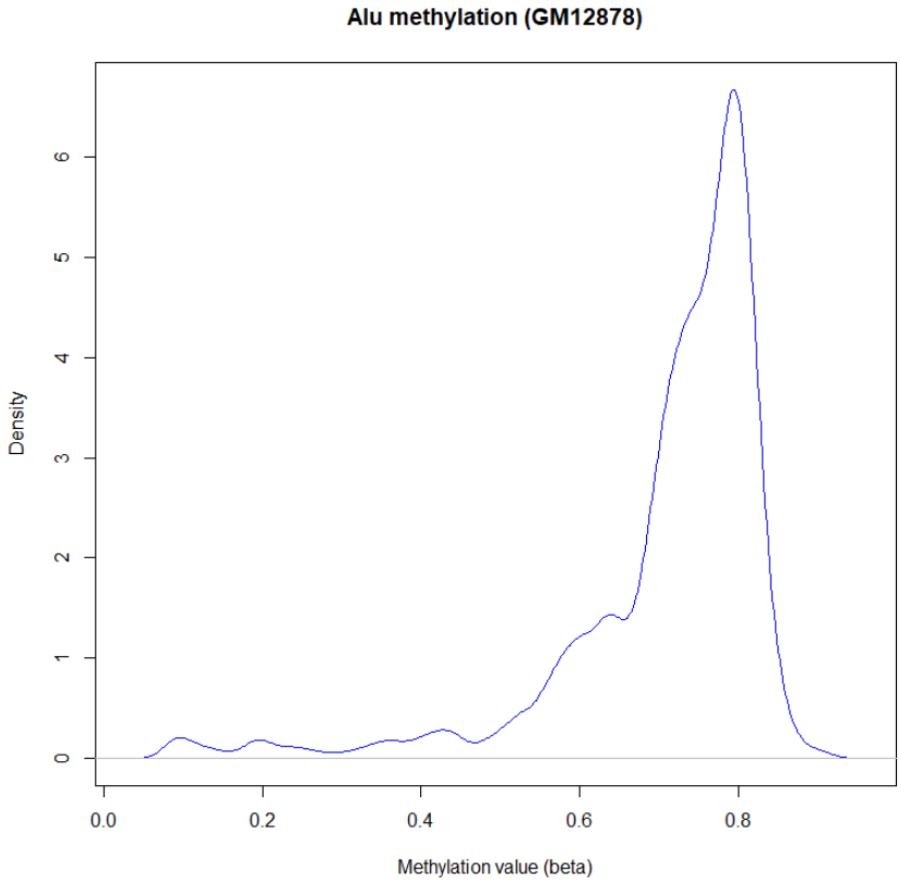

Figure 25 Example: Distribution of predicted methylation across CpGs in Alu.

SUMMARY AND CONCLUSIONS

Our prospective study of global DNA methylation and cancer found a series of complex relationships underlying Alu and LINE-1 methylation, cancer incidence and mortality, and time. Alu and LINE-1 hypermethylation of blood leukocytes may be a predictor of cancer incidence, in particular prostate cancer. Furthermore, Alu and LINE-1 methylation may serve as useful prognostic indicators of cancer mortality. The possible use of these two measures as cancer early detection and prognostic biomarkers, and whether such relationships are consistent with other measures of global DNA methylation, should be validated in larger prospective studies. In addition, the rate of change of Alu methylation and average LINE-1 methylation 10 years pre-diagnostically are interesting (if speculative) findings that need to be confirmed in future research. Taken together, these findings suggest that global DNA methylation plays a dynamic role in tumorigenesis and cancer progression, and identify several new avenues of research for the study of Alu and LINE-1 as biomarkers of cancer.

To better understand the role of DNA methylation in locus-specific RE and overcome the limitations in current bisulfite sequencing in RE, we developed a prediction algorithm and corresponding R package `REMP` (please refer to **Chapter 6** for detailed introduction) to predict locus-specific RE methylation by mining methylation information from neighboring CpG sites profiled in Infinium methylation arrays. We validated the reliability of our algorithm using both sequencing (i.e., NimbleGen) and EPIC array (covering over 850,000 CpGs) data, further verifying the algorithm's prediction performance by demonstrating the inverse relationships

between Alu/LINE-1 methylation and evolutionary age previously observed. More importantly, we tested the clinical use of our algorithm in TCGA data to examine epigenome-wide associations and distinguish tumor from normal tissues. Our algorithm may help address current challenges in studying the role of RE methylation in human diseases. It also directly addresses the assumption of a uniform methylation profile in RE with similar biological or pathological effects, which may have caused information loss in extant studies and hindered our understanding of the exact role that RE methylation plays in human diseases. Furthermore, as technologies for epigenomic profiling continue to improve, our algorithm can serve as an important framework for later expanding RE coverage. This will enhance our ability to investigate relationships between RE epigenetic features and complex traits/diseases in a highly cost-effective manner in large clinical and population studies.

Our algorithm was mainly developed based on the HM450 and EPIC arrays, since compared to other sequencing-based approaches the array-based data were the most robust for Alu/LINE-1 measurement (higher coverage in some sequencing platforms, e.g., WGBS, notwithstanding). In addition, the Infinium methylation array is the ideal source to provide reliable neighboring information for methylation prediction. Previous attempts at predicting methylation suggested that incorporating extensive neighboring information such as profiled CpG sites, genomic positions, DNA sequence properties, and *cis*-regulatory elements could yield highly accurate predictions (Zhang et al., 2015, Das et al., 2006, Zheng et al., 2013). However, in practice obtaining the requisite information is often impractical and infeasible. By leveraging the co-methylation features of neighboring CpGs and the structure of RE sequences, we devised a

simpler predictive strategy and achieved high predictive performance for our algorithm. Our algorithm only relies on predictors that are easily extractable from DNA methylation profiling data, minimizing dependence on a reference genome and preserving individual variability in the human epigenome.

The predictive power of our algorithm was further confirmed by testing Alu/LINE-1 methylation in relation to evolutionary age. As **Chapter 1** mentioned, Alu and LINE-1 propagated in mammalian genomes over the past 65 and 150 million years, respectively. This evolutionary process resulted in phylogenetic trees of Alu/LINE-1 subfamilies with different evolutionary ages (Kapitonov and Jurka, 1996, Smit et al., 1995). Evolutionary age of Alu/LINE-1 can be inferred by the divergence of the copies from the consensus sequence, as new base substitutions, insertions, or deletions accumulated in Alu/LINE-1 through the "copy and paste" retrotransposition activity. Older Alu/LINE-1 copies are in general inactive since more mutations have been induced, partially by CpG methylation, as the mutation rate of CpG sites is estimated to be 9.2 time faster than non-CpG loci in young Alu (Batzer et al., 1990). Younger Alu/LINE-1, especially current active ones, have less mutation and thus CpG methylation is active as a defense mechanism to suppress the retrotransposition activity. Therefore, DNA methylation level should be lower in older Alu/LINE-1 than ones in younger Alu/LINE-1. One of our previous studies confirmed this inverse relationship by bisulfite-PCR-pyrosequencing 10 differentially-evolved RE subfamilies (Byun et al., 2013). In accordance with these findings the current study also confirmed this hypothesis from a more comprehensive genome-wide perspective, which further supports the reliability of our prediction results. This demonstrates the potential utility of

our algorithm in studying more specific characteristics of RE methylation throughout the genome in connection with human diseases and other phenotypes, which may presently be impossible or impractical due to data limitations.

Our algorithm offers a more comprehensive perspective on the RE methylation landscape and biological implications of RE methylation on an epigenome-wide scale. The consistent enrichment of hypomethylated Alu and LINE-1 in regulatory regions (i.e., DNase and TFBS) across all four types of tested tumors highlights the potential *cis*-regulatory roles of Alu/LINE-1 methylation. This is supported by previous findings that RE derive a wide variety of gene regulatory regions, including DNase and TFBS in the human genome, demonstrating the effects of RE on regulating genes (Jordan et al., 2003, Thornburg et al., 2006, Marino-Ramirez and Jordan, 2006). The enrichment of hypomethylated Alu/LINE-1 in the histone modifications that we observed were largely consistent with a recent sequencing study of hypomethylated Alu in cancer cells (Jorda et al., 2017). Specifically, the enrichment of hypomethylated Alu and LINE-1 in H3K4me3 (a marker for transcriptional activation) and H3K9me3 (for transcriptional repression) suggests possible involvement of RE methylation in transcription activation events in tumor (Barski et al., 2007). Ward *et al.* demonstrated that Alu and LINE-1 are responsible for transcriptional activation and enriched in regions marked by H3K4me3 (Ward et al., 2013). DNA hypomethylation in Alu and LINE-1 in H3K9me3 could be an indicator of decreased H3K9me3, suggesting a less transcriptionally repressive function as H3K9me3 is shown to promote persistent DNA methylation in RE (Rose and Klose, 2014). Furthermore, hypomethylated Alu/LINE-1 in breast and colon cancer were overrepresented in H2A.Z, a histone variant that can

potentially alter nucleosome stability (Suto et al., 2000). It has been hypothesized that adequate genic methylation (mostly in RE) may stabilize translational control functions such as translation, ribosome biogenesis, RNA splicing, and protein localization by antagonizing H2A.Z deposition (Coleman-Derr and Zilberman, 2012). Thus hypomethylation of Alu/LINE-1 in H2A.Z may indicate dysfunction of translational control functions which are important to cancer etiology (Ruggero, 2013).

Our pathway-based analysis further supports the biological relevance of our predicted RE methylation. Olfactory transduction was one of the top-ranked pathways enriched by our predicted hypomethylated Alu in all four tumor tissues of interest. This pathway contains a large gene family of olfactory receptors (ORs), which have been found to be ectopically expressed in non-olfactory tissues (Kang and Koo, 2012) and for some ORs overexpressed in breast (Muranen et al., 2011), colon (Weber et al., 2017), lung (Giandomenico et al., 2013), and prostate tissues (Weng et al., 2006). In addition, we observed axon guidance pathway was significantly enriched with hypomethylated LINE-1 in breast, colon, and lung cancers. Axon guidance has been shown to play an important role in cancerogenesis (Chedotal et al., 2005). Further data analysis of the intronic locus-specific LINE-1 methylation and the host gene expression of *SEMA3A* supported the hypothesis that DNA methylation in intronic regions may potentially silence RE to maintain a gene's efficient transcription, and thus usually has a positive correlation with gene expression (Yang et al., 2014).

In the tumor-normal discrimination test, the improved AUC when using locus-specific Alu and LINE-1 methylation demonstrated the potential for information loss when using mean Alu and LINE-1 methylation (both widely used surrogate global methylation measures). In addition, the AUC using our extended Alu and LINE-1 methylation outperformed HM450 profiled methylation, further underscoring the valuable information added by our algorithm.

