ลักษณะทางจีโนมของการแสดงออกของยีนที่มีไลน์-๑ โดยโปรตีนอาร์โกนอต

นายชุมพล งามผิว

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต
สาขาวิชาชีวเวชศาสตร์ (สหสาขาวิชา)
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2558

GENOMIC CHARACTERISTICS OF ARGONAUTE PROTEINS MEDIATE GENES CONTAINING LINE-1 EXPRESSION

Mr. Chumpol Ngamphiw

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in Biomedical Sciences
(Interdisciplinary Program)
Graduate School
Chulalongkorn University
Academic Year 2015

Thesis Title	GENOMIC CHARACTERISTICS OF ARGONAUTE PROTEINS MEDIATE GENES CONTAINING LINE-1 EXPRESSION
By	Mr. Chumpol Ngamphiw
Field of Study	Biomedical Sciences
Thesis Advisor	Professor Apiwat Mutirangura, M.D., Ph.D.
Thesis Co-Advisor	Sissades Tongsima, Ph.D.

Accepted by the Graduate School, Chulalongkorn University in Partial Fulfillment of the Requirements for the Doctoral Degree

---Dean of the Graduate School

(Associate Professor Sunait Chutintaranond, Ph.D.)

THESIS COMMITTEE

---Chairman

(Professor Prasit Pavasant, D.D.S., Ph.D.)

---Thesis Advisor

(Professor Apiwat Mutirangura, M.D., Ph.D.)

---Thesis Co-Advisor

(Sissades Tongsima, Ph.D.)

---Examiner

(Assistant Professor Viroj Boonyaratanakornkit, Ph.D.)

---Examiner

(Trairak Pisitkun, M.D.)

---External Examiner

(Apichart Intarapanich, Ph.D.)

ชุมพล งามผิว : ลักษณะทางจีโนมของการแสดงออกของยีนที่มีไลน์-๑ โดยโปรตีนอาร์โกนอต (GENOMIC CHARACTERISTICS OF ARGONAUTE PROTEINS MEDIATE GENES CONTAINING LINE-1 EXPRESSION) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ศ. นพ. ดร.อภิวัฒน์ มุทิรางกูร, อ.ที่ปรึกษาวิทยานิพนธ์ร่วม: ดร.ศิษเฎศ ทองสิมา, 57 หน้า.

ปัจจุบันมีข้อมูลสนับสนุนมากขึ้นว่าการแสดงออกของไลน์-๑มีการทำงานร่วมกับส่วนประกอบทางพันธุกรรมอื่นๆ ในการแสดงบทบาทต่าง ๆ ตลอดช่วงวิวัฒนาการในสัตว์เลี้ยงลูกด้วยนม เช่น ยับยั้งการแสดงออกของยีนหนึ่งชุดในโครโมโซมเอ็กซ์ เกี่ยวข้องกับกระบวนการแลกเปลี่ยนชิ้นส่วนดีเอ็นเอในโครโมโซมที่เข้าคู่กัน และควบคุมการแสดงออกของยีน เป็นต้น วิทยานิพนธ์ฉบับนี้ได้ใช้วิธีทางสถิติและชีวสารสนเทศเพื่อศึกษาความสัมพันธ์ของไลน์-๑ที่มีระดับเมทิลเลชั่นลดลงซึ่งอาจไปควบคุมการแสดงออกของยีนได้ โดยใช้ข้อมูลไมโครอะเรย์จากฐานข้อมูลสาธารณะ Gene Expression Omnibus (GEO) ร่วมกับข้อมูลไลน์-๑ จากฐานข้อมูล L1Base จากผลการวิเคราะห์ทางสถิติพบว่ายีนที่มีการแสดงออกลดลงมีไลน์-๑ อยู่อย่างมีนัยสำคัญทั้งในมะเร็งชนิดต่างๆและเซลล์ปรกติที่ถูกทำให้ระดับเมทิลเลชั่นลดลง ยิ่งไปกว่านั้นพบว่าโปรตีนอาร์โกนอต-๒น่าจะมีการทำงานร่วมกันกับไลน์-๑ โดยมีความสัมพันธ์สอดคล้องกับการยับยั้งการแสดงออกของยีนในมะเร็ง ดังนั้นไลน์-๑ที่อยู่ในยีนและมีระดับเมทิลเลชั่นลดลงน่าจะมีการทำหน้าที่ลักษณะเดียวกับ siRNA ร่วมกับส่วนประกอบทางพันธุกรรมอื่นที่อยู่ใกล้กันส่งผลให้ยับยั้งการแสดงออกของยีนได้

ยิ่งกว่านั้นเราพบว่าลำดับเบสของไลน์-๑ ที่อยู่ในยีนมีการอนุรักษ์ให้มีลักษณะเหมือนเดิมมากกว่า เพื่อรักษาความสามารถในการแสดงออกทั้งในมนุษย์และหนู ไลน์-๑ที่อยู่ในยีนของหนูจะมีจำนวนโมโนเมอร์ซึ่งอยู่ในส่วน 5' UTR มากกว่า ซึ่งแสดงว่าโมโนเมอร์สำคัญในการควบคุมการแสดงออกของไลน์-๑ เราได้เปรียบเทียบการกระจายตัวของไลน์-๑ ในสิ่งมีชีวิตทั้งสองชนิดทั้งในออโตโซม โครโมโซมเอ็กซ์และโครโมโซมวาย พบว่ามีปริมาณหนาแน่นสุดในโครโมโซมเอ็กซ์ซึ่งบอกเป็นนัยถึงหน้าที่ในกระบวนการยับยั้งการแสดงออกของยีนในโครโมโซมเอ็กซ์ และท้ายสุดพบว่าไลน์-๑ที่อยู่ในยีนมีความสัมพันธ์อย่างมีนัยสำคัญกับการยับยั้งการแสดงออกของยีนในกระบวนการพัฒนาก่อนเป็นตัวอ่อนในสิ่งมีชีวิตทั้งสองด้วย ซึ่งน่าจะคล้ายกับกลไกควบคุมการแสดงออกของยีนในเซลล์มะเร็ง

สาขาวิชา ชีวเวชศาสตร์ ลายมือชื่อนิสิต ....

ปีการศึกษา 2558 ลายมือชื่อ อ.ที่ปรึกษาหลัก ....

ลายมือชื่อ อ.ที่ปรึกษาร่วม ....

5387842120 : MAJOR BIOMEDICAL SCIENCES

KEYWORDS: AGONAUTE PROTEINS / EMBRYOGENESIS / GENE SUPPRESSION / LINE-1 HYPOMETHYLATION / X-INACTIVATION

CHUMPOL NGAMPHIW: GENOMIC CHARACTERISTICS OF ARGONAUTE PROTEINS MEDIATE GENES CONTAINING LINE-1 EXPRESSION. ADVISOR: PROF. APIWAT MUTIRANGURA, M.D., Ph.D., CO-ADVISOR: SISSADES TONGSIMA, Ph.D., 57 pp.

There is increasing evidence that transcriptionally active LINE-1 (L1) could have been co-opted through mammalian genomes evolution to play various roles including X-inactivation, homologous recombination and gene regulation. In this dissertation, both statistical and bioinformatic methods were used to evaluate intragenic L1 hypomethylation, which may influences their host gene expressions. All experiments utilized array data from the Gene Expression Omnibus (GEO) database and L1 information from L1Base. Genome-wide statistical analysis between genes containing L1s and their corresponding expression profile showed that these genes are likely to be repressed in both cancer and hypomethylated normal cells. Furthermore, AGO2 potentially forms a complex with intronic L1 pre-mRNA that correlates with the down-regulation of cancer genes. Thus, hypomethylated intragenic L1s could act as a nuclear siRNA mediated *cis*-regulatory element that can repress genes.

Moreover, we found that intragenic L1 sequences have been conserved across evolutionary time with respect to transcriptional activity in human and mouse. The monomers located in the 5' UTR of mouse L1 (more monomers found more in *intragenic* regions of the host genome) suggesting their important role for controlling L1 expression. We then compared L1 distributions of both species across autosomes, X and Y chromosomes. The results agreeably reveal that L1 densities of both species located much denser in the X-chromosome, suggestive of X-inactivation role. A significant correlation was demonstrated between the presence of intragenic L1s and downregulated genes in the early embryogenesis of both species, suggestive of a similar role in regulating genes in cancers.

Field of Study: Biomedical Sciences   Student's Signature ....

Academic Year: 2015   Advisor's Signature ....

Co-Advisor's Signature ....

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my advisor, Prof. Apiwat Mutirangura, and my co-advisor, Dr. Sissades Tongsima for all their inspiration, guidance, suggestions, helps, and encouragement throughout my doctoral study. I also thank all members from the Biostatistics and Informatics Laboratory, Genome Technology Research Unit, National Center for Genetic Engineering and Biotechnology (BIOTEC) for their encouragement to complete the thesis. This research was supported by the 2011 Research Chair Grant from the National Science and Technology Development Agency (NSTDA) and the Center of Excellence in Molecular Genetics of Cancer and Human Diseases, Chulalongkorn University.

Finally, I am so thankful to my family, especially my mother, Mrs. Woranuch Ngamphiw and my late father, Mr. Payom Ngamphiw for their love, support and be there when I needed.

CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

LINE-1, L1	Long Interspersed Element-1
LOH	Loss of heterozygosity
GEO	Gene Expression Omnibus database
siRNA	small interfering RNA
AGO	Argonaute protein
RISC	RNA-induced silencing complex
Aza-dC	5-Aza-2'-deoxycytidine
ORF	Open reading frame
UTR	Untranslated region
CpG	–C–phosphate–G–
dsRNA	double-stranded RNA
RNAi	RNA interference
mRNA	messenger RNA
ZGA	Zygotic gene activation
Alu	Arthrobacter luteus restriction endonuclease
NCBI	National Center for Biotechnology Information
L1PA	primate L1
L1M	mammalian L1
MH chi-square	Mantel-Haenszel chi-square
OR	Odds ratio
HEK-293	Human Embryonic Kidney 293 cells
RBP	RNA-binding protein
CLIP	crosslinking and immunoprecipitation
hBECs	human bronchial epithelial cells
hMSCs	human mesenchymal stem cells
LIDs	L1 insertion dimorphisms

CHAPTER I

INTRODUCTION

Background and rationale

The Long INterspersed Element-1 (LINE-1 or L1) is a transposable DNA element, which amplifies itself in a genome by utilizing the reverse transcription process similar to that of retrovirus when integrating itself to the host genome. L1 DNAs are transcribed to RNAs, which later are converted to cDNA format using the underlying reverse transcriptase mechanism. These L1s enter a nuclease and integrate themselves into new genomic loci, causing the host genome to get bigger. Although L1s account for 18-20% of mammalian genomes [1, 2], most of them do not have retrotransposition activity owing to truncations in 5' regions, rearrangements, or mutations [3, 4]. However, active L1 retrotransposition is an important evolutionary driver of mammalian genome complexity, and is responsible for many heritable disorders [5].

Earlier L1s were thought to be selfish DNA elements in which their only function was to replicate and cause host genome to expand [5]. However, through an evolutionary process depending on the genomic context where they have inserted, these L1s may have acquired other important functions, such as spreading of X-inactivation [6-8], controlling of gene expression (acting as a *cis*-regulatory element), differentiating cells and repairing DNA [9, 10]. Interestingly, scientists also found that intragenic L1s are transcriptionally active during embryogenesis [11] and in cancer cells [12] as a result of hypomethylation. It is still unclear whether these special functions are conserved among mammalian genomes.

DNA methylation is a fundamental molecular characteristic of the human genome, and alteration of this epigenetic regulation is associated with cancers [13]. The effects of promoter methylation on chromatin configuration and gene transcription have been well documented [14]. Particularly, in cancer cells, global hypomethylation relative to the methylation level in normal cells is commonly observed [13, 15, 16]. Furthermore, hypomethylation markedly affects those loci with moderate to high-copy repeats [17]. This epigenetic phenomenon could play several significant roles in multistep tumorigenesis. Most commonly recognized effect of this hypomethylation event is the promotion of chromosomal instability [18], increasing a risk of chromosome breakage and recombination including high rate of loss of heterozygosity (LOH) [19, 20]. Recent report reveals that hypomethylation of a L1 activates alternate promoter of *MET* oncogene [21]. Due to hypomethylation is very common on those repeating sequences, the role of global L1 hypomethylation affecting gene expression profile has not yet been fully explained. Gene body methylation (DNA methylation within a gene) changes can certainly affect transcription of various transposable elements. In addition, unique methylated sequences in introns are frequently found in highly expressed genes [22]. Such methylation profile could hinder expression of these genes by forming heterochromatin that limits the efficiency of RNA polymerase [23]. These evidences implied that methylation of intragenic repetitive sequences may play roles in gene regulation.

L1s are randomly inserted across genome and are likely to skip those housekeeping genes [24]. A report in [25] shows that the insertion of active L1 sequences into introns of host-gene could lower gene expression. Methylation levels of L1s around 5' UTR were also reported to vary at each locus across different cell types [26]. This supports the notion that L1s could play a role in regulating gene expression. To study human or mouse L1s, researchers can now access the comprehensive L1base database [27, 28] that stores locations of L1s and their annotations. On the other hand, high quality microarray gene expression data under different methylation conditions are also freely available from the Gene Expression Omnibus (GEO) database [29, 30]. Thus computational-based experiments can be then formulated to statistically test whether the existence of L1 elements associate with expression of genes.

Although mammalian L1s are similar in nature, they are not identical. L1s from different mammalian species could regulate genes with different mechanisms. In the second part of this thesis, comparative studies between human and mouse

L1s are conducted. L1s were classified into two classes according to their locations in genome: intragenic (within the body of a gene model) and intergenic (outside the body of a gene model). These L1s are counted collectively from within autosomes, X chromosome and Y chromosome, respectively. Then, the differences between the distributions of mouse and human L1s over these chromosomal regions can be compared. Note that human and mouse L1s are different majorly in their 5' UTR regions where monomers uniquely present in mouse L1s and are involved in the mouse L1 transcriptional activities [4]. Hence, mouse L1s may make use of these monomers to further regulate gene expression. Knowing that epigenetics is heavily involved in mammalian embryogenesis, we want to perform statistical test if L1s from both human and mouse potentially play a role in regulating genes during the early stage of embryogenesis. Chi-square statistics is performed to check correlation between L1s and the expression of their host genes from both human and mouse using embryo gene expression data at different key stages, including zygote (one-cell), two-cell, four-cell, eight-cell, morula, and blastocyst.

Research questions

1. Could human intragenic L1s repress host genes in global hypomethylated environment? What are specific L1 characteristics that differentiate intragenic and intergenic L1s?
2. Could human L1s act like siRNA to repress expression of gene by forming RNA-induced silencing complex with AGO proteins?
3. What are the different L1 characteristics between mouse and human L1s? How many L1 characteristics are shared between the two species in terms of intra- and intergenic regions?
4. Do mammalian L1s regulate genes during early stage of embryogenesis?

Objectives

1. To statistically show human intragenic L1s associated with gene regulation and propose novel mechanism of gene regulation via collaborated function of AGO proteins and L1s.
2. To classify different characteristics between intragenic and intergenic L1s in human.
3. To classify different L1s' characteristics in mouse and compare distribution of L1s in both species
4. To statistically show L1s can regulate genes in early embryogenesis in both species other than down regulate genes in cancer.

Hypotheses

1. Human L1s can down regulate their host genes in global hypomethylated environment.
2. Gene regulation function of Intragenic L1s works as siRNAs where the underlying transcribed L1 RNAs are bounded by AGO proteins in nuclease to form RISC to silence the genes.
3. Both mouse and human are mammal; hence their L1 elements should be conserved.
4. Mammalian L1s regulate genes during early stage of embryogenesis.

Key words

Agonaute proteins; Early embryogenesis; Gene suppression; Long INterspersed Element-1; LINE-1; Hypomethylation; X-inactivation

Expected results

1. The statistical results performed on genes possessing L1s and hypomethylation in cancers show the association between hypomethylation L1 can repress host genes in many cancers.

2. The association analysis will show the presence of AGO1, AGO2, AGO3 and AGO4 can affect on gene expression in either direction up- or down-regulated.
3. The association analysis will show the presence of AGO bound on genes possessing L1s are more likely to regulated genes. Furthermore the analysis on AGO binding site locations and binding site sequences related to L1s will lead to yield novel mechanisms how AGO and L1s regulate gene expression.
4. Human intragenic L1s should be more conserve than intergenic LINE-1s for controlling host gene regulation.
5. Mammalian L1s should be conserved then mouse intragenic L1s characteristics should be more conserved similar in human for control gene regulation.
6. Intragenic L1s can control gene regulation in early stages of embryogenesis in both human and mouse.

Conceptual framework

Research question I : Could human intragenic L1s repress host genes in global hypomethylated environment? What are specific L1 characteristics that differentiate intragenic and intergenic L1s?

Experiment I :

- Using chi-square test to check with GEO microarray data if human L1s appear with statistical significant in those down-regulated genes under the global DNA hypomethylation condition when compared with normal, namely Aza-dC treated vs normal and cancer vs normal microarray experiments.
- Group human L1s into intragenic (within gene body) and intergenic (between genes) groups, and perform statistical comparisons of L1 key characteristics [27] from different groups by using chi-square and Student's *t*-test to determine their characteristic differences.

Research question II : Could human L1s act like siRNA to repress expression of gene by forming RNA-induced silencing complex with AGO proteins?

Experiment II :

- Assuming that AGO proteins are needed in this process, we need to test if presence of AGO proteins are associated with repressed genes containing human L1s. Chi-square statistics is used to test whether genes containing human L1s increase the expression level on the AGO knock down microarray experiment.
- To support that AGO proteins could form complexes with L1s, we perform chi-square test to find significant association between those genes with increased expression levels (containing L1s in the genes) and the reported AGO binding sites from CLIPz database.

Research question III : What are the different L1 characteristics between mouse and human L1s? How many L1 characteristics are shared between the two species in terms of intra- and intergenic regions?

Experiment III :

- Characterize mouse L1s into intragenic (within gene body) and intergenic (between genes) groups, and perform statistical comparisons of L1 key characteristics [28] from different groups by using chi-square and Student's *t*-test to determine their characteristic differences as well as human L1s.
- Compare the distribution of L1s between the two species across three groups: autosomal, X and Y chromosomes. Also compare the distributions in terms of intra vs intergenic categories.

Research question IV : Do mammalian L1s regulate genes during early stage of embryogenesis?

Experiment IV :

- For both human and mouse L1s, use chi-square to test association between present of L1s and the regulated genes between the starting stage, 1-cell and the other stages of embryogenesis: 2-cell vs. 1-cell, 4-cell vs. 1-cell, 8-cell vs. 1-cell, morula vs. 1-cell and blastocyst vs. 1-cell, respectively.
- Cross check with known genes that are active during early embryogenesis if the above genes with L1s belong to these active genes for both human and mouse L1s.

CHAPTER II

REVIEW OF RELATED LITERATURE

Background of Long INterspersed Element-1

The Long INterspersed Element-1 (LINE-1 or L1) is a retrotransposable element that has recently been widely studied in mammal. Such transposons were integrated to eukaryotic genomes since ancient time. Mammalian L1s are accounted for approximately 18–20% in the host genome. In particular, there are about 500000 copies of L1 in human genome while the mouse genome has almost 600000 copies [1, 2]. A full-length mammalian L1 is approximately 6000–8000 nucleotides. The size of a full-length human L1 (~6000 nucleotides) is shorter than that of mouse (~7000 nucleotides). Full-length L1s contains two open reading frames encoding proteins essential for retrotransposition, a smaller ORF1 and a larger ORF2 separated by ~60 bp intergenic region (Figure 1). Both ORFs are demarcated by 5' UTR and 3' UTR with a poly-A tail [31, 32].

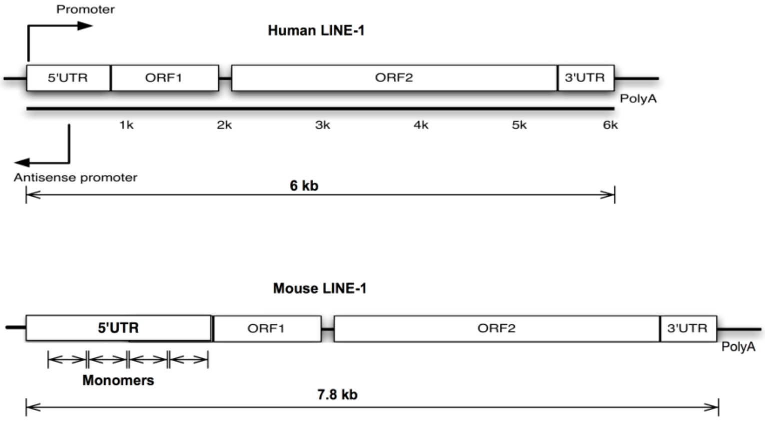


Figure 1 The structure of full-length L1s in human and mouse

There is a great deal of similarities between both ORFs of mouse and human L1s. The major difference of L1s between the two species is on the 5' UTR. Severynse and colleagues [4] proposed that the underlying differences in 5' UTR vary level of transcriptional activities between the two species. In particular, the 5' UTRs of human L1s house two internal promoters, sense and antisense promoters [3]. On the contrary, the 5' UTRs of mouse L1s contains ~200 bp repeated sequences, called monomers [4, 33]. The numbers of these monomers differ among mouse L1 families, which also affect the varying promoter activities in mouse. It has been reported that increasing monomer copies can raise the underlying transcriptional activity [4, 34].

Although L1s are abundant in human and mouse genomes, most of them lack of retrotransposition activity due to truncations in 5' regions, rearrangements, or mutations [3, 4]. According to L1base [27], there are almost 12000 full-length (>4500 nucleotides) human L1s, but only 145 of these are considered as potentially active. In contrast, full-length (>5000 nucleotides) L1s are more numerous in mouse, and the fraction of potentially active elements is considerably higher (16000 and 2382, respectively). Although most L1s are inactive, active L1 retrotransposition is an important evolutionary driver of mammalian genome complexity, and is responsible for heritable disorders [5].

It has been shown that L1s have several important roles in many high-level organisms. Considerable numbers of research works have reported several important activities of L1s, which include spreading of X-inactivation, retrotransposition activity, and controlling gene expression via a mechanism similar to that of siRNAs. X-inactivation or lyonization controls the amount of gene product from X chromosome. L1s were shown to locate more on X chromosome than on the other chromosomes indicating the special role of L1s [6, 7]. The genes that are inactivated via gene silencing mechanism are consequently richer in L1s than those genes that do escape the X-inactivation [8]. This suggests that L1s may associate with the spreading of X-inactivation; however, the data was not sufficient to deduce that L1s alone were involved in the propagation from X-inactivation center (Xic) [8].

Furthermore, L1s play important roles along the path of mammalian genome evolution and possess several critical physiological functions of the cells that may be similar among different organisms. L1s function as retrotransposon by inserting itself to various loci causing random mutations as well as altering the genome size and/or its genome GC content [5]. It was known that this retrotransposition mechanism causes various single-gene disorders such as Duchene muscular dystrophy (DMD) and Hemophilia, Breast cancer and Colon cancer, β-Thalassemia and others [5]. Although, the retrotransposition activity may have been the direct cause of such disorders, such an activity requires “intact” full-length L1s, which are only less than 200 elements in human and less than 2000 elements in mouse [27].

Intragenic LINE-1 methylation and gene regulation

Although L1s were thought to be selfish DNA elements, but there are several researches to support that intragenic L1s act as *cis*-regulatory elements to play a crucial role in cell differentiation and the maintenance of normal cell function. The decrease of intragenic L1 methylation level are associated with the reduced expression level of host genes containing L1s [9]. These methylation levels of intragenic L1s are tissue specific [26] that lead to different gene expression levels in different tissues, may be a consequence of the epigenetic modification of intragenic L1s [10, 35]. The 5' UTR of L1, which has rich of CpGs is controlled by DNA methylation, L1RNA is more transcribed when methylation level is decreased. Figure 2 shows the expression of intragenic L1s in hypermethylated, partialmethylated, and hypomethylated conditions. In hypermethylated condition, L1RNA is not transcribed, while in hypomethylated condition shows L1RNA is produced. Then the L1RNA and pre-mRNA complex is bound by AGO2 protein to suppress expression of mRNA [35].

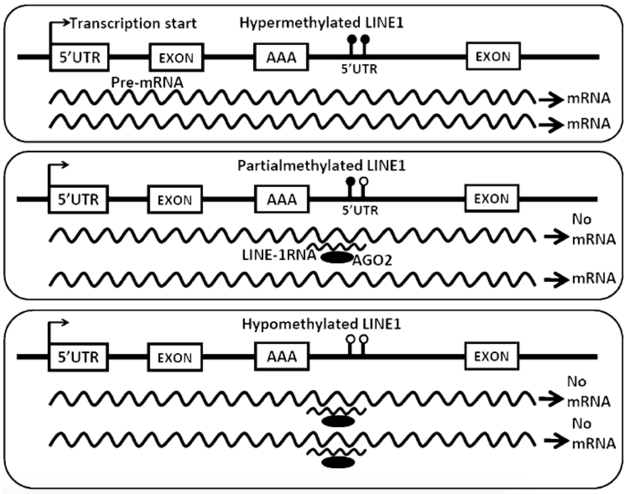


Figure 2 Intragenic L1s repress host gene expression via AGO2 [35]

Furthermore, our group reported that the orientation of intragenic L1s influence the regulation of their host genes for each siRNA experiment differently [9]. They showed some genes regulate genes containing L1s only when the intragenic L1 orientations are sense or antisense direction only, while some genes had significant results regardless of any direction. These results suggest that the intragenic L1 isoform changes or that some genes possess at least two different L1 regulation mechanisms: one mechanism promotes the gene expression, while the other mechanism suppresses other gene containing L1s [9].

siRNA mechanism silencing gene

Small interfering RNA or siRNA, knows as short interfering RNA or silencing RNA, is a 21-25 base pairs double-stranded RNA (dsRNA). There are several reports show that siRNA plays many roles, especially in the RNA interference (RNAi) pathway to mediates the degradation of mRNAs with sequences fully complementary to the siRNA [36]. The originated sequences in dsRNA from transgene- and virus-induced silencing in plants, repeat-associated transcripts (e.g. centromeres and transposons), convergent mRNA transcripts, other natural sense-antisense pairs, pseudogene, hairpin RNAs (hpRNAs), and induced from environment are processed by Dicer into siRNAs with two-base overhang on the 3' ends that direct silencing (Figure 3). These siRNA duplex forms are assembled into pre-RISC that requires Ago protein to cleave the passenger strand. The guide strand or antisense strand of siRNA and Ago are form to functional RNA-induced silencing complex (RISC) then bind to complementary sequence and prevent the expression of the targeted mRNA [37]. This siRNA mechanism silencing gene are widely used for studying the gene function in interested pathways, as well as for identifying and validating new drug targets by silencing the target gene expression [38, 39].

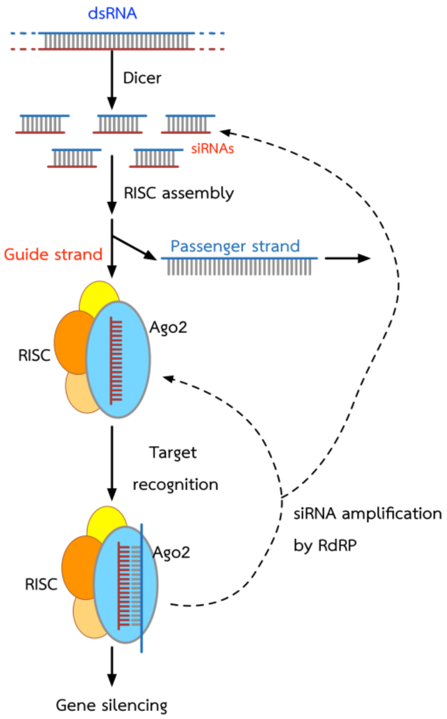


Figure 3 Mechanism of siRNA silencing gene [37]

Background of epigenetic event and global methylation change in early stage of embryogenesis

The DNA methylation is the essential epigenetic control mechanism in mammal. During the developmental processes of embryo, DNA methylation poses a fundamental epigenetic barrier to guide and restrict differentiation and prevent regression into an undifferentiated state that directs cells toward their future lineages [40]. Genomewide epigenetic reprogramming occurs at stages when developmental potency of cell changes. In normal developmental stage or disease situations, some cells have major epigenetic reprogramming dealing with the removal of epigenetic marks in the nucleus and then followed by establishment of a different set of marks. Especially, this process occurs at fertilization when many genetic marks are eradicated and replaced with embryonic marks important for early embryonic development and totipotency or pluripotency. This major reprogramming also takes place in primordial germ cells in which parental imprints are erased and totipotency is restored [41].

During preimplantation embryo development comprises four stages: fertilization, cell division, morula, and blastocyst formation [42]. In mouse embryo, after fertilization there are active and passive hypomethylation happen to control gene regulation such as imprinted genes in the parental specific manner (Figure4). This passive demethylation continues until forming to blastocyst. After that, *de novo* methylation happens and then maintain the methylation level until developed to Embryo.

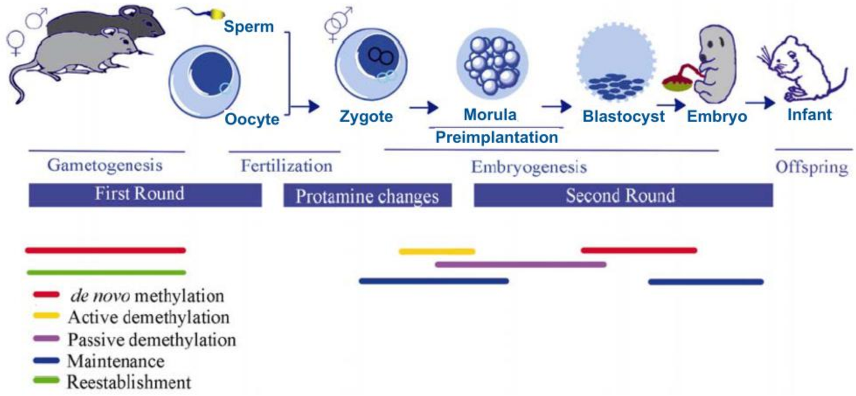


Figure 4 The methylation profile among embryonic development in mouse [43]

The zygotic gene activation (ZGA) that is the critical transition event from maternal to embryonic control of development [44]. ZGA is difference among species. In mouse, the ZGA is initiated during one-cell to two-cell stages that differ in human [44, 45]. In human embryo, the ZGA is inititated at later stages, day 3 of human embryo development between four-cell to eight-cell stages (Figure 5) [45]. The ZGA of each species correspond with the DNA methylation profiles, active demethylation in mouse embryo happens at one-cell to two-cell stages of development (Figure 4). Human embryo has different methylation profile, whereas human zygotic gene activation (ZGA) occurs between 4-cell to 8-cell stages. Interestingly, Fulka and colleagues [46] reported the changes of DNA methylation among developmental stages in human. They reported that methylation were decreased significantly between four-cell to eight-cell stages in human embryo development and rose again between murola and late blastocyst stage (Figure 6).