Several features and caveats of our algorithm are worth noting. First, our algorithm was not designed to cover whole-genome RE methylation, but rather to provide a reliable extension of RE methylation profiled using Infinium methylation arrays, which prioritize CpG interrogations in genes and functional regions. The predicted CpG sites maintained a similar genomic distribution as those profiled in the arrays, thus offering extended information on the biological roles of RE methylation in transcriptional regulation, identifying biomarkers of diseases, and devising useful clinical tools with minimal artificial bias. Second, our algorithm's performance can be influenced by methylation data quality and patterns of RE methylation in different tissues. However our algorithm allows for convenient evaluation and control of prediction reliability using the forest-based model to ensure prediction quality. Incorporating prediction reliability control may lead to missing data, posing potential challenges to downstream data analysis, however imputation techniques such as KNN-imputation (Troyanskaya et al., 2001) can be applied to obtain more complete data if needed. Third, the test of our algorithm's clinical utility was conducted only on TCGA HM450 data due to the lack of more advanced data, such as that from the EPIC array. Further investigations in larger human studies using such data to validate the clinical utility of our algorithm are warranted. Fourth, our algorithm was designed to predict

all types of RE methylation. However, our validation and clinical application tests only focused on the two most common human RE, Alu and LINE-1, due to their predominance throughout the human genome. The algorithm can be used on other human RE such as long terminal repeats and tandem repeats (Zhang et al., 2016).

In conclusion, our work reveals the spatiotemporal dynamics of DNA methylation in RE across human genome and its relations to cancer. The proposed algorithm can be applied to the widely-used methylation profiling platforms and extend RE CpG coverage in a highly cost-effective manner. More importantly it promotes genome-wide, locus-specific RE methylation association analyses in large human population and clinical studies by providing extended coverage of locus-specific RE methylation. This allows for more precise investigations into the tumorigenic (and potentially other etiological) roles of RE methylation, improving the accuracy of epigenetic studies. Our work may drive further investigations on how DNA methylation in RE may differ in their *cis*- and/or trans-effects on genomic stability, such as increasing mutation rates or aberrant gene expression, and identify novel RE loci that may exert important biological and pathological effects for cancer early detection and diagnosis.

REFERENCES

AKALIN, A., KORMAKSSON, M., LI, S., GARRETT-BAKELMAN, F. E., FIGUEROA, M. E., MELNICK, A. & MASON, C. E. 2012. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol,* 13**,** R87.

AMERICAN CANCER SOCIETY 2015. American Cancer Society: Cancer Facts and Figures 2015.

ANDREOTTI, G., KARAMI, S., PFEIFFER, R. M., HURWITZ, L., LIAO, L. M., WEINSTEIN, S. J., ALBANES, D., VIRTAMO, J., SILVERMAN, D. T., ROTHMAN, N. & MOORE, L. E. 2014. LINE1 methylation levels associated with increased bladder cancer risk in pre-diagnostic blood DNA among US (PLCO) and European (ATBC) cohort study participants. *Epigenetics,* 9**,** 404-15.

BABUSHOK, D. V. & KAZAZIAN, H. H., JR. 2007. Progress in understanding the biology of the human mutagen LINE-1. *Hum Mutat,* 28**,** 527-39.

BALASSIANO, K., LIMA, S., JENAB, M., OVERVAD, K., TJONNELAND, A., BOUTRON-RUAULT, M. C., CLAVEL-CHAPELON, F., CANZIAN, F., KAAKS, R., BOEING, H., MEIDTNER, K., TRICHOPOULOU, A., LAGLOU, P., VINEIS, P., PANICO, S., PALLI, D., GRIONI, S., TUMINO, R., LUND, E., BUENO-DE-MESQUITA, H. B., NUMANS, M. E., PEETERS, P. H., RAMON QUIROS, J., SANCHEZ, M. J., NAVARRO, C., ARDANAZ, E., DORRONSORO, M., HALLMANS, G., STENLING, R., EHRNSTROM, R., REGNER, S., ALLEN, N. E., TRAVIS, R. C., KHAW, K. T., OFFERHAUS, G. J., SALA, N., RIBOLI, E., HAINAUT, P., SCOAZEC, J. Y., SYLLA, B. S., GONZALEZ, C. A. & HERCEG, Z. 2011. Aberrant DNA methylation of cancer-associated genes in gastric cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC-EURGAST). *Cancer Lett,* 311**,** 85-95.

BARCHITTA, M., QUATTROCCHI, A., MAUGERI, A., VINCIGUERRA, M. & AGODI, A. 2014. LINE-1 hypomethylation in blood and tissue samples as an epigenetic marker for cancer risk: a systematic review and meta-analysis. *PLoS One,* 9**,** e109478.

BARIOL, C., SUTER, C., CHEONG, K., KU, S. L., MEAGHER, A., HAWKINS, N. & WARD, R. 2003. The relationship between hypomethylation and CpG island methylation in colorectal neoplasia. *Am J Pathol,* 162**,** 1361-71.

BARRY, K. H., MOORE, L. E., LIAO, L. M., HUANG, W. Y., ANDREOTTI, G., POULIN, M. & BERNDT, S. I. 2015. Prospective study of DNA methylation at LINE-1 and Alu in peripheral blood and the risk of prostate cancer. *Prostate,* 75**,** 1718-25.

BARSKI, A., CUDDAPAH, S., CUI, K., ROH, T. Y., SCHONES, D. E., WANG, Z., WEI, G., CHEPELEV, I. & ZHAO, K. 2007. High-resolution profiling of histone methylations in the human genome. *Cell,* 129**,** 823-37.

BATZER, M. A. & DEININGER, P. L. 2002. Alu repeats and human genomic diversity. *Nat Rev Genet,* 3**,** 370-9.

BATZER, M. A., KILROY, G. E., RICHARD, P. E., SHAIKH, T. H., DESSELLE, T. D., HOPPENS, C. L. & DEININGER, P. L. 1990. Structure and variability of recently inserted Alu family members. *Nucleic Acids Res,* 18**,** 6793-8.

BECK, C. R., GARCIA-PEREZ, J. L., BADGE, R. M. & MORAN, J. V. 2011. LINE-1 elements in structural variation and disease. *Annu Rev Genomics Hum Genet,* 12**,** 187-215.

BEISEL, C. & PARO, R. 2011. Silencing chromatin: comparing modes and mechanisms. *Nat Rev Genet,* 12**,** 123-35.

BELL, J. T., PAI, A. A., PICKRELL, J. K., GAFFNEY, D. J., PIQUE-REGI, R., DEGNER, J. F., GILAD, Y. & PRITCHARD, J. K. 2011. DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol,* 12**,** R10.

BIBIKOVA, M., BARNES, B., TSAN, C., HO, V., KLOTZLE, B., LE, J. M., DELANO, D., ZHANG, L., SCHROTH, G. P., GUNDERSON, K. L., FAN, J. B. & SHEN, R. 2011. High density DNA methylation array with single CpG site resolution. *Genomics,* 98**,** 288-95.

BIEMONT, C. & VIEIRA, C. 2006. Genetics: junk DNA as an evolutionary force. *Nature,* 443**,** 521-4.

BIRD, A. 2002. DNA methylation patterns and epigenetic memory. *Genes Dev,* 16**,** 6-21.

BOCK, C., PAULSEN, M., TIERLING, S., MIKESKA, T., LENGAUER, T. & WALTER, J. 2006. CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genet,* 2**,** e26.

BOCK, C., WALTER, J., PAULSEN, M. & LENGAUER, T. 2007. CpG island mapping by epigenome prediction. *PLoS Comput Biol,* 3**,** e110.

BREIMAN, L. 2001. Random Forests. *Mach. Learn.,* 45**,** 5-32.

BRENNAN, K. & FLANAGAN, J. M. 2012. Is there a link between genome-wide hypomethylation in blood and cancer risk? *Cancer Prev Res (Phila),* 5**,** 1345-57.

BRENNAN, K., GARCIA-CLOSAS, M., ORR, N., FLETCHER, O., JONES, M., ASHWORTH, A., SWERDLOW, A., THORNE, H., RIBOLI, E., VINEIS, P., DORRONSORO, M., CLAVEL-CHAPELON, F., PANICO, S., ONLAND-MORET, N. C., TRICHOPOULOS, D., KAAKS, R., KHAW, K. T., BROWN, R. & FLANAGAN, J. M. 2012. Intragenic ATM methylation in peripheral blood DNA as a biomarker of breast cancer risk. *Cancer Res,* 72**,** 2304-13.

BYUN, H. M., MOTTA, V., PANNI, T., BERTAZZI, P. A., APOSTOLI, P., HOU, L. & BACCARELLI, A. A. 2013. Evolutionary age of repetitive element subfamilies and sensitivity of DNA methylation to airborne pollutants. *Part Fibre Toxicol,* 10**,** 28.

CHEDOTAL, A., KERJAN, G. & MOREAU-FAUVARQUE, C. 2005. The brain within the tumor: new roles for axon guidance molecules in cancers. *Cell Death Differ,* 12**,** 1044-56.

CHEN, J. M., FEREC, C. & COOPER, D. N. 2006. LINE-1 endonuclease-dependent retrotranspositional events causing human genetic disease: mutation detection bias and multiple mechanisms of target gene disruption. *J Biomed Biotechnol,* 2006**,** 56182.

CLAMP, M., FRY, B., KAMAL, M., XIE, X., CUFF, J., LIN, M. F., KELLIS, M., LINDBLAD-TOH, K. & LANDER, E. S. 2007. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A,* 104**,** 19428-33.

COLAPRICO, A., SILVA, T. C., OLSEN, C., GAROFANO, L., CAVA, C., GAROLINI, D., SABEDOT, T. S., MALTA, T. M., PAGNOTTA, S. M., CASTIGLIONI, I., CECCARELLI, M., BONTEMPI, G. & NOUSHMEHR, H. 2016. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res,* 44**,** e71.

COLEMAN-DERR, D. & ZILBERMAN, D. 2012. DNA methylation, H2A.Z, and the regulation of constitutive expression. *Cold Spring Harb Symp Quant Biol,* 77**,** 147-54.

CORDAUX, R. & BATZER, M. A. 2009. The impact of retrotransposons on human genome evolution. *Nat Rev Genet,* 10**,** 691-703.

CORTES, C. & VAPNIK, V. 1995. Support-Vector Networks. *Mach. Learn.,* 20**,** 273-297.

COST, G. J., FENG, Q., JACQUIER, A. & BOEKE, J. D. 2002. Human L1 element target-primed reverse transcription in vitro. *EMBO J,* 21**,** 5899-910.

CRAVO, M., PINTO, R., FIDALGO, P., CHAVES, P., GLORIA, L., NOBRE-LEITAO, C. & COSTA MIRA, F. 1996. Global DNA hypomethylation occurs in the early stages of intestinal type gastric carcinoma. *Gut,* 39**,** 434-8.

CRISCIONE, S. W., ZHANG, Y., THOMPSON, W., SEDIVY, J. M. & NERETTI, N. 2014. Transcriptional landscape of repetitive elements in normal and cancer human cells. *BMC Genomics,* 15**,** 583.

DAS, R., DIMITROVA, N., XUAN, Z., ROLLINS, R. A., HAGHIGHI, F., EDWARDS, J. R., JU, J., BESTOR, T. H. & ZHANG, M. Q. 2006. Computational prediction of methylation status in human genomic sequences. *Proc Natl Acad Sci U S A,* 103**,** 10713-6.

DE KONING, A. P., GU, W., CASTOE, T. A., BATZER, M. A. & POLLOCK, D. D. 2011. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet,* 7**,** e1002384.

DELONG, E. R., DELONG, D. M. & CLARKE-PEARSON, D. L. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics,* 44**,** 837-45.

DEROO, L. A., BOLICK, S. C., XU, Z., UMBACH, D. M., SHORE, D., WEINBERG, C. R., SANDLER, D. P. & TAYLOR, J. A. 2014. Global DNA methylation and one-carbon metabolism gene polymorphisms and the risk of breast cancer in the Sister Study. *Carcinogenesis,* 35**,** 333-8.

DEWANNIEUX, M. & HEIDMANN, T. 2005. Role of poly(A) tail length in Alu retrotransposition. *Genomics,* 86**,** 378-81.

DI, J. Z., HAN, X. D., GU, W. Y., WANG, Y., ZHENG, Q., ZHANG, P., WU, H. M. & ZHU, Z. Z. 2011. Association of hypomethylation of LINE-1 repetitive element in blood leukocyte DNA with an increased risk of hepatocellular carcinoma. *J Zhejiang Univ Sci B,* 12**,** 805-11.

DU, P., ZHANG, X., HUANG, C. C., JAFARI, N., KIBBE, W. A., HOU, L. & LIN, S. M. 2010. Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics,* 11**,** 587.

DUHAIME-ROSS, A. 2014. Revved-up epigenetic sequencing may foster new diagnostics. *Nat Med,* 20**,** 2.

ECKHARDT, F., LEWIN, J., CORTESE, R., RAKYAN, V. K., ATTWOOD, J., BURGER, M., BURTON, J., COX, T. V., DAVIES, R., DOWN, T. A., HAEFLIGER, C., HORTON, R., HOWE, K., JACKSON, D. K., KUNDE, J., KOENIG, C., LIDDLE, J., NIBLETT, D., OTTO, T., PETTETT, R., SEEMANN, S., THOMPSON, C., WEST, T., ROGERS, J., OLEK, A., BERLIN, K. & BECK, S. 2006. DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat Genet,* 38**,** 1378-85.

EDWARDS, J. R., O'DONNELL, A. H., ROLLINS, R. A., PECKHAM, H. E., LEE, C., MILEKIC, M. H., CHANRION, B., FU, Y., SU, T., HIBSHOOSH, H., GINGRICH, J.

A., HAGHIGHI, F., NUTTER, R. & BESTOR, T. H. 2010. Chromatin and sequence features that define the fine and gross structure of genomic methylation patterns. *Genome Res,* 20**,** 972-80.

EHRLICH, M. 2002. DNA methylation in cancer: too much, but also too little. *Oncogene,* 21**,** 5400-13.

EHRLICH, M. 2009. DNA hypomethylation in cancer cells. *Epigenomics,* 1**,** 239-59.

EHRLICH, M., WOODS, C. B., YU, M. C., DUBEAU, L., YANG, F., CAMPAN, M., WEISENBERGER, D. J., LONG, T., YOUN, B., FIALA, E. S. & LAIRD, P. W. 2006. Quantitative analysis of associations between DNA hypermethylation, hypomethylation, and DNMT RNA levels in ovarian tumors. *Oncogene,* 25**,** 2636-45.

EKRAM, M. B., KANG, K., KIM, H. & KIM, J. 2012. Retrotransposons as a major source of epigenetic variations in the mammalian genome. *Epigenetics,* 7**,** 370-82.

ENCODE PROJECT CONSORTIUM 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489**,** 57-74.

ESTELLER, M. 2007. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat Rev Genet,* 8**,** 286-98.

EZKURDIA, I., JUAN, D., RODRIGUEZ, J. M., FRANKISH, A., DIEKHANS, M., HARROW, J., VAZQUEZ, J., VALENCIA, A. & TRESS, M. L. 2014. Multiple evidence strands suggest that there may be as few as 19,000 human protein-coding genes. *Hum Mol Genet,* 23**,** 5866-78.

FAN, S., ZHANG, M. Q. & ZHANG, X. 2008. Histone methylation marks play important roles in predicting the methylation status of CpG islands. *Biochem Biophys Res Commun,* 374**,** 559-64.

FANG, F., FAN, S., ZHANG, X. & ZHANG, M. Q. 2006. Predicting methylation status of CpG islands in the human brain. *Bioinformatics,* 22**,** 2204-9.

FENG, Q., MORAN, J. V., KAZAZIAN, H. H., JR. & BOEKE, J. D. 1996. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell,* 87**,** 905-16.

FESCHOTTE, C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet,* 9**,** 397-405.

FORTIN, J. & HANSEN, K. D. 2016. minfiDataEPIC: Example data for the Illumina Methylation EPIC array. . *R package version 1.0.0*.

GAO, Y., BACCARELLI, A., SHU, X. O., JI, B. T., YU, K., TARANTINI, L., YANG, G., LI, H. L., HOU, L., ROTHMAN, N., ZHENG, W., GAO, Y. T. & CHOW, W. H. 2012. Blood leukocyte Alu and LINE-1 methylation and gastric cancer risk in the Shanghai Women's Health Study. *Br J Cancer,* 106**,** 585-91.

GIANDOMENICO, V., CUI, T., GRIMELIUS, L., OBERG, K., PELOSI, G. & TSOLAKIS, A. V. 2013. Olfactory receptor 51E1 as a novel target for diagnosis in somatostatin receptor-negative lung carcinoids. *J Mol Endocrinol,* 51**,** 277-86.

GOFF, S. A., RICKE, D., LAN, T. H., PRESTING, G., WANG, R., DUNN, M., GLAZEBROOK, J., SESSIONS, A., OELLER, P., VARMA, H., HADLEY, D., HUTCHISON, D., MARTIN, C., KATAGIRI, F., LANGE, B. M., MOUGHAMER, T., XIA, Y., BUDWORTH, P., ZHONG, J., MIGUEL, T., PASZKOWSKI, U., ZHANG, S., COLBERT, M., SUN, W. L., CHEN, L., COOPER, B., PARK, S., WOOD, T. C., MAO, L., QUAIL, P., WING, R., DEAN, R., YU, Y., ZHARKIKH, A., SHEN, R.,

SAHASRABUDHE, S., THOMAS, A., CANNINGS, R., GUTIN, A., PRUSS, D., REID, J., TAVTIGIAN, S., MITCHELL, J., ELDREDGE, G., SCHOLL, T., MILLER, R. M., BHATNAGAR, S., ADEY, N., RUBANO, T., TUSNEEM, N., ROBINSON, R., FELDHAUS, J., MACALMA, T., OLIPHANT, A. & BRIGGS, S. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science,* 296**,** 92-100.

HADJIARGYROU, M. & DELIHAS, N. 2013. The intertwining of transposable elements and non-coding RNAs. *Int J Mol Sci,* 14**,** 13307-28.

HALLING, K. C., LAZZARO, C. R., HONCHEL, R., BUFILL, J. A., POWELL, S. M., ARNDT, C. A. & LINDOR, N. M. 1999. Hereditary desmoid disease in a family with a germline Alu I repeat mutation of the APC gene. *Hum Hered,* 49**,** 97-102.

HANCKS, D. C. & KAZAZIAN, H. H., JR. 2012. Active human retrotransposons: variation and disease. *Curr Opin Genet Dev,* 22**,** 191-203.

HANSEN, K. D., TIMP, W., BRAVO, H. C., SABUNCIYAN, S., LANGMEAD, B., MCDONALD, O. G., WEN, B., WU, H., LIU, Y., DIEP, D., BRIEM, E., ZHANG, K., IRIZARRY, R. A. & FEINBERG, A. P. 2011. Increased methylation variation in epigenetic domains across cancer types. *Nat Genet,* 43**,** 768-75.

HASLER, J., SAMUELSSON, T. & STRUB, K. 2007. Useful 'junk': Alu RNAs in the human transcriptome. *Cell Mol Life Sci,* 64**,** 1793-800.

HOU, L., WANG, H., SARTORI, S., GAWRON, A., LISSOWSKA, J., BOLLATI, V., TARANTINI, L., ZHANG, F. F., ZATONSKI, W., CHOW, W. H. & BACCARELLI, A. 2010. Blood leukocyte DNA hypomethylation and gastric cancer risk in a high-risk Polish population. *Int J Cancer,* 127**,** 1866-74.

HOU, L., ZHANG, X., TARANTINI, L., NORDIO, F., BONZINI, M., ANGELICI, L., MARINELLI, B., RIZZO, G., CANTONE, L., APOSTOLI, P., BERTAZZI, P. A. & BACCARELLI, A. 2011. Ambient PM exposure and DNA methylation in tumor suppressor genes: a cross-sectional study. *Part Fibre Toxicol,* 8**,** 25.

INTERNATIONAL HAPMAP, C. 2003. The International HapMap Project. *Nature,* 426**,** 789-96.

INTERNATIONAL HAPMAP, C. 2005. A haplotype map of the human genome. *Nature,* 437**,** 1299-320.

INTERNATIONAL HUMAN GENOME SEQUENCING, C. 2004. Finishing the euchromatic sequence of the human genome. *Nature,* 431**,** 931-45.

JAFFE, A. E., MURAKAMI, P., LEE, H., LEEK, J. T., FALLIN, M. D., FEINBERG, A. P. & IRIZARRY, R. A. 2012. Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International journal of epidemiology,* 41**,** 200-209.

JAMES, P., GIRIJADEVI, R., CHARLES, S. & PILLAI, M. R. 2013. MethFinder - A software package for prediction of human tissue-specific methylation status of CpG islands. *Bioinformation,* 9**,** 61-4.

JIANG, B., ZHANG, X. & CAI, T. 2008. Estimating the Confidence Interval for Prediction Errors of Support Vector Machine Classifiers. *J. Mach. Learn. Res.,* 9**,** 521-540.

JONES, P. A. & BAYLIN, S. B. 2002. The fundamental role of epigenetic events in cancer. *Nat Rev Genet,* 3**,** 415-28.

JORDA, M., DIEZ-VILLANUEVA, A., MALLONA, I., MARTIN, B., LOIS, S., BARRERA, V., ESTELLER, M., VAVOURI, T. & PEINADO, M. A. 2017. The epigenetic landscape

of Alu repeats delineates the structural and functional genomic architecture of colon cancer cells. *Genome Res,* 27**,** 118-132.

JORDAN, I. K., ROGOZIN, I. B., GLAZKO, G. V. & KOONIN, E. V. 2003. Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet,* 19**,** 68-72.