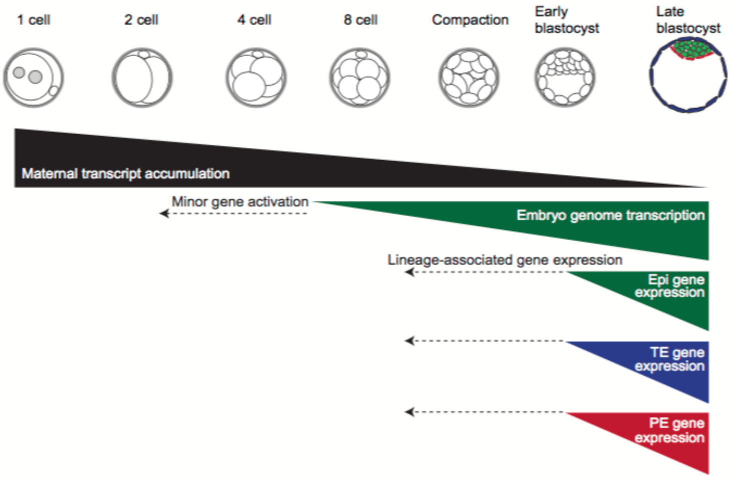


Figure 5 Genetic networks of preimplantation development in human [45]

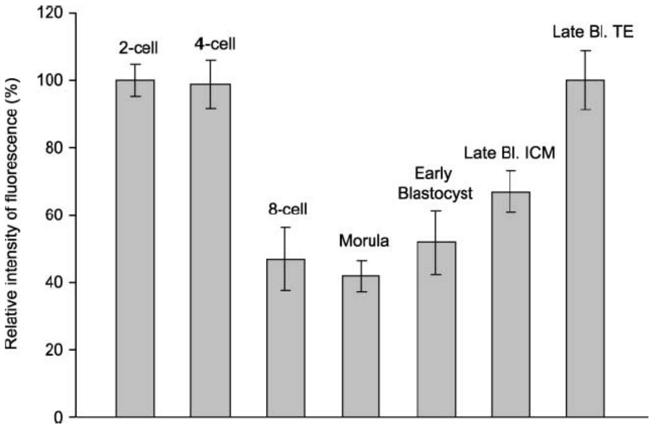


Figure 6 DNA methylation changed in human embryo development [46]

CHAPTER III

MATERIALS AND METHODS

Human and mouse LINE-1 information

Human and mouse L1 information were downloaded from the L1Base, which is a public database containing L1 elements residing in human and mouse reference genomes [27]. These L1 sequences include full-length intact L1s (putatively active with all functional elements necessary for retrotransposition present), full-length non-intact L1s (lacking some or mutated in functional sites, which reduce likelihood of mobilization), and intact ORF2 L1s (lacking ORF1 but may assist retrotransposition of *Alu*). For human, there are totally 11885 L1s including 145 full-length intact L1s, 11637 full-length non-intact L1s and 133 intact ORF2 L1s. While in mouse, there are 16508, 2382, 13660, and 466 L1s, respectively. L1Base provides their genomic locations and analytical important characteristics that associated with L1s activities.

These L1s were categorized into two groups, intragenic and intergenic, based on their genomic locations in NCBI *Homo sapiens* reference sequence (Refseq) build 36.3 and *Mus musculus* mouse Refseq build 35. The intragenic L1 group comprises L1s that are *totally* or *partially* located within the gene body, which from the first to the last exon of the largest transcript isoform (Figure 7). All other L1s are defined as intergenic L1, which shows graphical in Figure 7. An intragenic L1 is represented by a blue box, while the intergenic one is represented in a red box. The black box represents a gene (intragenic region) and the black line represents an area outside (intergenic region) the gene bodies. In human, 2535 (21.33%) of the total human L1 elements are intragenic, which located in 1454 human genes. While in mouse, 2594 elements or 15.71% of the total mouse L1s are intragenic L1s distributed over 1066 genes.

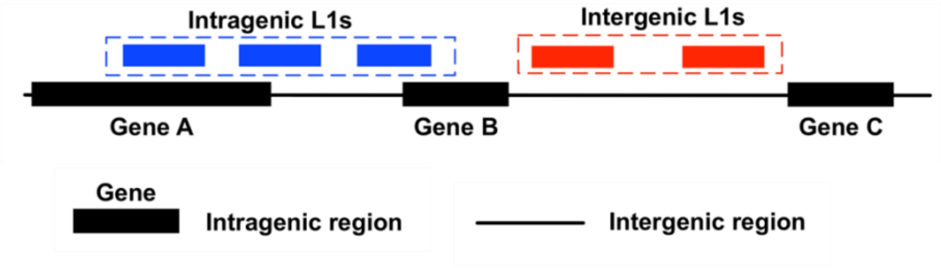
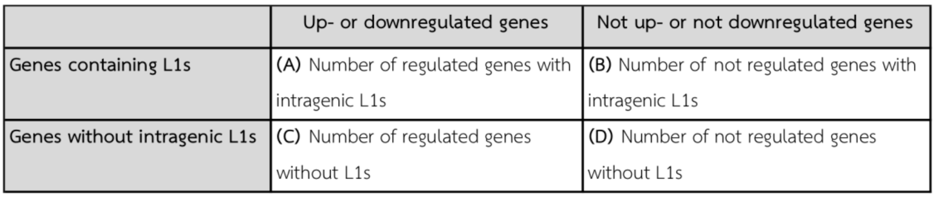


Figure 7 Definition of intragenic and intergenic L1s

Evaluate the association between gene regulation and human L1s

To evaluate if intragenic L1s can play a role in regulating host genes in human cancer. The gene expression libraries in many cancers and normal cells treated with 5-aza-2'-deoxycytidine (Aza-dC) for genome wide demethylation were downloaded from the Gene Expression Omnibus (GEO) database. These libraries compose GSE6631 (expression data from head and neck squamous cell carcinoma) [47], GSE9750 (cervical cancer) [48], GSE5816 (hypomethylated genes in lung cancer) [49], GSE14811 (hepatocellular carcinoma) [50], GSE1299 (breast cancer) [51], GSE3167 (bladder cancer) [52], GSE13911 (gastric cancer) [53], GSE6919 (prostate cancer) [54, 55], and GSE9764 (carcinoma associated fibroblast) [56]. mRNAs from these GEO experiments were classified as "up or down" regulation if they pass the significant p-value threshold in Student's *t*-test. The mRNAs that did not pass the cutoff threshold are classified as "not up- or not downregulated group". The list of samples in either group on every microarray experiments were provided as online supplementary documents of [12]. A Student's *t*-test statistics were calculated from the average and standard deviation of the differences between two conditions, such as cancer and normal cells. The paired *t*-test statistics was performed on experiments with paired samples data, while *t*-test with unequal variance [57] was performed on the others. These statistics was performed on all probes. Some probes represented more than one gene (homologous probes), a gene was counted as differentially expressed (up- or downregulated) by expression level pass the threshold from at least one unique probe. If a gene contained only homologous probes, there must be at least two homologous probes representing the same gene. Up- or downregulated genes were counted when representing probes were significantly different between test and control groups at p-value 0.01.

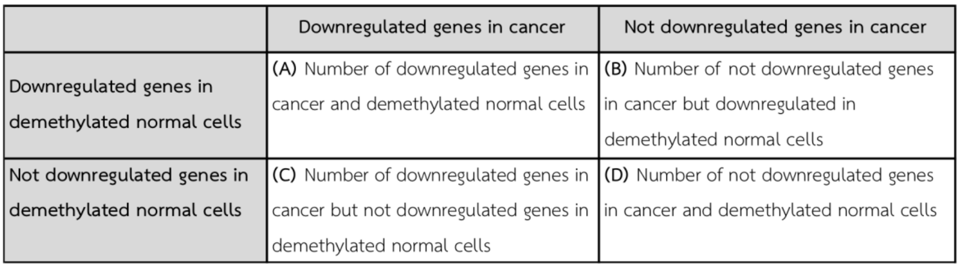
The chi-square statistics was used to evaluate intragenic L1s associated with gene regulation. The up- or downregulated genes and genes without significantly increased- or decreased-expression in each library were divided into two categories whether containing intragenic L1s or not. The chi-square analysis was formed into 2x2 table as shown in Figure 8. The statistical methods were performed by written in computer programming with Python language, using "stats" library in "scipy" module, which used to perform statistical analysis in Python.



	Up- or downregulated genes	Not up- or not downregulated genes
Genes containing L1s	**(A)** Number of regulated genes with intragenic L1s	**(B)** Number of not regulated genes with intragenic L1s
Genes without intragenic L1s	**(C)** Number of regulated genes without L1s	**(D)** Number of not regulated genes without L1s

Figure 8 Chi-square table to evaluate intragenic L1s influence host gene expression

Then to evaluate intragenic L1s can repress gene expression in demethylated normal cells in the same pattern as cancer with the chi-square table in Figure 8. Figure 9 shows the chi-square test to evaluate the association between gene downregulated in cancer and demethylated normal cells.



	Downregulated genes in cancer	Not downregulated genes in cancer
Downregulated genes in demethylated normal cells	**(A)** Number of downregulated genes in cancer and demethylated normal cells	**(B)** Number of not downregulated genes in cancer but downregulated in demethylated normal cells
Not downregulated genes in demethylated normal cells	**(C)** Number of downregulated genes in cancer but not downregulated genes in demethylated normal cells	**(D)** Number of not downregulated genes in cancer and demethylated normal cells

Figure 9 Chi-square table to evaluate downregulated genes in cancer also reduce expression in demethylated normal cells

Statistical analysis of L1 characteristics

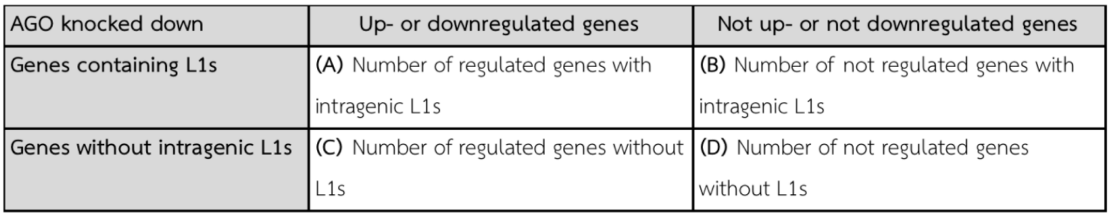
If human intragenic L1s regulate their host genes, the characteristics of L1 sequences may have differences between L1s in intragenic and intergenic regions. L1 characteristics downloaded from L1Base were classified into two groups, categorical and non-categorical characteristics. Most of the categorical characteristics were obtained by compared L1 sequences with full-length L1s (L1.2 or gi:M80343 for human and L1MdA2 or gi:M13002 for mouse). The intactness (conservation) from each of these characteristics were calculated by comparing them with the corresponding locus on the reference L1s. These characteristics can be used to predict the status of L1 activity [27, 28]. From this definition, conserved means conservation of protein functional motifs and RNA structural elements that altogether are necessary and sufficient for retrotransposition [27]. These functional motifs include ORF boundaries, promoter motifs, poly A terminator, and important amino acid residues [27, 28]. The non-categorical characteristics were presented in term of the continuous values such as a G-C content, number of ORF gaps, ORF stop codons, ORF frameshifts, and intactness score (the overall score represented conservation of predicted L1 sequences compared with full-length L1s), etc.

Human L1 sequences can be grouped into two major subfamilies according to their sequences in the 3' end of ORF2 [58], namely L1PA (primate L1s with 10668 elements) and L1M (mammalian L1s with 969 elements). Such subfamily information is thought to reflect L1 age by using the assumption that sequence divergence increases with age [59, 60]. The different age of L1s may confound the relative contributions of young and old elements to the intragenic and intergenic regions, the Mantel-Haenszel (MH) chi-square testing model was adopted [61] on categorical characteristics to adjust these confounding effects. MH chi-square operates by combining the chi-square tests performed separately on each L1 subfamilies. MH p-values and MH odds ratios (OR) between L1s located in the intragenic and intergenic region were then calculated for each characteristics. An OR greater than one indicates that the L1 status tested (conserved, etc.) has a higher probability to be intragenic than intergenic. For non-categorical (quantitative) characteristics, unpaired Student’s *t*-tests with unequal variances [57] were performed to compare the differentiation between intragenic and intergenic L1s. These statistical tests were conducted to test the null hypothesis that for a given feature of L1, there should not be much different between intragenic and intergenic L1s. There are 33 categorical and 18 non-categorical characteristics of human L1s. For chi-square tests, 2×2 contingency tables were constructed for every categorical characteristic, describing relationship between *groups* related to the host genome (intragenic/intergenic) and *condition,* e.g., conserved, CpG islands, and L1 functional characteristics.

Intragenic human L1s works as siRNAs by collaborated with AGO proteins to repress genes

To evaluated gene regulation function of intragenic human L1s act like siRNA to repress expression of gene by forming RNA-induced silencing complex (RISC) with Argonaute proteins (AGO). The analysis of transcripts regulated by Dicer and Argonaute proteins in human HEK-293 cells (GSE4246 [62]) was downloaded from GEO database. This microarray data composes the genes expression of embryonic kidney cells that knock down AGO (AGO1, AGO2, AGO3, and AGO4) and control. If AGO proteins involve to down regulate genes, the gene expressions when knock down AGO protein should be up regulate genes. If AGO proteins can up regulate genes, the gene expressions when knock down AGO should down regulate genes, in vice versa. The paired Student’s *t*-test with the p-value 0.05 cutoff was used to differentiate the upregulated genes from unchanged as well as downregulated genes.

If intragenic human L1s act as siRNA by forming with AGO proteins in RISC to repress genes, the AGO proteins should be binding closely with L1s. The chromosomal locations of AGO binding sites were retrieved from the CLIPZ database [63], which releases RNA-binding protein (RBP) binding site data generated by crosslinking and immunoprecipitation (CLIP) mapping technique. Only AGO binding sites longer than 18 base pairs were included in this study. The locations of AGO binding sites were mapped to human genome reference sequence build 36.3 and then targeted genes of those AGO were identified. The list of genes that contain AGO binding sites and also contain L1 were obtained from intersecting the set of genes that has at least one AGO target site with the set of L1-associated genes inferred from L1base as shown in Figure 10. The locations of AGO binding sites in relative to L1 sequences were also identified and were plotted to see their distribution.