JOYCE, B. T., GAO, T., LIU, L., ZHENG, Y., LIU, S., ZHANG, W., PENEDO, F., DAI, Q., SCHWARTZ, J., BACCARELLI, A. A. & HOU, L. 2015. Longitudinal Study of DNA Methylation of Inflammatory Genes and Cancer Risk. *Cancer Epidemiol Biomarkers Prev*.

JOYCE, B. T., GAO, T., ZHENG, Y., LIU, L., ZHANG, W., DAI, Q., SHRUBSOLE, M. J., HIBLER, E. A., CRISTOFANILLI, M., ZHANG, H., YANG, H., VOKONAS, P., CANTONE, L., SCHWARTZ, J., BACCARELLI, A. & HOU, L. 2016. Prospective changes in global DNA methylation and cancer incidence and mortality. *Br J Cancer,* 115**,** 465-72.

KANG, N. & KOO, J. 2012. Olfactory receptors in non-chemosensory tissues. *BMB Rep,* 45**,** 612-22.

KANNAN, S., CHERNIKOVA, D., ROGOZIN, I. B., POLIAKOV, E., MANAGADZE, D., KOONIN, E. V. & MILANESI, L. 2015. Transposable Element Insertions in Long Intergenic Non-Coding RNA Genes. *Front Bioeng Biotechnol,* 3**,** 71.

KAPITONOV, V. & JURKA, J. 1996. The age of Alu subfamilies. *J Mol Evol,* 42**,** 59-65.

KAPUSTA, A., KRONENBERG, Z., LYNCH, V. J., ZHUO, X., RAMSAY, L., BOURQUE, G., YANDELL, M. & FESCHOTTE, C. 2013. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. *PLoS Genet,* 9**,** e1003470.

KARAMI, S., ANDREOTTI, G., LIAO, L. M., PFEIFFER, R. M., WEINSTEIN, S. J., PURDUE, M. P., HOFMANN, J. N., ALBANES, D., MANNISTO, S. & MOORE, L. E. 2015. LINE1 methylation levels in pre-diagnostic leukocyte DNA and future renal cell carcinoma risk. *Epigenetics,* 10**,** 282-92.

KOLOSHA, V. O. & MARTIN, S. L. 1997. In vitro properties of the first ORF protein from mouse LINE-1 support its role in ribonucleoprotein particle formation during retrotransposition. *Proc Natl Acad Sci U S A,* 94**,** 10155-60.

KONKEL, M. K. & BATZER, M. A. 2010. A mobile threat to genome stability: The impact of non-LTR retrotransposons upon the human genome. *Semin Cancer Biol,* 20**,** 211-21.

KUHN, M. 2008. Caret package. *Journal of Statistical Software,* 28**,** 1-26.

LANDER, E. S., LINTON, L. M., BIRREN, B., NUSBAUM, C., ZODY, M. C., BALDWIN, J., DEVON, K., DEWAR, K., DOYLE, M., FITZHUGH, W., FUNKE, R., GAGE, D., HARRIS, K., HEAFORD, A., HOWLAND, J., KANN, L., LEHOCZKY, J., LEVINE, R., MCEWAN, P., MCKERNAN, K., MELDRIM, J., MESIROV, J. P., MIRANDA, C., MORRIS, W., NAYLOR, J., RAYMOND, C., ROSETTI, M., SANTOS, R., SHERIDAN, A., SOUGNEZ, C., STANGE-THOMANN, N., STOJANOVIC, N., SUBRAMANIAN, A., WYMAN, D., ROGERS, J., SULSTON, J., AINSCOUGH, R., BECK, S., BENTLEY, D., BURTON, J., CLEE, C., CARTER, N., COULSON, A., DEADMAN, R., DELOUKAS, P., DUNHAM, A., DUNHAM, I., DURBIN, R., FRENCH, L., GRAFHAM, D., GREGORY, S., HUBBARD, T., HUMPHRAY, S., HUNT, A., JONES, M., LLOYD, C., MCMURRAY, A., MATTHEWS, L., MERCER,

S., MILNE, S., MULLIKIN, J. C., MUNGALL, A., PLUMB, R., ROSS, M., SHOWNKEEN, R., SIMS, S., WATERSTON, R. H., WILSON, R. K., HILLIER, L. W., MCPHERSON, J. D., MARRA, M. A., MARDIS, E. R., FULTON, L. A., CHINWALLA, A. T., PEPIN, K. H., GISH, W. R., CHISSOE, S. L., WENDL, M. C., DELEHAUNTY, K. D., MINER, T. L., DELEHAUNTY, A., KRAMER, J. B., COOK, L. L., FULTON, R. S., JOHNSON, D. L., MINX, P. J., CLIFTON, S. W., HAWKINS, T., BRANSCOMB, E., PREDKI, P., RICHARDSON, P., WENNING, S., SLEZAK, T., DOGGETT, N., CHENG, J. F., OLSEN, A., LUCAS, S., ELKIN, C., UBERBACHER, E., FRAZIER, M., et al. 2001. Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

LI, T. H. & SCHMID, C. W. 2001. Differential stress induction of individual Alu loci: implications for transcription and retrotransposition. *Gene,* 276**,** 135-41.

LI, Y., ZHU, J., TIAN, G., LI, N., LI, Q., YE, M., ZHENG, H., YU, J., WU, H., SUN, J., ZHANG, H., CHEN, Q., LUO, R., CHEN, M., HE, Y., JIN, X., ZHANG, Q., YU, C., ZHOU, G., SUN, J., HUANG, Y., ZHENG, H., CAO, H., ZHOU, X., GUO, S., HU, X., LI, X., KRISTIANSEN, K., BOLUND, L., XU, J., WANG, W., YANG, H., WANG, J., LI, R., BECK, S., WANG, J. & ZHANG, X. 2010. The DNA methylome of human peripheral blood mononuclear cells. *PLoS Biol,* 8**,** e1000533.

LIAO, L. M., BRENNAN, P., VAN BEMMEL, D. M., ZARIDZE, D., MATVEEV, V., JANOUT, V., KOLLAROVA, H., BENCKO, V., NAVRATILOVA, M., SZESZENIA-DABROWSKA, N., MATES, D., ROTHMAN, N., BOFFETTA, P., CHOW, W. H. & MOORE, L. E. 2011. LINE-1 methylation levels in leukocyte DNA and risk of renal cell cancer. *PLoS One,* 6**,** e27361.

LIM, U., FLOOD, A., CHOI, S. W., ALBANES, D., CROSS, A. J., SCHATZKIN, A., SINHA, R., KATKI, H. A., CASH, B., SCHOENFELD, P. & STOLZENBERG-SOLOMON, R. 2008. Genomic methylation of leukocyte DNA in relation to colorectal adenoma among asymptomatic women. *Gastroenterology,* 134**,** 47-55.

LISANTI, S., OMAR, W. A., TOMASZEWSKI, B., DE PRINS, S., JACOBS, G., KOPPEN, G., MATHERS, J. C. & LANGIE, S. A. 2013. Comparison of methods for quantification of global DNA methylation in human cells and tissues. *PLoS One,* 8**,** e79044.

LISTER, R., PELIZZOLA, M., DOWEN, R. H., HAWKINS, R. D., HON, G., TONTI-FILIPPINI, J., NERY, J. R., LEE, L., YE, Z., NGO, Q. M., EDSALL, L., ANTOSIEWICZ-BOURGET, J., STEWART, R., RUOTTI, V., MILLAR, A. H., THOMSON, J. A., REN, B. & ECKER, J. R. 2009. Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature,* 462**,** 315-22.

LU, X.-J., XUE, H.-Y., QI, X., XU, J. & MA, S.-J. 2015. LINE-1 in cancer: multifaceted functions and potential clinical implications. *Genet Med*.

LUO, Y., LU, X. & XIE, H. 2014. Dynamic Alu methylation during normal development, aging, and tumorigenesis. *Biomed Res Int,* 2014**,** 784706.

MARINO-RAMIREZ, L. & JORDAN, I. K. 2006. Transposable element derived DNaseI-hypersensitive sites in the human genome. *Biol Direct,* 1**,** 20.

MARTIN, S. L. & BUSHMAN, F. D. 2001. Nucleic acid chaperone activity of the ORF1 protein from the mouse LINE-1 retrotransposon. *Mol Cell Biol,* 21**,** 467-75.

MAUMUS, F. & QUESNEVILLE, H. 2014. Deep investigation of Arabidopsis thaliana junk DNA reveals a continuum between repetitive elements and genomic dark matter. *PLoS One,* 9**,** e94101.

MEINSHAUSEN, N. 2006. Quantile regression forests. *Journal of Machine Learning Research,* 7**,** 983-999.

MEINSHAUSEN, N. 2016. quantregForest: Quantile Regression Forests. R package version 1.3-5.

MEISSNER, A., GNIRKE, A., BELL, G. W., RAMSAHOYE, B., LANDER, E. S. & JAENISCH, R. 2005. Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. *Nucleic acids research,* 33**,** 5868-5877.

MEISSNER, A., MIKKELSEN, T. S., GU, H., WERNIG, M., HANNA, J., SIVACHENKO, A., ZHANG, X., BERNSTEIN, B. E., NUSBAUM, C., JAFFE, D. B., GNIRKE, A., JAENISCH, R. & LANDER, E. S. 2008. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature,* 454**,** 766-70.

MIKI, Y., NISHISHO, I., HORII, A., MIYOSHI, Y., UTSUNOMIYA, J., KINZLER, K. W., VOGELSTEIN, B. & NAKAMURA, Y. 1992. Disruption of the APC gene by a retrotransposal insertion of L1 sequence in a colon cancer. *Cancer Res,* 52**,** 643-5.

MISHRA, R., THORAT, D., SOUNDARARAJAN, G., PRADHAN, S. J., CHAKRABORTY, G., LOHITE, K., KARNIK, S. & KUNDU, G. C. 2015. Semaphorin 3A upregulates FOXO 3a-dependent MelCAM expression leading to attenuation of breast tumor growth and angiogenesis. *Oncogene,* 34**,** 1584-95.

MOEN, E. L., ZHANG, X., MU, W., DELANEY, S. M., WING, C., MCQUADE, J., MYERS, J., GODLEY, L. A., DOLAN, M. E. & ZHANG, W. 2013. Genome-wide variation of cytosine modifications between European and African populations and the implications for complex traits. *Genetics,* 194**,** 987-96.

MOORE, L. E., PFEIFFER, R. M., POSCABLO, C., REAL, F. X., KOGEVINAS, M., SILVERMAN, D., GARCIA-CLOSAS, R., CHANOCK, S., TARDON, A., SERRA, C., CARRATO, A., DOSEMECI, M., GARCIA-CLOSAS, M., ESTELLER, M., FRAGA, M., ROTHMAN, N. & MALATS, N. 2008. Genomic DNA hypomethylation as a biomarker for bladder cancer susceptibility in the Spanish Bladder Cancer Study: a case-control study. *Lancet Oncol,* 9**,** 359-66.

MORGAN, H. D., SUTHERLAND, H. G., MARTIN, D. I. & WHITELAW, E. 1999. Epigenetic inheritance at the agouti locus in the mouse. *Nat Genet,* 23**,** 314-8.

MORGAN, M., CARLSON, M., TENENBAUM, D. & ARORA, S. 2016. Annotationhub: Client to access annotationhub resources. *R package version 2.6.4*.

MOUSE GENOME SEQUENCING, C., WATERSTON, R. H., LINDBLAD-TOH, K., BIRNEY, E., ROGERS, J., ABRIL, J. F., AGARWAL, P., AGARWALA, R., AINSCOUGH, R., ALEXANDERSSON, M., AN, P., ANTONARAKIS, S. E., ATTWOOD, J., BAERTSCH, R., BAILEY, J., BARLOW, K., BECK, S., BERRY, E., BIRREN, B., BLOOM, T., BORK, P., BOTCHERBY, M., BRAY, N., BRENT, M. R., BROWN, D. G., BROWN, S. D., BULT, C., BURTON, J., BUTLER, J., CAMPBELL, R. D., CARNINCI, P., CAWLEY, S., CHIAROMONTE, F., CHINWALLA, A. T., CHURCH, D. M., CLAMP, M., CLEE, C., COLLINS, F. S., COOK, L. L., COPLEY, R. R., COULSON, A., COURONNE, O., CUFF, J., CURWEN, V., CUTTS, T., DALY, M., DAVID, R., DAVIES, J., DELEHAUNTY, K. D., DERI, J., DERMITZAKIS, E. T.,

DEWEY, C., DICKENS, N. J., DIEKHANS, M., DODGE, S., DUBCHAK, I., DUNN, D. M., EDDY, S. R., ELNITSKI, L., EMES, R. D., ESWARA, P., EYRAS, E., FELSENFELD, A., FEWELL, G. A., FLICEK, P., FOLEY, K., FRANKEL, W. N., FULTON, L. A., FULTON, R. S., FUREY, T. S., GAGE, D., GIBBS, R. A., GLUSMAN, G., GNERRE, S., GOLDMAN, N., GOODSTADT, L., GRAFHAM, D., GRAVES, T. A., GREEN, E. D., GREGORY, S., GUIGO, R., GUYER, M., HARDISON, R. C., HAUSSLER, D., HAYASHIZAKI, Y., HILLIER, L. W., HINRICHS, A., HLAVINA, W., HOLZER, T., HSU, F., HUA, A., HUBBARD, T., HUNT, A., JACKSON, I., JAFFE, D. B., JOHNSON, L. S., JONES, M., JONES, T. A., JOY, A., KAMAL, M., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature,* 420**,** 520-62.

MURANEN, T. A., GRECO, D., FAGERHOLM, R., KILPIVAARA, O., KAMPJARVI, K., AITTOMAKI, K., BLOMQVIST, C., HEIKKILA, P., BORG, A. & NEVANLINNA, H. 2011. Breast tumors from CHEK2 1100delC-mutation carriers: genomic landscape and clinical implications. *Breast Cancer Res,* 13**,** R90.

NEALE, R. E., CLARK, P. J., FAWCETT, J., FRITSCHI, L., NAGLER, B. N., RISCH, H. A., WALTERS, R. J., CRAWFORD, W. J., WEBB, P. M., WHITEMAN, D. C. & BUCHANAN, D. D. 2014. Association between hypermethylation of DNA repetitive elements in white blood cell DNA and pancreatic cancer. *Cancer Epidemiol,* 38**,** 576-82.

NUSGEN, N., GOERING, W., DAUKSA, A., BISWAS, A., JAMIL, M. A., DIMITRIOU, I., SHARMA, A., SINGER, H., FIMMERS, R., FROHLICH, H., OLDENBURG, J., GULBINAS, A., SCHULZ, W. A. & EL-MAARRI, O. 2015. Inter-locus as well as intra-locus heterogeneity in LINE-1 promoter methylation in common human cancers suggests selective demethylation pressure at specific CpGs. *Clin Epigenetics,* 7**,** 17.

OSTERTAG, E. M. & KAZAZIAN, H. H., JR. 2001. Biology of mammalian L1 retrotransposons. *Annu Rev Genet,* 35**,** 501-38.

PHOKAEW, C., KOWUDTITHAM, S., SUBBALEKHA, K., SHUANGSHOTI, S. & MUTIRANGURA, A. 2008. LINE-1 methylation patterns of different loci in normal and cancerous cells. *Nucleic Acids Res,* 36**,** 5704-12.

POBSOOK, T., SUBBALEKHA, K., SANNIKORN, P. & MUTIRANGURA, A. 2011. Improved measurement of LINE-1 sequence methylation for cancer detection. *Clin Chim Acta,* 412**,** 314-21.

PRICE, E. M., COTTON, A. M., PENAHERRERA, M. S., MCFADDEN, D. E., KOBOR, M. S. & ROBINSON, W. 2012. Different measures of "genome-wide" DNA methylation exhibit unique properties in placental and somatic tissues. *Epigenetics,* 7**,** 652-63.

PRUITT, K. D., BROWN, G. R., HIATT, S. M., THIBAUD-NISSEN, F., ASTASHYN, A., ERMOLAEVA, O., FARRELL, C. M., HART, J., LANDRUM, M. J., MCGARVEY, K. M., MURPHY, M. R., O'LEARY, N. A., PUJAR, S., RAJPUT, B., RANGWALA, S. H., RIDDICK, L. D., SHKEDA, A., SUN, H., TAMEZ, P., TULLY, R. E., WALLIN, C., WEBB, D., WEBER, J., WU, W., DICUCCIO, M., KITTS, P., MAGLOTT, D. R., MURPHY, T. D. & OSTELL, J. M. 2014. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res,* 42**,** D756-63.

PUFULETE, M., AL-GHNANIEM, R., LEATHER, A. J., APPLEBY, P., GOUT, S., TERRY, C., EMERY, P. W. & SANDERS, T. A. 2003. Folate status, genomic DNA

hypomethylation, and risk of colorectal adenoma and cancer: a case control study. *Gastroenterology,* 124**,** 1240-8.

QU, G., DUBEAU, L., NARAYAN, A., YU, M. C. & EHRLICH, M. 1999. Satellite DNA hypomethylation vs. overall genomic hypomethylation in ovarian epithelial tumors of different malignant potential. *Mutat Res,* 423**,** 91-101.

RAMZY, II, OMRAN, D. A., HAMAD, O., SHAKER, O. & ABBOUD, A. 2011. Evaluation of serum LINE-1 hypomethylation as a prognostic marker for hepatocellular carcinoma. *Arab J Gastroenterol,* 12**,** 139-42.

RANGWALA, S. H., ZHANG, L. & KAZAZIAN, H. H., JR. 2009. Many LINE1 elements contribute to the transcriptome of human somatic cells. *Genome Biol,* 10**,** R100.

RAZIN, A. & RIGGS, A. D. 1980. DNA methylation and gene function. *Science,* 210**,** 604-10.

REID, S. & TIBSHIRANI, R. 2014. Regularization Paths for Conditional Logistic Regression: The clogitL1 Package. *J Stat Softw,* 58.

RITCHIE, M. E., PHIPSON, B., WU, D., HU, Y., LAW, C. W., SHI, W. & SMYTH, G. K. 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res,* 43**,** e47.

ROBERTSON, K. D. 2001. DNA methylation, methyltransferases, and cancer. *Oncogene,* 20**,** 3139-55.

ROBIN, X., TURCK, N., HAINARD, A., TIBERTI, N., LISACEK, F., SANCHEZ, J. C. & MULLER, M. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics,* 12**,** 77.

ROCHE DIAGNOSTICS 2014. Sequencing Solutions Technical Note: How To Evaluate NimbleGen SeqCap Epi Target Enrichment Data.

RODIC, N. & BURNS, K. H. 2013. Long interspersed element-1 (LINE-1): passenger or driver in human neoplasms? *PLoS Genet,* 9**,** e1003402.

ROSE, N. R. & KLOSE, R. J. 2014. Understanding the relationship between DNA methylation and histone lysine methylation. *Biochim Biophys Acta,* 1839**,** 1362-72.

RUGGERO, D. 2013. Translational control in cancer etiology. *Cold Spring Harb Perspect Biol,* 5.

RUSIECKI, J. A., BACCARELLI, A., BOLLATI, V., TARANTINI, L., MOORE, L. E. & BONEFELD-JORGENSEN, E. C. 2008. Global DNA hypomethylation is associated with high serum-persistent organic pollutants in Greenlandic Inuit. *Environ Health Perspect,* 116**,** 1547-52.

SCHNABLE, P. S., WARE, D., FULTON, R. S., STEIN, J. C., WEI, F., PASTERNAK, S., LIANG, C., ZHANG, J., FULTON, L., GRAVES, T. A., MINX, P., REILY, A. D., COURTNEY, L., KRUCHOWSKI, S. S., TOMLINSON, C., STRONG, C., DELEHAUNTY, K., FRONICK, C., COURTNEY, B., ROCK, S. M., BELTER, E., DU, F., KIM, K., ABBOTT, R. M., COTTON, M., LEVY, A., MARCHETTO, P., OCHOA, K., JACKSON, S. M., GILLAM, B., CHEN, W., YAN, L., HIGGINBOTHAM, J., CARDENAS, M., WALIGORSKI, J., APPLEBAUM, E., PHELPS, L., FALCONE, J., KANCHI, K., THANE, T., SCIMONE, A., THANE, N., HENKE, J., WANG, T., RUPPERT, J., SHAH, N., ROTTER, K., HODGES, J., INGENTHRON, E., CORDES, M., KOHLBERG, S., SGRO, J., DELGADO, B., MEAD, K., CHINWALLA, A., LEONARD, S., CROUSE, K., COLLURA, K., KUDRNA, D., CURRIE, J., HE, R., ANGELOVA, A., RAJASEKAR, S., MUELLER, T., LOMELI, R., SCARA, G., KO, A.,

DELANEY, K., WISSOTSKI, M., LOPEZ, G., CAMPOS, D., BRAIDOTTI, M., ASHLEY, E., GOLSER, W., KIM, H., LEE, S., LIN, J., DUJMIC, Z., KIM, W., TALAG, J., ZUCCOLO, A., FAN, C., SEBASTIAN, A., KRAMER, M., SPIEGEL, L., NASCIMENTO, L., ZUTAVERN, T., MILLER, B., AMBROISE, C., MULLER, S., SPOONER, W., NARECHANIA, A., REN, L., WEI, S., KUMARI, S., FAGA, B., LEVY, M. J., MCMAHAN, L., VAN BUREN, P., VAUGHN, M. W., et al. 2009. The B73 maize genome: complexity, diversity, and dynamics. *Science,* 326**,** 1112-5.

SCOTT, A. F., SCHMECKPEPER, B. J., ABDELRAZIK, M., COMEY, C. T., O'HARA, B., ROSSITER, J. P., COOLEY, T., HEATH, P., SMITH, K. D. & MARGOLET, L. 1987. Origin of the human L1 elements: proposed progenitor genes deduced from a consensus DNA sequence. *Genomics,* 1**,** 113-25.