AGO knocked down	Up- or downregulated genes	Not up- or not downregulated genes
Genes containing L1s	**(A)** Number of regulated genes with intragenic L1s	**(B)** Number of not regulated genes with intragenic L1s
Genes without intragenic L1s	**(C)** Number of regulated genes without L1s	**(D)** Number of not regulated genes without L1s

Figure 10 Chi-square table to evaluate gene containing L1s versus gene regulation in AGO knocked down cells

Conservation and distribution of human and mouse L1s

To evaluate L1s should be conserved in other mammals. The mouse L1s information retrieved from L1Base were also categorized into intragenic and intergenic groups. Mouse L1s can be classified into four subfamilies according to their monomer signatures located at 5' UTR of L1s [28]. These subfamilies are F (2602 elements), A (6336 elements), T_F (4940 elements), and G_F (1622 elements). There are 42 categorical functional characteristics and 11 non-categorical L1 characteristics. The MH chi-square and Student’s *t*-test were performed on categorical and non-categorical characteristics as same as in human L1s.

L1s mapped to human and mouse genomes were classified into three classes, namely L1s in autosomes (chromosome 1 to 22 in human and chromosome 1 to 19 in mouse), the X chromosome and the Y chromosome, respectively. L1 density was calculated as L1 counts per million base pairs (cMbp) of the host chromosomal regions. The genome-wide distributions and densities of intragenic and intergenic L1s were calculated separately for the two species. The comparison of L1 distributions in human and mouse genome were presented with bar graphs separately.

Analysis of intragenic L1s regulating gene expression during embryogenesis

To test the hypothesis that intragenic L1s regulate genes in other physiological cellular processes such as embryogenesis in mammalian species, the publicly available microarray data from different stages of preimplantation embryonic development, namely one-cell, two-cell, four-cell, eight-cell, morula and blastocyst stages (GEO accession number GSE18290 [64]) was analyzed. The gene regulation profiles of one-cell stage were compared with all other stages for human and mouse. Differentially expressed genes between each developmental stage and the one-cell stage were identified using paired Student's *t*-test. Paired *t*-statistics were calculated from the average and standard deviation of differences between paired samples of each developmental stage and the one-cell stage. Genes with p-values less than 0.05 were considered as differentially expressed. Chi-square analysis was then performed to test if genes containing L1 sequences are associated with up regulation with respect to the one-cell stage. The 2x2 contingency tables were constructed with rows of number of genes with L1 present and L1 absent, and columns of number of upregulated genes and the rest as shown in Figure 8. Similar 2x2 contingency tables were also constructed for testing L1 association with downregulated genes in which columns were constructed as downregulated genes and the rest. Chi-square tests were performed for both human and mouse between each pair of time-points. Thresholds for statistical significance were p-value $<$ 1.0E-03 and OR $>$ 1.0.

Finally, the significant regulated genes containing L1s were identified. The downregulated homologous genes in both species were listed. The gene function and associated pathway were also identified to support the hypothesis.

CHAPTER IV

RESULTS

Human intragenic L1s repress genes in global hypomethylated environment

This study evaluates the association between gene regulations and intragenic L1s in global hypomethylated environments. These microarray data compose nine cancer types, the genes in each microarray experiments are classified into upregulated, downregulated, or not regulated. The association tests between gene possessing L1s and gene regulations were performed by chi-square test in Figure 8. Table 1 shows chi-square tests of gene expression in gastric cancer (GSE13911). Genes possessing intragenic L1s were found less likely to be upregulated with Odds Ratio (OR) = 0.64, p-value = 2.6572E-08, 95% confidence interval = 0.54-0.75 (Table 1A). Moreover, expression of genes containing L1s were more commonly decreased (OR = 1.61, p-value = 9.4992E-16, 95% CI = 1.43-1.81; Table 1B). Intragenic L1s may control hundreds of genes. Among 1458 genes containing L1s, 1158 genes were not upregulated and 459 genes were downregulated (Table 1A and 1B).

Table 1 Chi-square tests compare between gene possessing L1s and gene regulation in gastric cancer

GSE13911 (Microsatellite instable gastric cancer vs normal stomach epithelium)

	Up	Not up
L1	181	1158
no L1	3710	15095

Odds Ratio = 0.64, 95% confidence interval = 0.54 - 0.75

Sum chi = 30.94298, p-value = 2.6572E-08

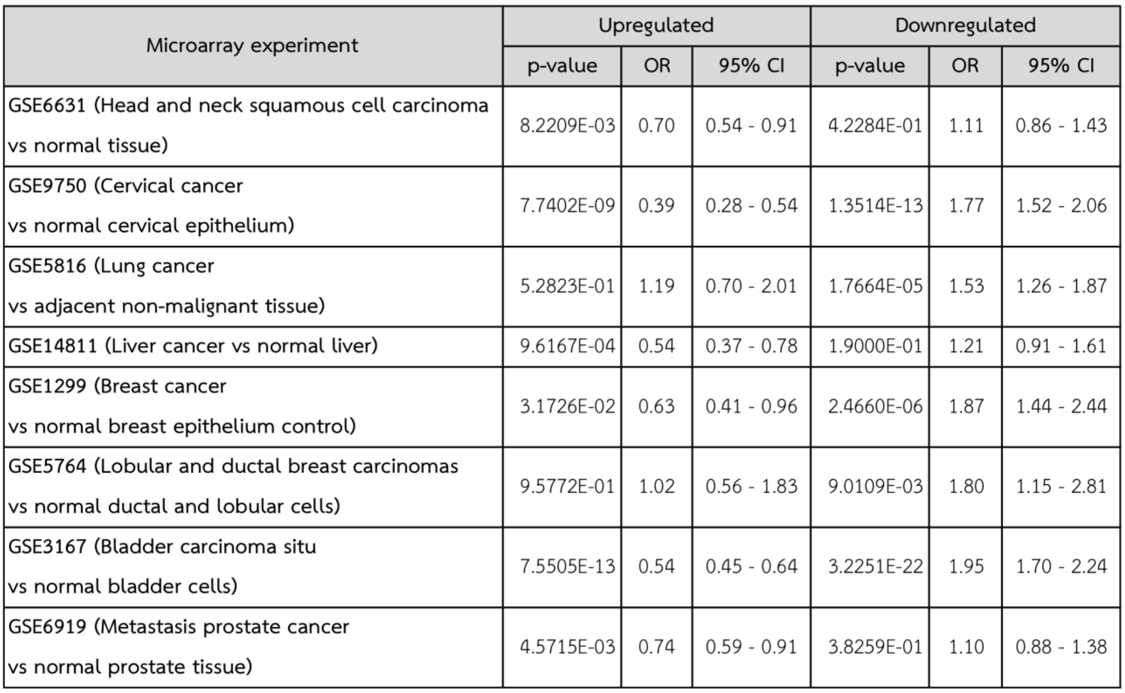
	Down	Not down
L1	459	880
no L1	4594	14211

Odds Ratio = 1.61, 95% confidence interval = 1.43 - 1.81

Sum chi = 64.53169, p-value = 9.4992E-16

The analysis on other cancers were performed and showed in Table 2. These results showed five in eight cancers comprise cervical cancer, lung cancer, breast cancer, lobular and ductal breast carcinomas, and bladder carcinoma situ were also more likely to decrease expression in genes containing L1s. Moreover, genes with higher expression levels in those cancers were less likely to possess L1s. Therefore, intragenic L1s may repress host genes in these cancers. In prostate cancer, head and neck squamous cell carcinoma, and liver cancer, L1s are not statistically significant associated with gene regulation.

Table 2 Chi-square tests were performed on multiple cancer types



Microarray experiment	Upregulated			Downregulated		
	p-value	OR	95% CI	p-value	OR	95% CI
GSE6631 (Head and neck squamous cell carcinoma vs normal tissue)	8.2209E-03	0.70	0.54 - 0.91	4.2284E-01	1.11	0.86 - 1.43
GSE9750 (Cervical cancer vs normal cervical epithelium)	7.7402E-09	0.39	0.28 - 0.54	1.3514E-13	1.77	1.52 - 2.06
GSE5816 (Lung cancer vs adjacent non-malignant tissue)	5.2823E-01	1.19	0.70 - 2.01	1.7664E-05	1.53	1.26 - 1.87
GSE14811 (Liver cancer vs normal liver)	9.6167E-04	0.54	0.37 - 0.78	1.9000E-01	1.21	0.91 - 1.61
GSE1299 (Breast cancer vs normal breast epithelium control)	3.1726E-02	0.63	0.41 - 0.96	2.4660E-06	1.87	1.44 - 2.44
GSE5764 (Lobular and ductal breast carcinomas vs normal ductal and lobular cells)	9.5772E-01	1.02	0.56 - 1.83	9.0109E-03	1.80	1.15 - 2.81
GSE3167 (Bladder carcinoma situ vs normal bladder cells)	7.5505E-13	0.54	0.45 - 0.64	3.2251E-22	1.95	1.70 - 2.24
GSE6919 (Metastasis prostate cancer vs normal prostate tissue)	4.5715E-03	0.74	0.59 - 0.91	3.8259E-01	1.10	0.88 - 1.38

Our group measured intragenic L1 methylation and host gene's mRNA level to explore the relation pattern between L1 methylation levels and gene expression [12]. They previously evaluated methylation levels of 17 intragenic L1 loci and found that the L1 methylation levels in some loci are strongly correlated in cancer cells, suggesting locus specific mechanism [18].

Loss of methylation in normal cell represses genes that harbor L1s

An analysis of gene expression in human bronchial epithelial cells (hBECs) and human mesenchymal stem cells (hMSCs) after genome wide demethylation by Aza-dC treatment demonstrated a greater prevalence of intragenic L1s in downregulated genes in Table 3 and Table 4 (OR = 1.61, p-value = 1.0575E-03 and OR = 1.59, p-value = 6.2736E-03, respectively), interestingly, a similar pattern as found in cancers.

Table 3 Chi-square tests compare between gene possessing L1s and gene regulation in hBEC cells treated with high dose Aza-dC

GSE5816 (hBEC high dose vs human bronchial epithelium)

	Up	Not up
L1	34	1341
no L1	504	18697

Odds Ratio = 0.94, 95% confidence interval = 0.66 - 1.34

Sum chi = 0.11663, p-value = 7.3272E-01

	Down	Not down
L1	54	1321
no L1	476	18725

Odds Ratio = 1.61, 95% confidence interval = 1.21 - 2.14

Sum chi = 10.72413, p-value = 1.0575E-03

Table 4 Chi-square tests compare between gene possessing L1s and gene regulation in hMSCs treated with Aza-dC

GSE9764 (5-azadeoxycytidine treated vs untreated human mesenchymal stem cells)

	Up	Not up
L1	35	1340
no L1	427	18770

Odds Ratio = 1.15, 95% confidence interval = 0.81 - 1.63

Sum chi = 0.60280, p-value = 4.3751E-01

	Down	Not down
L1	39	1336
no L1	346	18851

Odds Ratio = 1.59, 95% confidence interval = 1.14 - 2.22

Sum chi = 7.46999, p-value = 6.2736E-03

The genome wide hypomethylation regulated genes in cancer was further explored. The chi-square test for determining the significance of overlap between downregulated genes in demethylated hBECs and in lung cancer was performed. Genes which were downregulated in Aza-dC treatment on hBECs were found to preferentially have lower mRNA levels in the cancerous cells of the lung (p-value = 2.4401E-28; OR = 3.43); Table 5A). This supports the hypothesis that hypomethylation down regulates genes in cancer.

Interestingly, hypomethylation down regulates both groups of genes, with L1s (p-value=1.8525E-02; OR=2.50; Table 5B) and without L1s (p-value=1.3879E-26; OR = 3.52; Table 5C). Therefore, it is possible that in addition to L1 there are other DNA methylated gene body elements that regulated gene expression. This hypothesis is supported by a recent report that, in gene body, unique methylated sequences are more prevalence in highly expressed genes [22].

Table 5 Chi-square test compare between genes that down regulate both in lung cancer and hBECs treated with Aza-dC

A All genes GSE5816 (Lung cancer and Aza-dC hBECs)

Lung cancer vs hBECs

		Down	Not down
Aza-dC	Down	92	438
vs hBECs	Not down	1155	18884

Odds Ratio = 3.43, 95% confidence interval = 2.72 - 4.33

Sum chi = 121.88993, p-value = 2.4401E-28

B Genes with intragenic L1s

Lung cancer vs hBECs

		Down	Not down
Aza-dC	Down	10	44
vs hBECs	Not down	110	1211

Odds Ratio = 2.50, 95% confidence interval = 1.23 - 5.11

Sum chi = 5.54578, p-value = 1.8525E-02

C Genes without intragenic L1s

Lung cancer vs hBECs

		Down	Not down
Aza-dC	Down	82	394
vs hBECs	Not down	1045	17673

Odds Ratio = 3.52, 95% confidence interval = 2.75 - 4.50

Sum chi = 113.87503, p-value = 1.3879E-26

Our group measured the methylation and RNA levels that showed an inverse correlation between genome wide L1 methylation and L1 RNA. That finding supports the hypothesis that L1 hypomethylation increases L1 RNA transcription [65]. Intronic genes have been proposed to form aberrant RNA complexes with host genes and consequently inactivate host gene transcription [66].

Conservation of human intragenic L1 sequences to control gene regulation

Human L1 sequence variances and other characteristics downloaded from L1Base [27] were classified into intragenic and intergenic L1s. The MH chi-square tested was performed on 33 categorical characteristics while Student's *t*-test was performed on 18 non-categorical characteristics. The MH chi-square and student's *t*-test p-value measurements were presented in -$\log_{10}$(p-value), where the higher number represents more significant value. Human intragenic L1s are more likely conserved than intergenic L1s (Figure 11A), there are 15 categorical characteristics were passed the significant threshold (p-value=1.0E-03). The green and orange bars in Figure 11A represent conserved and mutated features, respectively. These colored bars are aligned with L1 structure shown below the graphs. The significant characteristics were widely distributed along L1 structure including 5' UTR, ORF1, and ORF2. The bars marked with an asterisk (*) indicate the features calculated for the entire L1 sequence. The left side of Figure 11A shows intragenic L1s are more conserved, while right side shows intergenic L1s are more mutated. Figure 11B shows the significant non-categorical characteristics. The blue columns indicate that more of these features appear in the intragenic L1s than that of intergenic ones. The red columns indicate that there are more of such features in the intergenic L1s than that of intragenic ones. The left panel shows GC content, intactness score, ORF1 and ORF2 codon adaptation indexes (CAI) were more prevalence in intragenic. Intergenic L1s contain more A and T nucleotides, frameshifts, and stop codons in both ORF were shown in right panel. This results show intragenic L1 sequences have been conserved across evolutionary time with respect to transcriptional activity for mammalian DNA methylation. These findings implied physiological functions of intragenic L1 methylation. The detail of statistical results were provided as online supplement in [67]. The précis of human L1 characteristics were summarized in Appendix.

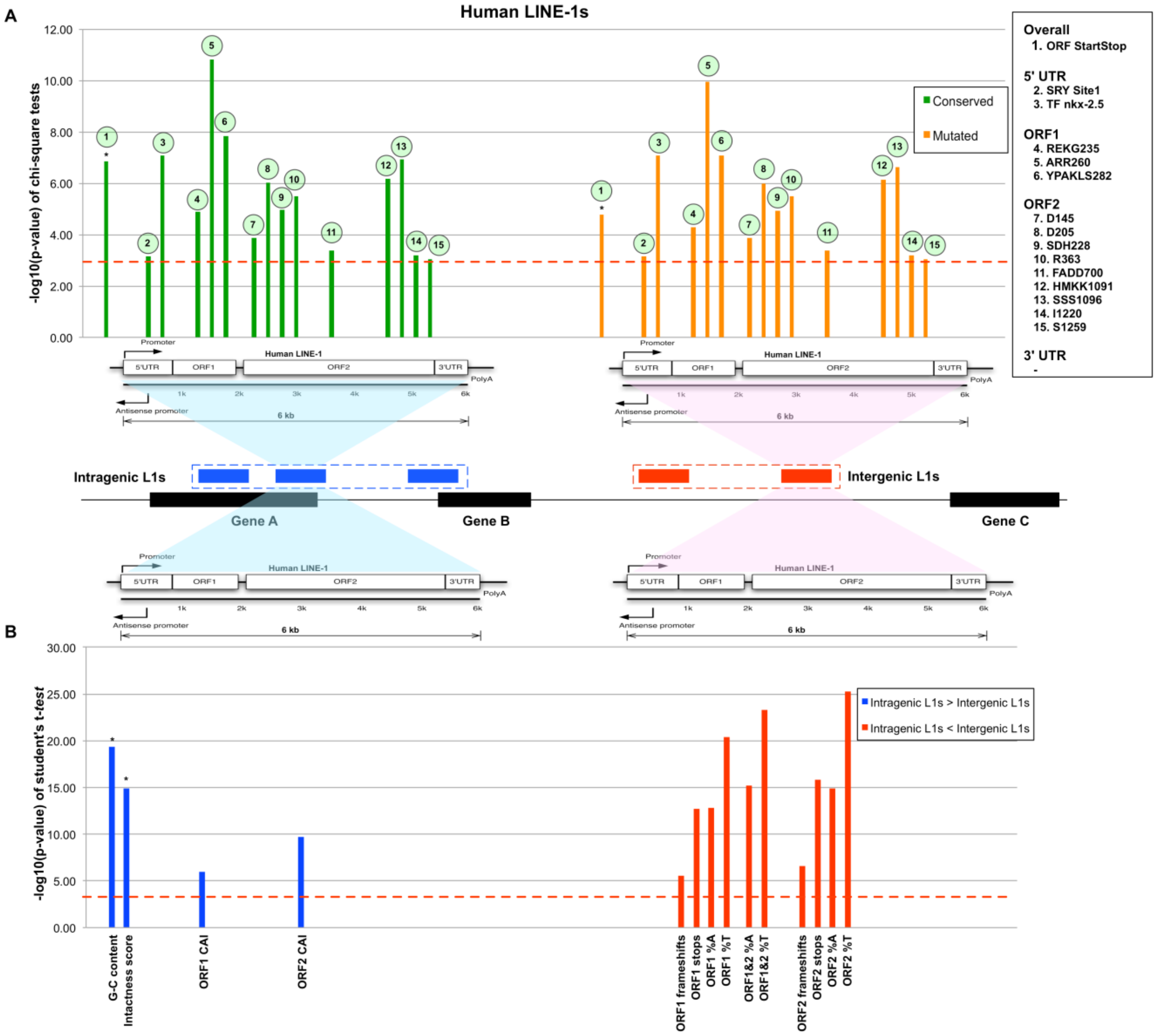


Figure 11 The comparison between intragenic and intergenic human L1 sequences.

Human intragenic LINE-1 elements repress transcription in cancer cells through AGO2

To evaluate if intragenic L1 RNA reduces host gene mRNA via AGO2, AGO2 protein deprivation will result to increase mRNA levels of genes hosting L1s. Analysis of mRNA microarray of AGO knock down experiment (GSE4246) was performed. Table 6 demonstrates that the limited expression of AGO protein in a human embryonic kidney cell line resulted in an expression pattern of gene containing L1s that was opposite from that observed during L1 hypomethylation; namely, they were more likely to be upregulated in AGO2 knocked down (OR = 1.48, p-value = 9.1036E-05;

Table 6B). Other AGO knocked down experiments (AGO1, AGO3, and AGO4) were not upregulated gene containing L1s (Table 6). AGO1 and AGO4 have opposite results with AGO2 that likely to downregulate genes containing L1s. This suggested that AGO2 preferentially limits the concentration of mRNAs derived from genes containing L1s. AGO1 and AGO4 should play a role that opposite to AGO2.

Table 6 Chi-square test compare between genes containing L1s and gene regulations when knocked down AGO proteins

GSE4246 (HEK-293 cell lines with depleted AGO proteins)

AGO1 knocked down

	Up	Not up
L1	108	735
no L1	2071	8834

Odds Ratio = 0.63, 95% confidence interval = 0.51 - 0.77

Sum chi = 19.78137, p-value = 8.6825E-06

	Down	Not down
L1	158	685
no L1	1607	9298

Odds Ratio = 1.33, 95% confidence interval = 1.11 - 1.60

Sum chi = 9.83745, p-value = 1.7099E-03

AGO2 knocked down

	Up	Not up
L1	126	709
no L1	1161	9698

Odds Ratio = 1.48, 95% confidence interval = 1.22 - 1.81

Sum chi = 15.31408, p-value = 9.1036E-05

	Down	Not down
L1	61	774
no L1	1176	9683

Odds Ratio = 0.65, 95% confidence interval = 0.50 - 0.85

Sum chi = 10.18163, p-value = 1.4185E-03

AGO3 knocked down

	Up	Not up
L1	47	791
no L1	736	10058

Odds Ratio = 0.81, 95% confidence interval = 0.60 - 1.10

Sum chi = 1.81346, p-value = 1.7809E-01

	Down	Not down
L1	79	759
no L1	810	9984

Odds Ratio = 1.28, 95% confidence interval = 1.01 - 1.64

Sum chi = 4.07408, p-value = 4.3546E-02

AGO4 knocked down

	Up	Not up
L1	84	757
no L1	2465	8467

Odds Ratio = 0.38, 95% confidence interval = 0.30 - 0.48

Sum chi = 72.62709, p-value = 1.5661E-17

	Down	Not down
L1	172	669
no L1	1612	9320

Odds Ratio = 1.49, 95% confidence interval = 1.25 - 1.77

Sum chi = 19.77658, p-value = 8.7043E-06

The AGO2 binding sites from CLIPZ database were downloaded and mapped to human reference sequence (NCBI Refseq build 36.3). The genes containing AGO2 binding sites were identified and used to association with genes that upregulated in AGO2 knocked down experiment. Table 7 shows the 2x2 contingency table displaying a chi-square test of association between the presence of L1 and AGO2 binding sites. AGO2 binding sites were found in all L1-containing genes that were upregulated in AGO2 knocked down cells (124 out of 126 genes, OR = 17.91, p = 3.5138E-08).

Table 7 Chi-square test compare between upregulated genes containing L1s and AGO2 binding

	All genes	GSE5816 (Lung cancer and Aza-dC hBECs)	
		Lung cancer vs hBECs	
		Down	Not down
Aza-dC	Down	92	438
vs hBECs	Not down	1155	18884

Odds Ratio = 3.43, 95% confidence interval = 2.72 - 4.33

Sum chi = 121.88993, p-value = 2.4401E-28

Focusing on these upregulated genes, the distribution of AGO2 binding sites related on L1 can be found. Numbers of AGO2 binding sites were counted if they are found in the vicinity of L1. Particularly, a histogram by counting the number of AGO2 binding sites located within 600 kb upstream and downstream of L1 was created by using the 25-kb interval size (Figure 12). Figure 12A and Figure 12B demonstrate the frequency distribution of AGO2 binding sites with respect to antisense and sense L1, respectively. Interestingly, hundreds of AGO2 binding sites were found at hypothetical locations presenting double strand RNA between pre-mRNA and 5' or 3' L1 transduction sequence. These were sequences nearby L1 at 5' direction from L1 toward gene transcriptional start sites. Therefore, AGO2 preferentially regulates genes containing L1s by targeting intragenic L1 RNAs with sequence complementary to pre-mRNAs.

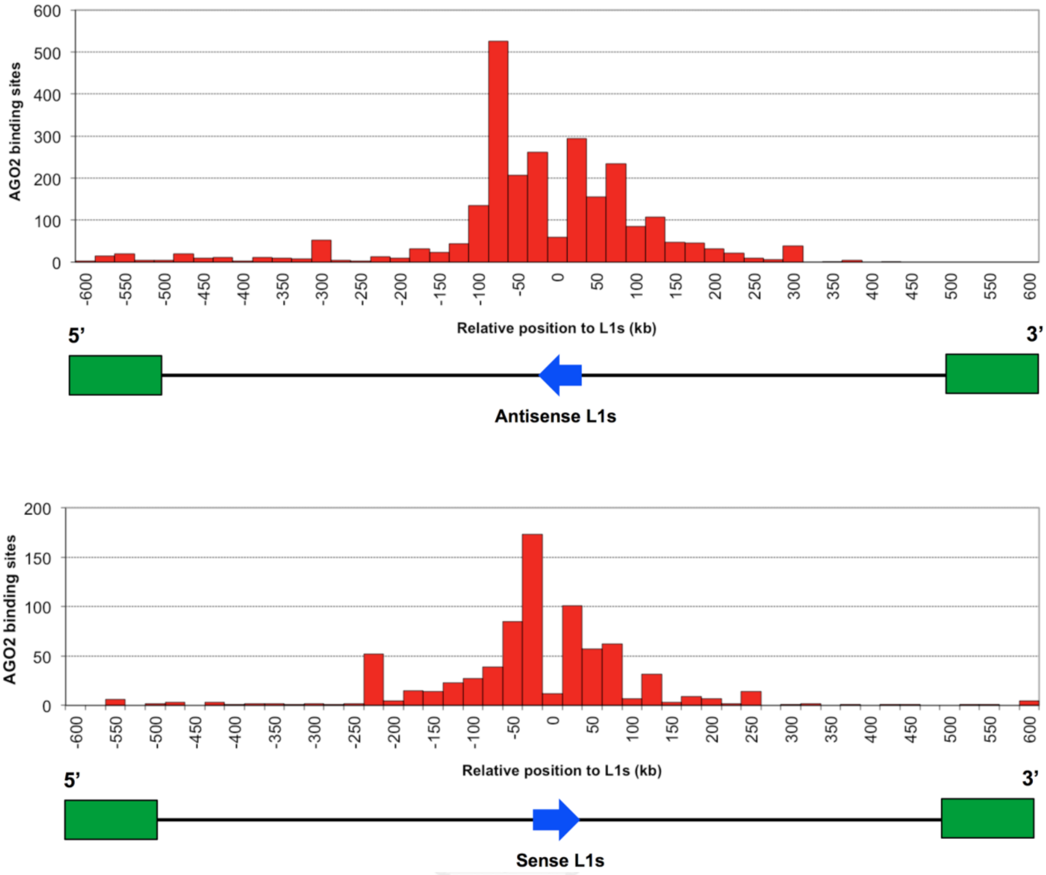


Figure 12 The frequency distribution of AGO2 binding sites corresponding with the location of antisense and sense L1s

Conservation of mouse L1s

Previous experiment showed that intragenic human L1s are more conserved than intergenic ones. The greater conservation of human intragenic L1 sequences may reflect functions dependent on L1 transcription. The conservation and distinction of intragenic and intergenic L1 sequences in mouse and human may be similar. The conservation of mouse L1s were also tested by Mantel-Haenszel chi-square and unequal variance Student’s *t*-tests, the detail information of mouse L1 characteristics were summarized in Appendix. The completed statistical results were provided as online supplement in [67]. Figure 13A shows that intragenic L1s are significantly more conserved in mouse as well as human. In mouse intragenic L1s, conserved features are distributed along the structure of L1 except for the 5' UTR.

Only one functional feature, the SA-154 acceptor splice site on antisense mouse L1 sequences, is poorly conserved among intragenic mouse L1. There are three conserved features in ORF1, six in ORF2 and one in the 3' UTR. For both mouse and human, intragenic L1s have significantly higher intactness score and GC contents than that of intergenic L1s (Figure 13B). The number of monomer and monomer splice sites are significantly greater in intragenic L1s. These analyses indicate that many important features of mouse L1 sequences are well conserved in intragenic L1s. Furthermore, the significantly higher number of monomer repeats (> 3 copies on average) in mouse intragenic L1s suggests their main roles in regulating transcriptional activities as reported in [4, 34]. Therefore, like human intragenic L1s, the conservation of structural features could suggest a similar transcriptional role.

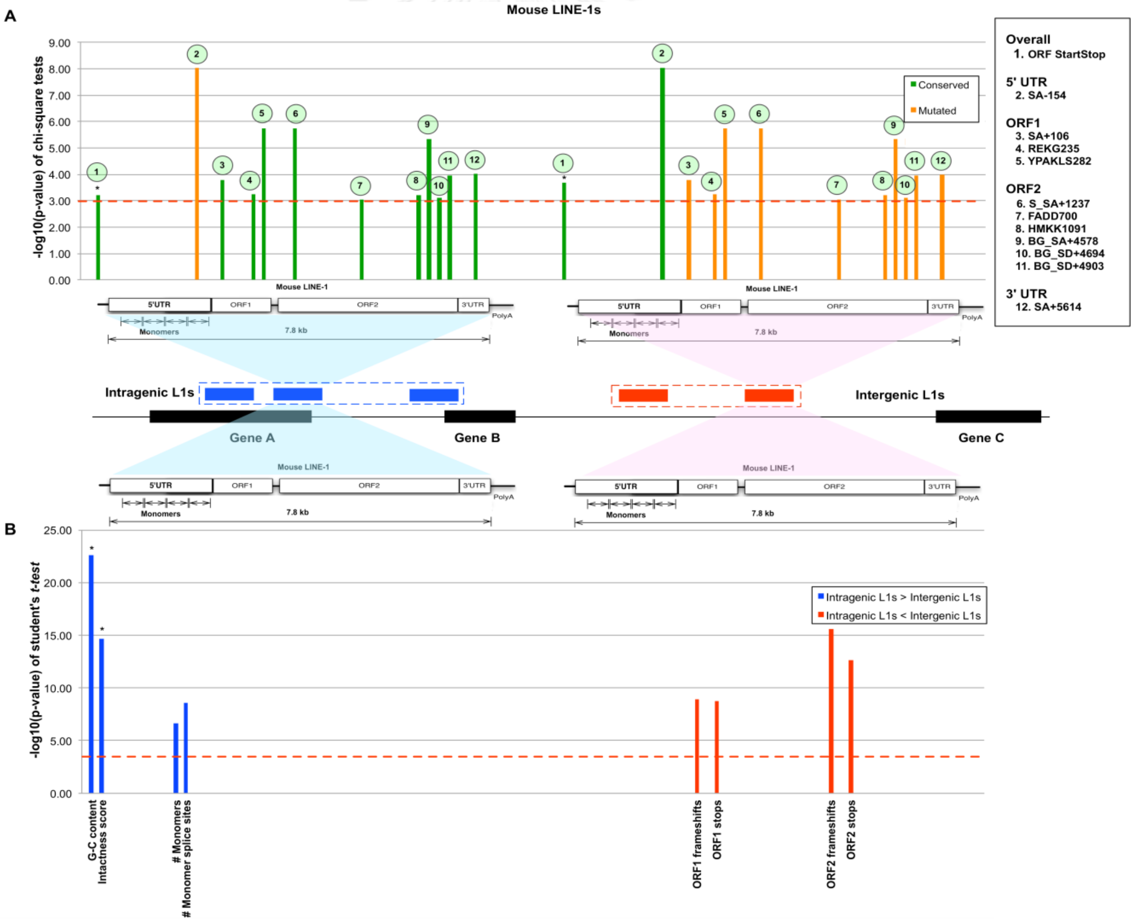


Figure 13 The comparison between intragenic and intergenic mouse L1 sequences

Comparison of L1 chromosomal distributions

The densities of the mouse and human L1s in term of number of L1s per Mbp (cMbp) on autosome, X and Y chromosomes were represented with a bar graph (grey columns present distribution of mouse L1s while black columns present human L1s) in Figure 14A. Except for the Y chromosome, L1 density is much greater in mouse than that of human. Intragenic L1 density is lower than intergenic for autosomes and X chromosome of both species, whereas the density of intragenic L1s is greater in the Y chromosome of both species. Figure 14B showed two side-by-side bar graphs comparing intragenic (blue columns) vs. intergenic (red columns) L1s on mouse and human genomes. The denseness of intragenic L1s in Y-chromosome (ChrY) cannot be explained by the compactness of Y-chromosome. The percentage of intergenic region is always larger than intragenic region on all chromosomes and is largest for the Y chromosome. In the human genome on average, 58.95% of autosomes are intergenic whereas the intergenic contents of sex chromosomes are higher (68.68% and 94.26% for X and Y respectively). Intergenic contents in mouse are similar (68.30% average of autosomes, 78.54% for X and 96.35% for Y).