SLOTKIN, R. K. & MARTIENSSEN, R. 2007. Transposable elements and the epigenetic regulation of the genome. *Nat Rev Genet,* 8**,** 272-85.

SMIT, A. F., TOTH, G., RIGGS, A. D. & JURKA, J. 1995. Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. *J Mol Biol,* 246**,** 401-417.

SMIT, A. F. A., HUBLEY, R. & GREEN, P. 2013. RepeatMasker Open-4.0.

SMITH, T. F. & WATERMAN, M. S. 1981. Identification of common molecular subsequences. *J Mol Biol,* 147**,** 195-7.

SOARES, J., PINTO, A. E., CUNHA, C. V., ANDRE, S., BARAO, I., SOUSA, J. M. & CRAVO, M. 1999. Global DNA hypomethylation in breast carcinoma: correlation with prognostic factors and tumor progression. *Cancer,* 85**,** 112-8.

STEVENS, M., CHENG, J. B., LI, D., XIE, M., HONG, C., MAIRE, C. L., LIGON, K. L., HIRST, M., MARRA, M. A. & COSTELLO, J. F. 2013. Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. *Genome research,* 23**,** 1541-1553.

SU, M., HAN, D., BOYD-KIRKUP, J., YU, X. & HAN, J. D. 2014. Evolution of Alu elements toward enhancers. *Cell Rep,* 7**,** 376-85.

SUTO, R. K., CLARKSON, M. J., TREMETHICK, D. J. & LUGER, K. 2000. Crystal structure of a nucleosome core particle containing the variant histone H2A.Z. *Nat Struct Biol,* 7**,** 1121-4.

SWERGOLD, G. D. 1990. Identification, characterization, and cell specificity of a human LINE-1 promoter. *Mol Cell Biol,* 10**,** 6718-29.

SZPAKOWSKI, S., SUN, X., LAGE, J. M., DYER, A., RUBINSTEIN, J., KOWALSKI, D., SASAKI, C., COSTA, J. & LIZARDI, P. M. 2009. Loss of epigenetic silencing in tumors preferentially affects primate-specific retroelements. *Gene,* 448**,** 151-67.

TAFT, R. J., PHEASANT, M. & MATTICK, J. S. 2007. The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays,* 29**,** 288-99.

TAJUDDIN, S. M., AMARAL, A. F., FERNANDEZ, A. F., CHANOCK, S., SILVERMAN, D. T., TARDON, A., CARRATO, A., GARCIA-CLOSAS, M., JACKSON, B. P., TORANO, E. G., MARQUEZ, M., URDINGUIO, R. G., GARCIA-CLOSAS, R., ROTHMAN, N., KOGEVINAS, M., REAL, F. X., FRAGA, M. F. & MALATS, N. 2014. LINE-1 methylation in leukocyte DNA, interaction with phosphatidylethanolamine N-methyltransferase variants and bladder cancer risk. *Br J Cancer,* 110**,** 2123-30.

TANGKIJVANICH, P., HOURPAI, N., RATTANATANYONG, P., WISEDOPAS, N., MAHACHAI, V. & MUTIRANGURA, A. 2007. Serum LINE-1 hypomethylation as a potential prognostic marker for hepatocellular carcinoma. *Clin Chim Acta,* 379**,** 127-33.

TEUGELS, E., DE BRAKELEER, S., GOELEN, G., LISSENS, W., SERMIJN, E. & DE GREVE, J. 2005. De novo Alu element insertions targeted to a sequence common to the BRCA1 and BRCA2 genes. *Hum Mutat,* 26**,** 284.

THORNBURG, B. G., GOTEA, V. & MAKALOWSKI, W. 2006. Transposable elements as a significant source of transcription regulating signals. *Gene,* 365**,** 104-10.

TREANGEN, T. J. & SALZBERG, S. L. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet,* 13**,** 36-46.

TREANGEN, T. J. & SALZBERG, S. L. 2012. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat Rev Genet,* 13**,** 36-46.

TROYANSKAYA, O., CANTOR, M., SHERLOCK, G., BROWN, P., HASTIE, T., TIBSHIRANI, R., BOTSTEIN, D. & ALTMAN, R. B. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics,* 17**,** 520-5.

ULLU, E., ESPOSITO, V. & MELLI, M. 1982. Evolutionary conservation of the human 7 S RNA sequences. *J Mol Biol,* 161**,** 195-201.

VAN HOESEL, A. Q., VAN DE VELDE, C. J. H., KUPPEN, P. J. K., LIEFERS, G. J., PUTTER, H., SATO, Y., ELASHOFF, D. A., TURNER, R. R., SHAMONKI, J. M., DE KRUIJF, E. M., VAN NES, J. G. H., GIULIANO, A. E. & HOON, D. S. B. 2012. Hypomethylation of LINE-1 in primary tumor has poor prognosis in young breast cancer patients: A retrospective cohort study. *Breast Cancer Research and Treatment,* 134**,** 1103-1114.

VERT, J. P., TSUDA, K. & SCHOLKOPF, B. 2004. *Kernel Methods in Computational Biology*, MIT Press.

WALTERS, R. J., WILLIAMSON, E. J., ENGLISH, D. R., YOUNG, J. P., ROSTY, C., CLENDENNING, M., WALSH, M. D., PARRY, S., AHNEN, D. J., BARON, J. A., WIN, A. K., GILES, G. G., HOPPER, J. L., JENKINS, M. A. & BUCHANAN, D. D. 2013. Association between hypermethylation of DNA repetitive elements in white blood cell DNA and early-onset colorectal cancer. *Epigenetics,* 8**,** 748-55.

WARD, M. C., WILSON, M. D., BARBOSA-MORAIS, N. L., SCHMIDT, D., STARK, R., PAN, Q., SCHWALIE, P. C., MENON, S., LUKK, M., WATT, S., THYBERT, D., KUTTER, C., KIRSCHNER, K., FLICEK, P., BLENCOWE, B. J. & ODOM, D. T. 2013. Latent regulatory potential of human-specific repetitive elements. *Mol Cell,* 49**,** 262-72.

WEBER, L., AL-REFAE, K., EBBERT, J., JAGERS, P., ALTMULLER, J., BECKER, C., HAHN, S., GISSELMANN, G. & HATT, H. 2017. Activation of odorant receptor in colorectal cancer cells leads to inhibition of cell proliferation and apoptosis. *PLoS One,* 12**,** e0172491.

WEISENBERGER, D. J., CAMPAN, M., LONG, T. I., KIM, M., WOODS, C., FIALA, E., EHRLICH, M. & LAIRD, P. W. 2005. Analysis of repetitive element DNA methylation by MethyLight. *Nucleic Acids Res,* 33**,** 6823-36.

WENG, J., WANG, J., HU, X., WANG, F., ITTMANN, M. & LIU, M. 2006. PSGR2, a novel G-protein coupled receptor, is overexpressed in human prostate cancer. *Int J Cancer,* 118**,** 1471-80.

WICKER, T., ROBERTSON, J. S., SCHULZE, S. R., FELTUS, F. A., MAGRINI, V., MORRISON, J. A., MARDIS, E. R., WILSON, R. K., PETERSON, D. G., PATERSON, A. H. & IVARIE, R. 2005. The repetitive landscape of the chicken genome. *Genome Res,* 15**,** 126-36.

WOO, H. D. & KIM, J. 2012. Global DNA hypomethylation in peripheral blood leukocytes as a biomarker for cancer risk: a meta-analysis. *PLoS One,* 7**,** e34615.

WU, H. C., WANG, Q., YANG, H. I., TSAI, W. Y., CHEN, C. J. & SANTELLA, R. M. 2012. Global DNA methylation levels in white blood cells as a biomarker for hepatocellular carcinoma risk: a nested case-control study. *Carcinogenesis,* 33**,** 1340-5.

XIE, H., WANG, M., BONALDO MDE, F., SMITH, C., RAJARAM, V., GOLDMAN, S., TOMITA, T. & SOARES, M. B. 2009. High-throughput sequence-based epigenomic analysis of Alu repeats in human cerebellum. *Nucleic Acids Res,* 37**,** 4331-40.

XIE, H., WANG, M., DE ANDRADE, A., BONALDO MDE, F., GALAT, V., ARNDT, K., RAJARAM, V., GOLDMAN, S., TOMITA, T. & SOARES, M. B. 2011. Genome-wide quantitative assessment of variation in DNA methylation patterns. *Nucleic Acids Res,* 39**,** 4099-108.

XING, J., ZHANG, Y., HAN, K., SALEM, A. H., SEN, S. K., HUFF, C. D., ZHOU, Q., KIRKNESS, E. F., LEVY, S., BATZER, M. A. & JORDE, L. B. 2009. Mobile elements create structural variation: analysis of a complete human genome. *Genome Res,* 19**,** 1516-26.

YANG, A. S., ESTECIO, M. R., DOSHI, K., KONDO, Y., TAJARA, E. H. & ISSA, J. P. 2004. A simple method for estimating global DNA methylation using bisulfite PCR of repetitive DNA elements. *Nucleic Acids Res,* 32**,** e38.

YANG, X., HAN, H., DE CARVALHO, D. D., LAY, F. D., JONES, P. A. & LIANG, G. 2014. Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell,* 26**,** 577-90.

YU, G., WANG, L. G., HAN, Y. & HE, Q. Y. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS,* 16**,** 284-7.

ZHANG, W., SPECTOR, T. D., DELOUKAS, P., BELL, J. T. & ENGELHARDT, B. E. 2015. Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements. *Genome Biol,* 16**,** 14.

ZHANG, Z., ZHENG, Y., ZHANG, X., LIU, C., JOYCE, B. T., KIBBE, W. A., HOU, L. & ZHANG, W. 2016. Linking short tandem repeat polymorphisms with cytosine modifications in human lymphoblastoid cell lines. *Hum Genet,* 135**,** 223-32.

ZHENG, H., WU, H., LI, J. & JIANG, S. W. 2013. CpGIMethPred: computational model for predicting methylation status of CpG islands in human genome. *BMC Med Genomics,* 6 Suppl 1**,** S13.

ZHENG, Y., JOYCE, B. T., LIU, L., ZHANG, Z., KIBBE, W. A., ZHANG, W. & HOU, L. 2017a. Prediction of genome-wide DNA methylation in repetitive elements. *Nucleic Acids Research*.

ZHENG, Y., LIU, L., ZHANG, W., KIBBE, W. & HOU, L. 2017b. REMP: Repetitive Element Methylation Prediction. R package version 1.0.0. *Bioconductor*.

ZHU, Z. Z., SPARROW, D., HOU, L., TARANTINI, L., BOLLATI, V., LITONJUA, A. A., ZANOBETTI, A., VOKONAS, P., WRIGHT, R. O., BACCARELLI, A. & SCHWARTZ, J. 2011. Repetitive element hypomethylation in blood leukocyte DNA and

cancer incidence, prevalence, and mortality in elderly individuals: the Normative Aging Study. *Cancer Causes Control,* 22**,** 437–47.

APPENDIX: REFERENCE MANUAL FOR R PACKAGE REMP

RE Annotation Database Initialization

Description

`initREMP` is used to initialize annotation database for RE methylation prediction. Two major RE types in human, Alu element (Alu) and LINE-1 (L1) are available.

Usage

```
initREMP(arrayType = c("450k", "EPIC"), REtype = c("Alu", "L1"),
RE = NULL, ncore = NULL, BPPARAM = NULL, export = FALSE,
work.dir = tempdir(), verbose = FALSE)
```

Arguments

`arrayType`	Illumina methylation array type. Currently "`450k`" and "`EPIC`" are supported. Default = "`450k`".
`REtype`	Type of RE. Currently "`Alu`" and "`L1`" are supported.
`RE`	A `GRanges` object containing user-specified RE genomic location information. If NULL, the function will retrieve RepeatMasker RE database from `AnnotationHub` (build hg19).
`ncore`	Number of cores to run parallel computation. By default max number of cores available in the machine will be utilized. If `ncore = 1`, no parallel computing is allowed.
`BPPARAM`	An optional `BiocParallelParam` instance determining the parallel back-end to be used during evaluation. If not specified, default back-end in the machine will be used.
`export`	Logical. Should the returned `REMParcel` object be saved to local machine? See ***Details***.
`work.dir`	Path to the directory where the generated data will be saved. Valid when `export = TRUE`. If not specified and `export = TRUE`, temporary directory `tempdir()` will be used.
`verbose`	Logical parameter. Should the function be verbose?

Details

Currently we support two major types of RE in human, Alu and L1. The main purpose of `initREMP` is to generate and annotate CpG/RE data using the refSeq Gene annotation database (provided by `AnnotationHub`). These annotation data are crucial to RE methylation prediction in `remp`. Once generated, the data can be reused in the future (data can be very large). Therefore, we recommend user to save the output from `initREMP` to the local machine, so that user only need to run this function once as long as there is no change on the RE database. To minimize the size of resulting data file, the generated annotation data are only for REs that contain RE-CpGs with neighboring profiled CpGs. By default, the neighboring CpGs are confined within 1200 bp flanking window. This window size can be modified using `remp_options`.

Value

An `REMParcel` object containing data needed for RE methylation prediction.

See Also

See `remp` for RE methylation prediction.

Groom methylation data to fix potential data issues

Description

`grooMethy` is used to automatically detect and fix data issues including zero β value, missing value, and infinite value.

Usage

```
grooMethy(methyDat, impute = TRUE, mapGenome = FALSE, verbose =
FALSE)
```

Arguments

`methyDat`	An `RatioSet`, `GenomicRatioSet`, `DataFrame`, data.table, data.frame, or matrix of Illumina methylation data.
`impute`	If `TRUE`, K-Nearest Neighbouring imputation will be applied to fill the missing values. If the imputed value is out of the original range, mean value will be used instead. Default = `TRUE`.
`mapGenome`	Logical parameter. If `TRUE`, function will return a `GenomicRatioSet` object instead of `RatioSet`.
`verbose`	Logical parameter. Should the function be verbose?

Details

For methylation data in β value, if zero value exist, the logit transformation from beta to M value will produce negative infinite value. Therefore, zero β value will be replaced with the smallest non-zero β value found in the dataset. `grooMethy` can also handle missing value (i.e. `NA` or `NaN`) using KNN-imputation (see `impute.knn`). Infinite value will be also treated as missing value for imputation. If original dataset is in β value, `grooMethy` will first transform it to M value before imputation is carried out. Since there is possibility that KNN-imputation could produce imputed data that are not reliable (i.e. values that are out of the original data range across samples), `grooMethy` will try to replace the unreliable imputation (if any) by the average of the original methylation data across samples. Please note that `grooMethy` is also embedded in `remp` so user can run `remp` directly without explicitly running `grooMethy`.

Value

An `RatioSet` or `GenomicRatioSet` containing β value and M value of the methylation data.

Repetitive element methylation prediction

Description

`remp` is used to predict genomewide methylation levels of locus-specific repetitive elements (RE). Two major RE types in human, Alu element (Alu) and LINE-1 (L1) are available.

Usage

```
remp(methyDat, REtype = c("Alu", "L1"), parcel = NULL, work.dir
= tempdir(), groom = TRUE, win = 1000, method = c("rf",
"svmLinear", "svmRadial", "naive"), autoTune = TRUE, param =
NULL, ncore = NULL, BPPARAM = NULL, verbose = FALSE)
```

Arguments

`methyDat`	An `RatioSet`, `GenomicRatioSet`, `DataFrame`, data.table, data.frame, or matrix of methylation dataset. See ***Details***.
`REtype`	Type of RE. Currently "`Alu`" and "`L1`" are supported.
`parcel`	An `REMParcel` object containing necessary data to carry out the prediction. If `NULL`, function will search the `.rds` data file in `work.dir` exported by `initREMP` (with `export = TRUE`) or `saveParcel`.
`work.dir`	Path to the directory where the annotation data generated by `initREMP` are saved. Valid when the argument parcelis missing. If not specified, temporary directory `tempdir()` will be used. If specified, the directory path has to be the same as the one specified in `initREMP` or in `saveParcel`.
`groom`	Should the function run `grooMethy` implicitly to check and fix the data? Default = `TRUE`. If `grooMethy` has been run in advance, let `groom = FALSE` can avoid repeated check.
`win`	An integer specifying window size to confine the upstream and downstream flanking region centered on the predicted CpG in RE for prediction. Default = `1000`. See ***Details***.
`method`	Name of model/approach for prediction. Currently "`rf`" (random forest), "`svmLinear`" (SVM with linear kernel),"`svmRadial`" (SVM with linear kernel), and "`naive`" (carrying over methylation values of the closest CpG site) are available. Default = "`rf`" (random forest). Names of the machine learning models will be passed to the argument method in train (package `caret`). See ***Details***.

autoTune	Logical parameter. If TRUE, a 3-time repeated 5-fold cross validation will be performed to determine the best model parameter. If FALSE, the param option (see below) must be specified. Default = TRUE. See ***Details***.
param	A number or a vector specifying the model tuning parameter(s). For random forest, param represents 'mtry'; for SVM, param represents 'Cost' (for linear kernel) or 'Sigma' and 'Cost' (for radial basis function kernel). This parameter is valid only when autoTune = FALSE.
ncore	Number of cores to run parallel computation. By default, max number of cores available in the machine will be utilized. If ncore = 1, no parallel computing is allowed.
BPPARAM	An optional BiocParallelParam instance determining the parallel back-end to be used during evaluation. If not specified, default back-end in the machine will be used.
verbose	Logical parameter. Should the function be verbose?

Details

Before running remp, user should make sure the methylation data have gone through proper quality control, background correction, and normalization procedures. Both β value and M value are allowed. Rows represents probes and columns represents samples. Please make sure to have row names that specify the Illumina probe ID (i.e. cg00000029). Parameter win = 1000 is based on previous findings showing that neighboring CpGs are more likely to be co-modified within 1000 bp. User can specify narrower window size for slight improvement of prediction accuracy at the cost of less predicted RE. Window size greater than 1000 is not recommended as the machine learning models would not be able to learn much useful information for prediction but introduce noise. Random Forest model (method = "rf") is recommended as it offers more accurate prediction and it also enables prediction reliability functionality. Prediction reliability is estimated by conditional standard deviation using Quantile Regression Forest (see package quantregForest for more details). Please note that if parallel computing is allowed, parallel random forest (parRF in package caret) will be used automatically. If autoTune = TRUE, preset tuning parameter search grid can be access and modified using remp_options.

Value

An REMProduct object containing prediction results.

See Also

See initREMP to prepare necessary annotation database before running remp.

REMParcel instances

Description

`REMParcel` is a container class to organize required datasets for RE methylation prediction generated from `initREMP` and used in `remp`.

Usage

```
REMParcel(REtype = "Unknown", platform = "Unknown", RefGene =
GRanges(), RE = GRanges(), RECpG = GRanges(), ILMN = GRanges())

getRefGene(object)
getRE(object)
getRECpG(object)
getILMN(object, ...)
saveParcel(object, ...)

## S4 method for signature 'REMParcel'
saveParcel(object, work.dir = tempdir(),
  verbose = FALSE, ...)

## S4 method for signature 'REMParcel'
getRefGene(object)

## S4 method for signature 'REMParcel'
getRE(object)

## S4 method for signature 'REMParcel'
getRECpG(object)

## S4 method for signature 'REMParcel'
getILMN(object, REonly = FALSE)
```

Arguments

`REtype`	Type of RE ("`Alu`" or "`L1`").
`platform`	Illumina methylation profiling platform ("`450k`" or "`EPIC`").

`RefGene`	refSeq gene annotation data, which can be obtained by `fetchRefSeqGene`.
`RE`	Annotated RE genomic range data, which can be obtained by `fetchRMSK` and annotated by `GRannot`.
`RECpG`	Genomic range data of annotated CpG site identified in RE DNA sequence, which can be obtained by `findRECpG` and annotated by `GRannot`.
`ILMN`	Illumina CpG probe genomic range data.
`object`	A `REMParcel` object.
`...`	For `saveParcel`: other parameters to be passed to the `saveRDS` method. See `saveRDS`.
`work.dir`	For `saveParcel`: path to the directory where the generated data will be saved. If not specified, temporary directory `tempdir()` will be used.
`verbose`	For `saveParcel`: logical parameter. Should the function be verbose?
`REonly`	For `getILMN`: see ***Accessors***.

Value

An object of class `REMParcel` for the constructor.

Accessors

`getRefGene(object)`

Return RefSeq gene annotation data.

`getRE(object)`

Return RE genomic location data for prediction (annotated by refSeq gene database).

`getRECpG(object)`

Return RE-CpG genomic location data for prediction.

`getILMN(object, REonly = FALSE)`

Return Illumina CpG probe genomic location data for prediction (annotated by refSeq gene database). If `REonly = TRUE`, only probes within RE region are returned.

Utilities

`saveParcel(object, work.dir = tempdir(), verbose = FALSE, ...)`

Save the object to local machine.

REMProduct instances

Description

Class `REMProduct` is to maintain RE methylation prediction results. `REMProduct` inherits Bioconductor's `RangedSummarizedExperiment` class.