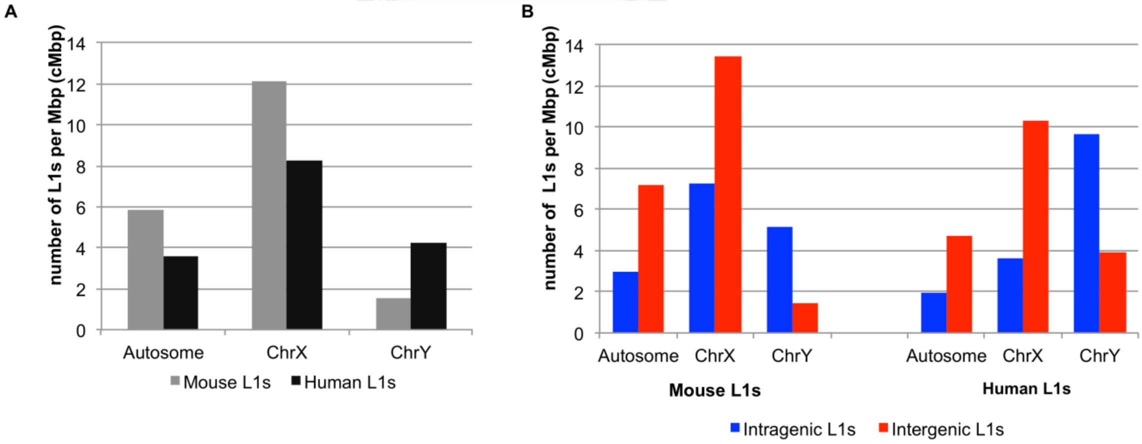


Figure 14 Distribution of mouse and human L1s over their genomes

Intragenic L1s regulate genes in early embryogenesis

L1s are expressed in early embryogenesis [11], and L1 products are essential for development [68]. It is not known, however, if expression of intragenic L1 regulates expression of gene pre-mRNA in embryogenesis similar to what was reported in cancer [12]. The analysis of microarray expression data from human and mouse early embryonic stages were performed whether changes in expression are associated with intragenic L1s.

In human, the observed number of genes with intragenic L1 and downregulated relative to the one-cell stage are significantly higher than expected for eight-cell, morula, and blastocyst. In contrast, no significant association was found for upregulated genes and intragenic L1s (Table 8). Significantly higher than expected numbers of downregulated genes with intragenic L1 were also found for all stages except blastocyst in mouse (Table 9).

Table 8 Intragenic L1s control gene expression in human early embryogenesis.

Chi-square test of association between human intragenic L1s and differential embryo gene expression stage (between cell division stage)						
		2-cell vs. 1-cell	**4-cell vs. 1-cell**	**8-cell vs. 1-cell**	**Morula vs. 1-cell**	**Blastocyst vs. 1-cell**
Up	**p-value**	1.5217E-01	3.6563E-03	1.9267E-03	3.4991E-06	1.7847E-04
	OR	1.29	1.33	0.75	0.70	0.76
	95%CI	0.94-1.51	1.10-1.60	0.63-0.90	0.60-0.81	0.66-0.88
Down	**p-value**	8.9550E-01	8.1226E-01	**3.2387E-24**	**1.7497E-24**	**1.4097E-18**
	OR	0.98	1.02	**1.80**	**1.83**	**1.70**
	95%CI	0.72-1.34	0.85-1.24	**1.61-2.02**	**1.63-2.05**	**1.51-1.91**

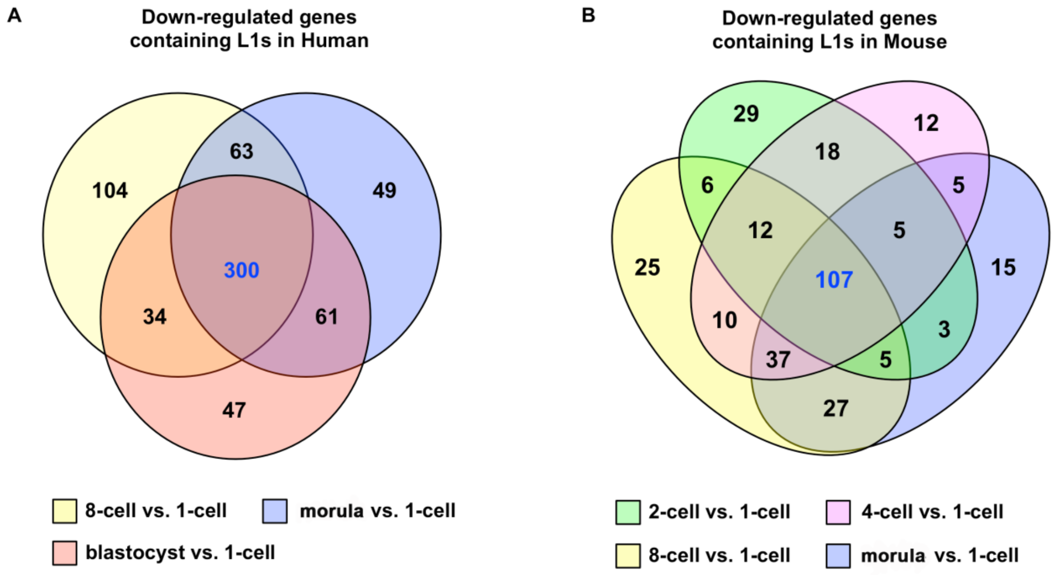
Bold items indicate differential stages that pass the threshold (OR>1.0 and p-value<1.0E-03).

Table 9 Intragenic L1s control gene expression in mouse early embryogenesis.

Chi-square test of association between mouse intragenic L1s and differential embryo gene expression stage (between cell division stage)						
		2-cell vs. 1-cell	**4-cell vs. 1-cell**	**8-cell vs. 1-cell**	**Morula vs. 1-cell**	**Blastocyst vs. 1-cell**
Up	**p-value**	9.0033E-20	7.0438E-18	1.0780E-21	1.7495E-13	1.1192E-04
	OR	0.17	0.32	0.29	0.42	0.48
	95%CI	0.11-0.26	0.24-0.42	0.23-0.38	0.33-0.53	0.32-0.70
Down	**p-value**	**6.7540E-07**	**9.3628E-09**	**1.7002E-10**	**6.5540E-07**	1.2723E-02
	OR	**1.57**	**1.65**	**1.73**	**1.55**	1.27
	95%CI	**1.31-1.87**	**1.39-1.96**	**1.46-2.05**	**1.30-1.85**	1.05-1.53

Bold items indicate differential stages that pass the threshold (OR>1.0 and p-value<1.0E-03).

Among the stages with significant association of intragenic L1 and down-regulation, 300 genes are commonly down regulated among human stages whereas 107 are common among mouse stages (Figure 15). Figure 15A showed the intersection of three gene sets in human genome. A colored circle represents each gene set. The numbers in yellow, blue, and red circles indicate the numbers of associated human genes in “8-cell vs. 1-cell”, “morula vs. 1-cell”, and “blastocyst vs. 1-cell” differential expression stages, respectively. The intersection of four gene sets in mouse genome were shown in Figure 15B. Each gene set is represented by a colored oval. The numbers in green, pink, yellow and blue ovals indicate the numbers of associated mouse genes in “2-cell vs. 1-cell”, “4-cell vs. 1-cell”, “8-cell vs. 1-cell”, and “morula vs. 1-cell” differential expressions stages, respectively. Among the genes in these two intersection sets, 14 are orthologous between human and mouse, according to the mouse genome database [69]. Figure 15C showed the list of mouse-human orthologous genes found in both mouse and human intersection gene sets. Each orthologous gene pair indicates the mouse gene name followed by the human gene name. The numbers in parentheses present the corresponding gene ids. By using Gene Ontology [70] and GeneCards [71], the molecular functions of these orthologous genes are listed in Table 10.



C

Orthologous down-regulated genes containing L1s

Kcnq1(16535) - **KCNQ1**(3784), **Rad51l1**(19363) - **RAD51B**(5890), **Rabgap1l**(29809) - **RABGAP1L**(9910), **Fut8**(53618) - **FUT8**(2530), **Pde3a**(54611) - **PDE3A**(5139), **Lmbr1**(56873) - **LMBR1**(64327), **Vav3**(57257) - **VAV3**(10451), **Rsrc1**(66880) - **RSRC1**(51319), **Ccdc132**(73288) - **CCDC132**(55610), **Tusc3**(80286) - **TUSC3**(7991), **Hivep1**(110521) - **HIVEP1**(3096), **Rims2**(116838) - **RIMS2**(9699), **Tox**(252838) - **TOX**(9760), **Cntn4**(269784) - **CNTN4**(152330)

Figure 15 The down-regulated gene sets at differential gene expression stages in early embryogenesis that pass the chi-square tests

Table 10 Molecular functions of 14 orthologous down-regulated genes during early embryogenesis in human and mouse.

Mouse gene	Mouse GeneID	Mouse gene function	Human gene	Human GeneID	Human gene function	Function from GeneCards
Kcnq1	16535	calmodulin binding, delayed rectifier potassium channel activity, ion channel activity, outward rectifier potassium channel activity, potassium channel activity, voltage-gated ion channel activity, voltage-gated potassium channel activity	KCNQ1	3784	calmodulin binding, delayed rectifier potassium channel activity, outward rectifier potassium channel activity, voltage-gated potassium channel activity	Probably important in cardiac repolarization. Associates with KCNE1 (MinK) to form the I(Ks) cardiac potassium current. Elicits a rapidly activating, potassium-selective outward current. Muscarinic agonist oxotremorine-M strongly suppresses KCNQ1/KCNE1 current in CHO cells in which cloned KCNQ1/KCNE1 channels were coexpressed with M1 muscarinic receptors. May associate also with KCNE3 (MiRP2) to form the potassium channel that is important for cyclic AMP-stimulated intestinal secretion of chloride ions, which is reduced in cystic fibrosis and pathologically stimulated in cholera and other forms of secretory diarrhea.
Rad51l1	19363	ATP binding, DNA binding, DNA-dependent ATPase activity, nucleoside-triphosphatase activity, nucleotide binding	RAD51B	5890	ATP binding, DNA binding, DNA-dependent ATPase activity, protein binding	Involved in the homologous recombination repair (HRR) pathway of double-stranded DNA breaks arising during DNA replication or induced by DNA-damaging agents. May promote the assembly of presynaptic RAD51 nucleoprotein filaments. The RAD51B-RAD51C dimer exhibits single-stranded DNA-dependent ATPase activity. The BCDX2 complex binds single-stranded DNA, single-stranded gaps in duplex DNA and specifically to nicks in duplex DNA.
Rabgap1l	29809	GTPase activator activity, Rab GTPase activator activity, Rab GTPase binding	RABGAP1L	9910	Rab GTPase activator activity, Rab GTPase binding	
Fut8	53618	SH3 domain binding, alpha-(1->6)-fucosyltransferase activity, glycoprotein 6-alpha-L-fucosyltransferase activity, transferase activity, transferase activity, transferring glycosyl groups	FUT8	2530	SH3 domain binding, glycoprotein 6-alpha-L-fucosyltransferase activity	Catalyzes the addition of fucose in alpha 1-6 linkage to the first GlcNAc residue, next to the peptide chains in N-glycans.
Pde3a	54611	3',5'-cyclic-AMP phosphodiesterase activity, 3',5'-cyclic-AMP phosphodiesterase activity, 3',5'-cyclic-nucleotide phosphodiesterase activity, cAMP binding, cGMP-inhibited cyclic-nucleotide phosphodiesterase activity, catalytic activity, hydrolase activity, metal ion binding, phosphoric diester hydrolase activity	PDE3A	5139	3',5'-cyclic-AMP phosphodiesterase activity, cAMP binding, cGMP-inhibited cyclic-nucleotide phosphodiesterase activity, metal ion binding	Cyclic nucleotide phosphodiesterase with a dual-specificity for the second messengers cAMP and cGMP, which are key regulators of many important physiological processes (By similarity).
Lmbr1	56873	-	LMBR1	64327	-	Putative membrane receptor.
Vav3	57257	Rac guanyl-nucleotide exchange factor activity, Rho guanyl-nucleotide exchange factor activity, epidermal growth factor receptor binding, guanyl-nucleotide exchange factor activity, metal ion binding, phospholipid binding, protein binding	VAV3	10451	GTPase activator activity, Rac guanyl-nucleotide exchange factor activity, SH3/SH2 adaptor activity, epidermal growth factor receptor binding, metal ion binding, phospholipid binding, protein binding	Exchange factor for GTP-binding proteins RhoA, RhoG and, to a lesser extent, Rac1. Binds physically to the nucleotide-free states of those GTPases. Plays an important role in angiogenesis. Its recruitment by phosphorylated EPHA2 is critical for EFNA1-induced RAC1 GTPase activation and vascular endothelial cell migration and assembly (By similarity). May be important for integrin-mediated signaling, at least in some cell types. In osteoclasts, along with SYK tyrosine kinase, required for signaling through integrin alpha-v/beta-1 (ITAGV-ITGB1), a crucial event for osteoclast proper cytoskeleton organization and function. This signaling pathway involves RAC1, but not RHO, activation. Necessary for proper wound healing. In the course of wound healing, required for the phagocytotic cup formation preceding macrophage phagocytosis of apoptotic neutrophils. Responsible for integrin beta-2 (ITGB2)-mediated macrophage adhesion and, to a lesser extent, contributes to beta-3 (ITGB3)-mediated adhesion. Does not affect integrin beta-1 (ITGB1)-mediated adhesion (By similarity).
Rsrc1	66880	protein binding	RSRC1	51319	protein binding	Plays a role in pre-mRNA splicing. Involved in both constitutive and alternative pre-mRNA splicing. May have a role in the recognition of the 3' splice site during the second step of splicing.
Ccdc132	73288	molecular_function	CCDC132	55610	-	-
Tusc3	80286	magnesium ion transmembrane transporter activity	TUSC3	7991	dolichyl-diphosphooligosaccharide-protein glycotransferase activity, magnesium ion transmembrane transporter activity	Magnesium transporter. May be involved in N-glycosylation through its association with N-oligosaccharyl transferase.
Hivep1	110521	DNA binding, DNA binding, HMG box domain binding, metal ion binding, nucleic acid binding, protein binding, zinc ion binding	HIVEP1	3096	DNA binding, protein binding, zinc ion binding	This protein specifically binds to the DNA sequence 5'-GGGACTTTCC-3' which is found in the enhancer elements of numerous viral promoters such as those of SV40, CMV, or HIV-1. In addition, related sequences are found in the enhancer elements of a number of cellular promoters, including those of the class I MHC, interleukin-2 receptor, and interferon-beta genes. It may act in T-cell activation. Involved in activating HIV-1 gene expression. Isoform 2 and isoform 3 also bind to the IPCS (IRF1 and p53 common sequence) DNA sequence in the promoter region of interferon regulatory factor 1 and p53 genes and are involved in transcription regulation of these genes. Isoform 2 does not activate HIV-1 gene expression. Isoform 2 and isoform 3 may be involved in apoptosis
Rims2	116838	Rab GTPase binding, ion channel binding, metal ion binding, protein binding, protein domain specific binding, protein heterodimerization activity	RIMS2	9699	Rab GTPase binding, metal ion binding, protein binding	Rab effector involved in exocytosis. May act as scaffold protein.
Tox	252838	DNA binding	TOX	9760	DNA binding	May play a role in regulating T-cell development (By similarity).
Cntn4	269784	-	CNTN4	152330	-	Contactins mediate cell surface interactions during nervous system development. Has some neurite outgrowth-promoting activity. May be involved in synaptogenesis.

CHAPTER V

DISCUSSION

We conducted genome-wide statistical analyses using publicly available microarray expression datasets that report gene expression level from different disease conditions or biological process-related mechanisms. The present findings show that genes containing intragenic L1s are more likely to be repressed in many cancers and that downregulation level depends on the degree of L1 hypomethylation. Levels of L1 hypomethylation vary at each locus of a tumor and they may change throughout the multistage carcinogenesis process. In general, more advanced stages of cancer are associated with a greater degree of hypomethylation [15, 72-76]. Therefore, intragenic L1s may promote cancer progression in part due to increasing degrees of gene repression and numbers of repressed genes.

Even though L1s are still active retrotransposons and the new insertion can be identified as L1 insertion dimorphisms (LIDs). These LIDs were reported to cause many diseases [5]. In human, the majority of LIDs are truncated and localized in intergenic regions [77-79]. Therefore, the number of long LIDs in intragenic region are low. Moreover, L1Base hosts locations of L1s from both human (human genome version 36) and mouse (mouse genome version 35). Most L1s in L1Base are not newly inserted and very few may represent common LIDs. This supports our statistical analysis between L1 locations in host genes and their expression profiles that LIDs should not regulate these genes.

Many studies have reported the methylation of tumor suppressor gene promoters in cancer cells; this epigenetic-based regulation has become a potential candidate for biomarker and therapeutic target development. As genome-wide hypomethylation is common to many cancer types [15], our results support that the global hypomethylation could down regulate some genes in cancer. Furthermore, there may be several hypomethylation-mediated *cis*-suppressor elements, including intragenic L1s. Hence, the hypomethylation sites and repressed genes described here represent a vast number of molecular targets and diagnostic markers.

However, not all intragenic L1s can repress gene expression in cancer, and intragenic L1s may regulate genes through several distinct mechanisms. Although human L1 sequence analysis showed that there are high number of conserved intragenic L1s, their sequences and distributions vary markedly. The methylation levels of some L1 loci are independent of genome wide L1 hypomethylation in cancer [26]. Rangwala et al. [80] reported that levels of L1 RNA vary in normal cells [26], suggestive of other factors influencing L1 expression. Therefore, it is important to further explore L1 and genome characteristics that may determine their repression properties in cancer cells.

This work also established a new finding that L1s could be complex with AGO2 to silence genes under a global hypomethylation environment. We hypothesized that L1 hypomethylation increases L1 RNA levels, which is then transcribed in both direction from sense and antisense promoters and formed dsRNAs. These dsRNAs are converted to siRNAs that form RISC complexes with AGO2, resulting in disruption of mRNA processing of genes containing hypomethylated L1s. This hypothesis is supported by the experiment in [81] in which AGO2 was reported to commonly target L1 RNA and was proposed to prevent retrotransposition events [82]. However, the role of the L1-RNA-AGO2 complexes derived from the majority of retrotranspositionally incompetent elements was unknown. Moreover, there are several mechanisms by which RISC can regulate gene expression [83].

Our experiments show strong statistical support that intragenic L1 hypomethylation represses genes via a post-transcriptionally mechanism, based on siRNA and AGO2. However, it is possible that there are other mechanisms that should be further explored such as the interference with the elongating RNA Pol2 transcribing their host genes or formation of chromatin complex in relation with L1 methylation level.

Based on the available data of mouse L1 sequences from L1Base, the conservation of mouse L1 sequences were analyzed. This statistical result shows mouse intragenic L1s also have more conserved than intergenic ones. But these conservations in mouse are less conserved than human. The lower conservation of mouse L1 is particularly marked in the 5' UTR, in which variation in monomer repeats was shown previously to control L1 promoter activity [4, 34]. The significantly higher number of monomer in intragenic compared with intergenic L1s suggests that intragenic L1s are more transcriptionally active. The difference in mechanism of transcriptional control in human and mouse L1 may suggest that the transcriptionally active L1s have acquired biologically important functions independently in different mammalian lineages, i.e., convergent evolution [84].

The L1 density is greater in mouse than human, including L1s within genes. The intergenic L1s density is greater in autosome and X-chromosome but less in Y-chromosome. On the X-chromosome, L1 density is highest in mouse and human. This result supports the thought that L1s act as boosters for spreading of inactive genes from the center of inactivation [7, 8]. For autosomal and X-chromosomes, intergenic L1 densities are much higher than intragenic ones. The lower density of intragenic versus intergenic L1 in both species suggests that L1 retrotransposition into genes is likely to be deleterious in some genes and would selected against in evolution [85]. This purifying selection in the X and autosomes could be facilitated by recombination of homologous chromosomes or homologous recombination DNA break repair. Y-chromosome is hemizygote and majority of the chromosome lacks homologous recombination. If the role of intergenic L1s is related to increasing of homologous recombination rate, intergenic L1s in Y-chromosome may have no function and can be considered as junk DNA. Rearrangements and deletion mutations of intergenic L1s in Y-chromosome should not affect fitness and the L1s should be continuously lost during evolution. In contrast to intergenic L1s, intragenic L1s possess gene regulatory function and should be conserved [9, 11, 12]. As a result, in Y-chromosome, intragenic L1 density is higher than intergenic for both human and mouse.

Although these tests are suggestive for possible function of intragenic L1s in human and mouse such as X-inactivation, there are alternative explanations that do not require L1s to have functions. For example, some rodent species thought to lack potentially mobile L1 still have X-inactivation [86]. In addition, accumulation of L1 elements in X still continues even when X-inactivation is not needed in *Tokudaia osimensis*, an XO species [87]. The reason for conservation of intragenic L1s could stem from the genomic context that these elements are located, i.e., genic regions are likely to be more constrained by background selection, hence conservation of intragenic L1s does not necessarily imply function. Therefore, apart from direct testing for function, e.g., L1 ablation by genome editing tool, comparison among a greater range of mammalian species could provide insights into putative functions of conserved L1s. This is because a recent intragenic L1 element is unlikely to have a function and is tolerated because it has minor phenotypic consequence. On the other hand, if an intragenic L1 element has persisted for a long evolutionary time, it may have acquired a new function which can be constrained by purifying selection.

The greater conservation and possible activity of intragenic L1 in both human and mouse is suggestive of function. We investigated whether intragenic L1 might play a role in gene regulation in early embryogenesis. Significant associations were found for down regulated genes with intragenic L1 and down regulation of the genes, starting from the two-cell to the morula stage in mouse, whereas associations were significant for eight-cell to blastocyst in human. The different “L1 associated with down regulation” (LaD) profiles align well with the varying zygotic gene activations and the levels of global hypomethylation among mammals [41]. In particular, human zygotic activation starts during the four to eight-cell divisions, whereas starts early from two-cell division in mouse. Furthermore, mouse embryos undergo demethylation after fertilization to become hypomethylated, and establish new methylation patterns at the blastocyst stage [88]. The mouse LaD pattern thus agrees with the global hypomethylation profile during zygotic activation (Table 9). Human embryogenesis differs from mouse in the timing of zygotic activation [41, 45] and the human LaD pattern aligns with the slower onset of activation in mouse (Table 8).

Although the timing of zygotic gene activation differs between mouse and human, intragenic L1 appears to be important for controlling gene expression in both species. Among the orthologous genes obtained from intersecting the mouse and human LaD gene sets (Table 10), two genes have previously been reported with roles in embryogenesis. *Kcnq1* was reported to be a paternally imprinted gene that is down regulated during embryogenesis development [89]. The Cyclic GMP-Inhibited Phosphodiesterase 3A (*PDE3A*) gene functions in the cGMP-PKG signaling pathway [90]. *PDE3A* must be inhibited to allow expression of other important genes during physiological development. Hence, under the global hypomethylation state during zygotic activity, intragenic L1 may be expressed which down regulates these genes, perhaps by the same AGO2-dependent mechanism as described in cancer cells [12].

REFERENCES

1. Lander, E.S., et al., Initial sequencing and analysis of the human genome. Nature, 2001. 409(6822): p. 860-921.
2. Mouse Genome Sequencing, C., et al., Initial sequencing and comparative analysis of the mouse genome. Nature, 2002. 420(6915): p. 520-62.
3. Hancks, D.C. and H.H. Kazazian, Jr., Active human retrotransposons: variation and disease. Curr Opin Genet Dev, 2012. 22(3): p. 191-203.
4. Severynse, D.M., C.A. Hutchison, 3rd, and M.H. Edgell, Identification of transcriptional regulatory activity within the 5' A-type monomer sequence of the mouse LINE-1 retroposon. Mamm Genome, 1992. 2(1): p. 41-50.
5. Cordaux, R. and M.A. Batzer, The impact of retrotransposons on human genome evolution. Nat Rev Genet, 2009. 10(10): p. 691-703.
6. Lyon, M.F., X-chromosome inactivation: a repeat hypothesis. Cytogenet Cell Genet, 1998. 80(1-4): p. 133-7.
7. Bailey, J.A., et al., Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. Proc Natl Acad Sci U S A, 2000. 97(12): p. 6634-9.
8. Wang, Z., et al., Evidence of influence of genomic DNA sequence on human X chromosome inactivation. PLoS Comput Biol, 2006. 2(9): p. e113.
9. Wanichnopparat, W., et al., Genes associated with the cis-regulatory functions of intragenic LINE-1 elements. BMC Genomics, 2013. 14: p. 205.
10. Khowutthitham, S., et al., Intragenic long interspersed element-1 sequences promote promoter hypermethylation in lung adenocarcinoma, multiple myeloma and prostate cancer. Genes & Genomics, 2012. 34(5): p. 517-528.
11. Kano, H., et al., L1 retrotransposition occurs mainly in embryogenesis and creates somatic mosaicism. Genes Dev, 2009. 23(11): p. 1303-12.
12. Aporntewan, C., et al., Hypomethylation of intragenic LINE-1 represses transcription in cancer cells through AGO2. PLoS One, 2011. 6(3): p. e17934.

13. Feinberg, A.P. and B. Tycko, The history of cancer epigenetics. Nat Rev Cancer, 2004. 4(2): p. 143-53.

14. Herman, J.G., Epigenetic changes in cancer and preneoplasia. Cold Spring Harb Symp Quant Biol, 2005. 70: p. 329-33.

15. Chalitchagorn, K., et al., Distinctive pattern of LINE-1 methylation level in normal tissues and the association with carcinogenesis. Oncogene, 2004. 23(54): p. 8841-6.

16. Feinberg AP, V.B., Hypomethylation distinguishes genes of some human cancers from their normal counterparts. Nature, 1982. 301: p. 89-92.

17. Ehrlich, M., DNA methylation in cancer: too much, but also too little. Oncogene, 2002. 21(35): p. 5400-13.

18. Hoffmann, M.J. and W.A. Schulz, Causes and consequences of DNA hypomethylation in human cancer. Biochem Cell Biol, 2005. 83(3): p. 296-321.

19. Kongruttanachok, N., et al., Replication independent DNA double-strand break retention may prevent genomic instability. Mol Cancer, 2010. 9: p. 70.

20. Pornthanakasem, W., et al., LINE-1 methylation status of endogenous DNA double-strand breaks. Nucleic Acids Res, 2008. 36(11): p. 3667-75.

21. Wolff, E.M., et al., Hypomethylation of a LINE-1 promoter activates an alternate transcript of the MET oncogene in bladders with cancer. PLoS Genet, 2010. 6(4): p. e1000917.

22. Ball, M.P., et al., Targeted and genome-scale strategies reveal gene-body methylation signatures in human cells. Nat Biotechnol, 2009. 27(4): p. 361-8.

23. Lorincz, M.C., et al., Intragenic DNA methylation alters chromatin structure and elongation efficiency in mammalian cells. Nat Struct Mol Biol, 2004. 11(11): p. 1068-75.

24. Eller, C.D., et al., Repetitive sequence environment distinguishes housekeeping genes. Gene, 2007. 390(1-2): p. 153-65.

25. Han, J.S., S.T. Szak, and J.D. Boeke, Transcriptional disruption by the L1 retrotransposon and implications for mammalian transcriptomes. Nature, 2004. 429(6989): p. 268-74.

26. Phokaew, C., et al., LINE-1 methylation patterns of different loci in normal and cancerous cells. Nucleic Acids Res, 2008. 36(17): p. 5704-12.

27. Penzkofer, T., T. Dandekar, and T. Zemojtel, L1Base: from functional annotation to prediction of active LINE-1 elements. Nucleic Acids Res, 2005. 33(Database issue): p. D498-500.

28. Zemojtel, T., et al., Exonization of active mouse L1s: a driver of transcriptome evolution? BMC Genomics, 2007. 8: p. 392.

29. Edgar, R., M. Domrachev, and A.E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res, 2002. 30(1): p. 207-10.

30. Barrett, T., et al., NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res, 2009. 37(Database issue): p. D885-90.

31. Fanning, T.G. and M.F. Singer, LINE-1: a mammalian transposable element. Biochim Biophys Acta, 1987. 910(3): p. 203-12.