Usage

```
REMProduct(REtype = "Unknown", platform = "Unknown", win =
"Unknown", predictModel = "Unknown", QCModel = "Unknown", rempM
= NULL, rempB = NULL, rempQC = NULL, cpgRanges = GRanges(),
sampleInfo = DataFrame(), REannotation = GRanges(), RECpG =
GRanges(), regionCode = DataFrame(), refGene = GRanges(), varImp
= DataFrame(), REStats = DataFrame(), GeneStats = DataFrame())

rempM(object)
rempB(object)
rempQC(object)
annotation(object)
imp(object)
stats(object)
details(object)
decodeAnnot(object, ...)
trim(object, ...)

## S4 method for signature 'REMProduct'
rempM(object)

## S4 method for signature 'REMProduct'
rempB(object)

## S4 method for signature 'REMProduct'
rempQC(object)

## S4 method for signature 'REMProduct'
imp(object)

## S4 method for signature 'REMProduct'
```

```
annotation(object)

## S4 method for signature 'REMProduct'
stats(object)

## S4 method for signature 'REMProduct,missing'
plot(x, type = c("individual", "overall"), ...)

## S4 method for signature 'REMProduct'
details(object)

## S4 method for signature 'REMProduct'
decodeAnnot(object, type = c("symbol", "entrez"), ncore = NULL,
BPPARAM = NULL)

## S4 method for signature 'REMProduct'
trim(object, threshold = 1.7, missingRate = 0.2)
```

Arguments

`REtype`	Type of RE ("`Alu`" or "`L1`").
`platform`	Illumina methylation profiling platform ("`450k`" or "`EPIC`").
`win`	Flanking window size of the predicting RE-CpG.
`predictModel`	Name of the model used for prediction.
`QCModel`	Name of the model used for prediction quality evaluation.
`rempM`	Predicted methylation level in M value.
`rempB`	Predicted methylation level in β value (optional).
`rempQC`	Prediction quality scores, which is available only when Random Forest model is used in `remp`.
`cpgRanges`	Genomic ranges of the predicting RE-CpG.
`sampleInfo`	Sample information.
`REannotation`	Annotation data for the predicting RE.
`RECpG`	Annotation data for the RE-CpG profiled by Illumina platform.

`regionCode`	Internal index code defined in refSeq Gene database for gene region indicators.
`refGene`	refSeq gene annotation data, which can be obtained by `fetchRefSeqGene`.
`varImp`	Importance of the predictors.
`REStats`	RE coverage statistics, which is internally generated in `remp`.
`GeneStats`	Gene coverage statistics, which is internally generated in `remp`.
`object`	A `REMProduct` object.
`...`	For plot: `graphical parameters` to be passed to the plot method.
`x`	For plot: an `REMProduct` object.
`type`	For plot and `decodeAnnot`: see ***Utilities***.
`ncore`	For `decodeAnnot`: number of cores to run parallel computation. By default max number of cores available in the machine will be utilized. If `ncore = 1`, no parallel computing is allowed (not recommended).
`BPPARAM`	For `decodeAnnot`: an optional `BiocParallelParam` instance determining the parallel back-end to be used during evaluation. If not specified, default back-end in the machine will be used.
`threshold`	For `trim`: see ***Utilities***.
`missingRate`	For `trim`: see ***Utilities***.

Value

An object of class `REMProduct` for the constructor.

Accessors

`rempM(object)`

Return M value of the prediction.

`rempB(object)`

Return β value of the prediction.

`rempQC(object)`

Return prediction quality scores.

`imp(object)`

Return relative importance of predictors.

`stats(object)`

Return RE and gene coverage statistics.

`annotation(object)`

Return annotation data for the predicted RE.

Utilities

`plot(x, type = c("individual", "overall"), ...)`

Make a density plot of predicted methylation (β values) in the `REMProduct` object x. If type = "individual", density curves will be plotted for each of the samples; If type = "overall", one density curve of the mean methylation level across the samples will be plotted. Default type = "`individual`".

`details(object)`

Display detailed descriptive statistics of the prediction results.

`decodeAnnot(object, type = c("symbol", "entrez")), ncore = NULL, BPPARAM = NULL`

Decode the RE annotation data by Gene Symbol (when type = "Symbol") or Entrez Gene (when type = "`Entrez`"). Default type = "`Symbol`". Annotation data are provided by `org.Hs.eg.db`.

`trim(object, threshold = 1.7, missingRate = 0.2)`

Any predicted CpG values with quality score < threshold (default = `1.7`) will be replaced with NA. CpGs contain more than `missingRate * 100` will be re-evaluated.

Get RefSeq gene database

Description

`fetchRefSeqGene` is used to obtain refSeq gene database provided by `AnnotationHub`.

Usage

```
fetchRefSeqGene(ah, mainOnly = FALSE, verbose = FALSE)
```

Arguments

`ah`	An `AnnotationHub` object. Use `AnnotationHub()` to retrieve information about all records in the hub.
`mainOnly`	Logical parameter. See ***details***.
`verbose`	Logical parameter. Should the function be verbose?

Details

When `mainOnly = FALSE`, only the transcript location information will be returned, Otherwise, a `GRangesList` object containing gene regions information will be added. Gene regions include: 2000 base pair upstream of the transcript start site (`$tss`)), 5'UTR (`$fiveUTR`)), coding sequence (`$cds`)), exon (`$exon`)), and 3'UTR (`$threeUTR`)).
The index column is an internal index that is used to facilitate data referral, which is meaningless for external use.

Value

A single `GRanges` (for main refSeq Gene database) object or a list incorporating both `GRanges` object (for main refSeq Gene database) and `GRangesList` object (for gene regions data).

Get RE database from RepeatMasker

Description

`fetchRMSK` is used to obtain specified RE database from RepeatMasker Database provided by `AnnotationHub`.

Usage

```
fetchRMSK(ah, REtype, verbose = FALSE)
```

Arguments

`ah`	An `AnnotationHub` object. Use `AnnotationHub()` to retrieve information about all records in the hub.
`REtype`	Type of RE. Currently "`Alu`" and "`L1`" are supported.
`verbose`	Logical parameter. Should the function be verbose?

Value

A `GRanges` object containing RE database. `'name'` column indicates the RE subfamily; `'score'` column indicate the SW score; `'Index'` is an internal index for RE to facilitate data referral, which is meaningless for external use.

Find RE-CpG genomic location given RE ranges information

Description

`findRECpG` is used to obtain RE-CpG genomic location data.

Usage

```
findRECpG(RE.hg19, REtype = c("Alu", "L1"), be = NULL, verbose = FALSE)
```

Arguments

`RE.hg19`	an `GRanges` object of RE genomic location database. This can be obtained by `fetchRMSK`.
`REtype`	Type of RE. Currently "`Alu`" and "`L1`" are supported.
`be`	A `BiocParallel` object containing back-end information that is ready for parallel computing. This can be obtained by `getBackend`.
verbose	logical parameter. Should the function be verbose?

Details

CpG site is defined as 5'-C-p-G-3'. It is reasonable to assume that the methylation status across all CpG/CpG dyads are concordant. Maintenance methyltransferase exhibits a preference for hemimethylated CpG/CpG dyads (methylated on one strand only). As a result, methylation status of CpG sites in both forward and reverse strands are usually consistent. Therefore, to accommodate the cytosine loci in both strands, the returned genomic ranges cover the 'CG' sequence with width of 2. The 'stand' information indicates the strand of the RE.

Value

A `GRanges` object containing identified RE-CpG genomic location data.

Get BiocParallel back-end

Description

`getBackend` is used to obtain `BiocParallel` Back-end to allow parallel computing.

Usage

```
getBackend(ncore, BPPARAM = NULL, verbose = FALSE)
```

Arguments

`ncore`	Number of cores to run parallel computation. By default, max number of cores available in the machine will be utilized. If `ncore = 1`, no parallel computing is allowed.
`BPPARAM`	An optional `BiocParallelParam` instance determining the parallel back-end to be used during evaluation. If not specified, default back-end in the machine will be used.
`verbose`	Logical parameter. Should the function be verbose?

Value

A `BiocParallel` object that can be used for parallel computing.

Get methylation data of GM12878 profiled by Illumina 450k array or EPIC array

Description

`getGM12878` is used to obtain public available methylation profiling data of HapMap LCL sample GM12878.

Usage

```
getGM12878(arrayType = c("450k", "EPIC"), mapGenome = FALSE)
```

Arguments

`arrayType`	Illumina methylation array type. Currently "450k" and "EPIC" are supported. Default = "`450k`".
`mapGenome`	logical parameter. If `TRUE`, function will return a `GenomicRatioSet` object instead of a `RatioSet` object.

Details

Illumina 450k data were sourced and curated from ENCODE http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibMethyl450/wgEncodeHaibMethyl450Gm12878SitesRep1.bed.gz. Illumina EPIC data were obtained from data package `minfiDataEPIC`.

Value

An `RatioSet` or `GenomicRatioSet` containing β value and M value of the methylation data.

Annotate genomic ranges data with gene region information.

Description

`GRannot` is used to annotate a `GRanges` dataset with gene region information using refSeq Gene database

Usage

```
GRannot(object.GR, refgene.hg19, verbose = FALSE)
```

Arguments

`object.GR`	An `GRanges` object of a genomic location database.
`refgene.hg19`	A complete refSeq Gene database returned by `fetchRefSeqGene` (with parameter `mainOnly = FALSE`).
`verbose`	Logical parameter. Should the function be verbose?

Details

The annotated gene region information includes: protein coding gene (`InNM`), noncoding RNA gene (`InNR`), 2000 base pair upstream of the transcript start site (`InTSS`), 5'UTR (`In5UTR`), coding sequence (`InCDS`), exon (`InExon`), and 3'UTR (`In3UTR`). The intergenic and intron regions can then be represented by the combination of these region data. The number shown in these columns represent the row number or `'index'` column in the main refSeq Gene database obtained by `fetchRefSeqGene`.

Value

A `GRanges` or a `GRangesList` object containing refSeq Gene database.

Configure options for REMP package

Description

Tools to manage global setting options for `REMP` package.

Usage

```
remp_options(...)
remp_reset()
```

Arguments

`...`	Option names to retrieve option values or [key]=[value] pairs to set options.

Value

`NULL`

Supported options

The following options are supported

`.default.AluFamily`

A list of Alu subfamily to be included in the prediction.

`.default.L1Family`

A list of L1 subfamily to be included in the prediction.

`.default.GM12878.450k.URL`

URL to download GM12878 450k methylation profiling data.

`.default.AH.repeatmasker.hg19`

`AnnotationHub` data ID linked to RepeatMasker database (build hg19)

`.default.AH.refgene.hg19`

`AnnotationHub` data ID linked to refSeq gene database (build hg19)

`.default.TSS.upstream`

Define the upstream range of transcription start site region.

`.default.TSS.downstream`

Define the downstream range of transcription start site region.

`.default.max.flankWindow`

Define the max size of the flanking window surrounding the predicted RE-CpG.

`.default.450k.annotation`

A character string associated with the Illumina 450k array annotation dataset.

`.default.epic.annotation`

A character string associated with the Illumina EPIC array annotation dataset.

`.default.genomicRegionColNames`

Define the names of the genomic regions for prediction.

`.default.predictors`

Define the names of predictors for RE methylation prediction.

`.default.mtry.tune`

Define the default mtry parameter for Random Forest model.

`.default.C.svmLinear.tune`

Define the default C (Cost) parameter for Support Vector Machine (SVM) using linear kernel.

`.default.sigma.svmRadial.tune`

Define the default sigma parameter for SVM using Radial basis function kernel.

`.default.C.svmRadial.tune`

Define the default C (Cost) parameter for SVM using Radial basis function kernel.