32. Kazazian, H.H., Jr. and J.V. Moran, The impact of L1 retrotransposons on the human genome. Nat Genet, 1998. 19(1): p. 19-24.

33. Schichman, S.A., et al., L1 A-monomer tandem arrays have expanded during the course of mouse L1 evolution. Mol Biol Evol, 1993. 10(3): p. 552-70.

34. DeBerardinis, R.J. and H.H. Kazazian, Jr., Analysis of the promoter from an expanding mouse retrotransposon subfamily. Genomics, 1999. 56(3): p. 317-23.

35. Kitkumthorn, N. and A. Mutirangura, Long interspersed nuclear element-1 hypomethylation in cancer: biology and clinical applications. Clin Epigenetics, 2011. 2(2): p. 315-30.

36. Morris, K.V., siRNA-mediated transcriptional gene silencing: the potential mechanism and a possible role in the histone code. Cell Mol Life Sci, 2005. 62(24): p. 3057-66.

37. Carthew, R.W. and E.J. Sontheimer, Origins and Mechanisms of miRNAs and siRNAs. Cell, 2009. 136(4): p. 642-55.

38. Agrawal, N., et al., RNA interference: biology, mechanism, and applications. Microbiol Mol Biol Rev, 2003. 67(4): p. 657-85.

39. Semizarov, D., P. Kroeger, and S. Fesik, siRNA-mediated gene silencing: a global genome view. Nucleic Acids Res, 2004. 32(13): p. 3836-45.

40. Messerschmidt, D.M., B.B. Knowles, and D. Solter, DNA methylation dynamics during epigenetic reprogramming in the germline and preimplantation embryos. Genes Dev, 2014. 28(8): p. 812-28.

41. Morgan, H.D., et al., Epigenetic reprogramming in mammals. Hum Mol Genet, 2005. 14 Spec No 1: p. R47-58.

42. Shi, L. and J. Wu, Epigenetic regulation in mammalian preimplantation embryo development. Reprod Biol Endocrinol, 2009. 7: p. 59.

43. Golbabapour, S., M.A. Abdulla, and M. Hajrezaei, A concise review on epigenetic regulation: insight into molecular mechanisms. Int J Mol Sci, 2011. 12(12): p. 8661-94.

44. Schultz, R.M., Regulation of zygotic gene activation in the mouse. Bioessays, 1993. 15(8): p. 531-8.

45. Niakan, K.K., et al., Human pre-implantation embryo development. Development, 2012. 139(5): p. 829-41.

46. Fulka, H., et al., DNA methylation pattern in human zygotes and developing embryos. Reproduction, 2004. 128(6): p. 703-8.

47. Kuriakose, M.A., et al., Selection and validation of differentially expressed genes in head and neck cancer. Cell Mol Life Sci, 2004. 61(11): p. 1372-83.

48. Scotto, L., et al., Identification of copy number gain and overexpressed genes on chromosome arm 20q by an integrative genomic approach in cervical cancer: potential role in progression. Genes Chromosomes Cancer, 2008. 47(9): p. 755-65.

49. Shames, D.S., et al., A genome-wide screen for promoter methylation in lung cancer identifies novel methylation markers for multiple malignancies. PLoS Med, 2006. 3(12): p. e486.

50. Kim, B.Y., et al., Feature genes of hepatitis B virus-positive hepatocellular carcinoma, established by its molecular discrimination approach using prediction analysis of microarray. Biochim Biophys Acta, 2004. 1739(1): p. 50-61.

51. Mecham, B.H., et al., Sequence-matched probes produce increased cross-platform consistency and more reproducible biological results in microarray-based gene expression measurements. Nucleic Acids Res, 2004. 32(9): p. e74.

52. Dyrskjot, L., et al., Gene expression in the urinary bladder: a common carcinoma in situ gene expression signature exists disregarding histopathological classification. Cancer Res, 2004. 64(11): p. 4040-8.

53. D'Errico, M., et al., Genome-wide expression profile of sporadic gastric cancers with microsatellite instability. Eur J Cancer, 2009. 45(3): p. 461-9.

54. Chandran, U.R., et al., Gene expression profiles of prostate cancer reveal involvement of multiple molecular pathways in the metastatic process. BMC Cancer, 2007. 7: p. 64.

55. Yu, Y.P., et al., Gene expression alterations in prostate cancer predicting tumor aggression and preceding development of malignancy. J Clin Oncol, 2004. 22(14): p. 2790-9.

56. Mishra, P.J., et al., Carcinoma-associated fibroblast-like differentiation of human mesenchymal stem cells. Cancer Res, 2008. 68(11): p. 4331-9.

57. Welch, B.L., The generalisation of student's problems when several different population variances are involved. Biometrika, 1947. 34(1-2): p. 28-35.

58. Smit, A.F., et al., Ancestral, mammalian-wide subfamilies of LINE-1 repetitive sequences. J Mol Biol, 1995. 246(3): p. 401-417.

59. Medstrand, P., L.N. van de Lagemaat, and D.L. Mager, Retroelement distributions in the human genome: variations associated with age and proximity to genes. Genome Res, 2002. 12(10): p. 1483-95.

60. Goodier, J.L., et al., A novel active L1 retrotransposon subfamily in the mouse. Genome Res, 2001. 11(10): p. 1677-85.

61. dos Santos Silva, I., Dealing with confounding in the analysis, in Cancer Epidemiology: Principles and Methods, I. dos Santos Silva, Editor. 1999, International Agency for Research on Cancer: Lyon, France.

62. Schmitter, D., et al., Effects of Dicer and Argonaute down-regulation on mRNA levels in human HEK293 cells. Nucleic Acids Res, 2006. 34(17): p. 4801-15.

63. Khorshid, M., C. Rodak, and M. Zavolan, CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. Nucleic Acids Res, 2011. 39(Database issue): p. D245-52.

64. Xie, D., et al., Rewirable gene regulatory networks in the preimplantation embryonic development of three mammalian species. Genome Res, 2010. 20(6): p. 804-15.

65. Hata, K. and Y. Sakaki, Identification of critical CpG sites for repression of L1 transcription by DNA methylation. Gene, 1997. 189(2): p. 227-34.

66. Katayama, S., et al., Antisense transcription in the mammalian transcriptome. Science, 2005. 309(5740): p. 1564-6.

67. Ngamphiw, C., S. Tongsima, and A. Mutirangura, Roles of intragenic and intergenic L1s in mouse and human. PLoS One, 2014. 9(11): p. e113434.

68. Beraldi, R., et al., Expression of LINE-1 retroposons is essential for murine preimplantation development. Mol Reprod Dev, 2006. 73(3): p. 279-87.

69. Eppig, J.T., et al., The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. Nucleic Acids Res, 2012. 40(Database issue): p. D881-6.

70. Ashburner, M., et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet, 2000. 25(1): p. 25-9.

71. Rebhan, M., et al., GeneCards: a novel functional genomics compendium with automated data mining and query reformulation support. Bioinformatics, 1998. 14(8): p. 656-64.

72. Ogino, S., et al., A cohort study of tumoral LINE-1 hypomethylation and prognosis in colon cancer. J Natl Cancer Inst, 2008. 100(23): p. 1734-8.

73. Pattamadilok, J., et al., LINE-1 hypomethylation level as a potential prognostic factor for epithelial ovarian cancer. Int J Gynecol Cancer, 2008. 18(4): p. 711-7.

74. Tangkijvanich, P., et al., Serum LINE-1 hypomethylation as a potential prognostic marker for hepatocellular carcinoma. Clin Chim Acta, 2007. 379(1-2): p. 127-33.

75. Smith, I.M., et al., DNA global hypomethylation in squamous cell head and neck cancer associated with smoking, alcohol consumption and stage. Int J Cancer, 2007. 121(8): p. 1724-8.

76. Shuangshoti, S., et al., Line-1 hypomethylation in multistage carcinogenesis of the uterine cervix. Asian Pac J Cancer Prev, 2007. 8(2): p. 307-9.

77. Sheen, F.M., et al., Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. Genome Res, 2000. 10(10): p. 1496-508.

78. Badge, R.M., R.S. Alisch, and J.V. Moran, ATLAS: a system to selectively identify human-specific L1 insertions. Am J Hum Genet, 2003. 72(4): p. 823-38.

79. Pornthanakasem, W. and A. Mutirangura, LINE-1 insertion dimorphisms identification by PCR. Biotechniques, 2004. 37(5): p. 750, 752.

80. Rangwala, S.H., L. Zhang, and H.H. Kazazian, Jr., Many LINE1 elements contribute to the transcriptome of human somatic cells. Genome Biol, 2009. 10(9): p. R100.

81. Watanabe, T., et al., Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. Nature, 2008. 453(7194): p. 539-43.

82. Yang, N. and H.H. Kazazian, Jr., L1 retrotransposition is suppressed by endogenously encoded small interfering RNAs in human cultured cells. Nat Struct Mol Biol, 2006. 13(9): p. 763-71.

83. Pratt, A.J. and I.J. MacRae, The RNA-induced silencing complex: a versatile gene-silencing machine. J Biol Chem, 2009. 284(27): p. 17897-901.

84. Waters, P.D., et al., Evolutionary history of LINE-1 in the major clades of placental mammals. PLoS One, 2007. 2(1): p. e158.

85. Boissinot, S., A. Entezam, and A.V. Furano, Selection against deleterious LINE-1-containing loci in the human lineage. Mol Biol Evol, 2001. 18(6): p. 926-35.

86. Cantrell, M.A., B.C. Carstens, and H.A. Wichman, X chromosome inactivation and Xist evolution in a rodent lacking LINE-1 activity. PLoS One, 2009. 4(7): p. e6252.

87. Scott, L.A., et al., X accumulation of LINE-1 retrotransposons in Tokudaia osimensis, a spiny rat with the karyotype XO. Cytogenet Genome Res, 2006. 112(3-4): p. 261-9.

88. Reik, W., W. Dean, and J. Walter, Epigenetic reprogramming in mammalian development. Science, 2001. 293(5532): p. 1089-93.

89. Lewis, A., et al., Epigenetic dynamics of the Kcnq1 imprinted domain in the early embryo. Development, 2006. 133(21): p. 4203-10.

90. Li, M., et al., The role of cilostazol, a phosphodiesterase 3 inhibitor, on oocyte maturation and subsequent pregnancy in mice. PLoS One, 2012. 7(1): p. e30649.

APPENDIX

Definitions of human L1 characteristics

Human L1 sequences were downloaded from L1Base. These elements were annotated with important features for L1 activities. We group them according to where these features can be found, namely, 5' UTR, ORF1, ORF2 and 3' UTR. We put the overall features, e.g., G-C content and cannot be placed according to a specific location on L1 in a group called “Overall”. Detailed information on finding of each feature can be found from the cited references. The measurement outputs are in two forms, categorical (e.g., conserved/mutated, L1M/L1PA for chi-square test) and non-categorical (e.g., %A, %T for student’s *t*-test).

Overall

- ORF StartStop: check the presence of valid methionine start and stop codons in both ORF1 and ORF2 in the form of (M*, M*). This feature is reported as conserved, ORF1 conserved, ORF2 conserved, or mutated.
- Find TSDs : search for target-site-duplications (TSD) flanking the L1 element. These TSDs, which span more than 10 nucleotides (nt), will be considered as valid TSDs. This feature is reported as the number of valid TSDs.
- CpG Islands : count the number of annotated CpG islands. This feature is reported as the number of CpG islands.
- G-C Content : calculate the percentage of G-C of the L1 element in a 50nt window. This feature is reported as G-C percentage.
- Intactness Score: calculate the overall score of features. Every intact feature (conserved) awards one point. This feature is reported as the total points earned.

5' UTR

- Ta1-nd/d : check for the presence of Ta1 subfamily where “nd” is no deletion and “d” is for deletion of Ta1 located in 5' UTR.

- Runx3 Site, Runx3 ASP : check for the presence of the intact RUNX3 and RUNX3 Anti-Sense-Promoter (ASP) binding motifs, respectively.
- SRY Site 1, SRY Site 2 : check for the presence of the intact first and second SRY (sex-determining region Y) binding motifs, respectively.
- YY1 BoxA+BoxA : check for the presence of an intact YY1 binding motif.
- TF nkx-2.5, TF nkx-2.5B : check for the presence of the first transcription factor Nkx-2.5 and second Nkx-2.5B sites, respectively.

ORF1

- ORF1 conserved: check for the conservation of amino acid sequence of ORF1. If the final score passes the significance cutoff value, it will be reported as "conserved", otherwise mutated.
- REKG235, ARR260, YPAKLS282 : these features are short amino acid sequences at different locations. It is said to be conserved if an amino acid sequence matches otherwise mutated. For example, check the amino acids position 235 to 238 if they match REKG (REKG235), see note below.
- ORF1 gaps, ORF1 frameshifts, ORF1 stops: count the number of gaps, frameshifts and stop codons (TAA, TAG, TGA) in ORF1, respectively.
- ORF1 %A, ORF1 %T, ORF1 CAI : calculate nucleotide percentages of A (%A), T (%T), and the Codon Adaptation Index (CAI) of ORF1, respectively.

ORF2

- ORF2 conserved: check for the conservation of amino acid sequence of ORF2. If the final score passes the significance cutoff value, it will be reported as "conserved", otherwise mutated.
- Ta0/Ta1 SSVs : determine the shared sequence variant (SSV) subfamily of L1s from Ta0-1 locus including Ta-0/L1PA2, Ta-1, and L1PA5.

- L1M/L1PA Discrimination : check if ORF2 contains either mammalian L1 (L1M) or primate L1 (L1PA).
- N14, E43, Y115, D145, N147, T192, D205, SDH228, R363, FADD700, HMKK1091, SSS1096, I1220, S1259 : check for the conservation of amino acid residues at these particular loci (see note below)
- ORF2 gaps, ORF2 frameshifts, ORF2 stops: count the number of gaps, frameshifts and stop codons (TAA, TAG, TGA) in ORF2.
- ORF2 %A, ORF2 %T, ORF2 CAI : calculate nucleotide percentages of A (%A), T (%T), and the Codon Adaptation Index (CAI) of ORF2, respectively.
- ORF1&2 %A, ORF1&2 %T : calculate %A and %T for ORF1&2, respectively.

3' UTR

- Ta SSVs : determine Ta families, including GAGA (L1PA2-L1PA5), GAGG or GCGA (intermediate L1), ACGA or ACGG (preTa L1), and ACAG (Ta Element).
- Poly A Signal : test for conservation of poly A patterns. The consensus ‘AATAAA’ or ‘AATTAAA’ are considered as the two valid patterns.

Note

- Amino acid residues changes: check for the intactness of amino acid residues on the ORF1 and ORF2. For example, REKG235 in ORF1 refer to checking of the amino acid residues starting at 235 on the ORF1 start codon whether they match the sequence ‘R-E-K-G’, respectively. N14 in ORF2 refer to checking the residue ‘N’ at the position 14 on the ORF2 start codon.

Definitions of mouse L1 characteristics

Mouse L1 sequences were downloaded from L1Base. These elements were annotated with important features for L1 activities. Some features in mouse L1 differ from that of human, e.g., monomer is only available in mouse. We group them according to where these features can be found, namely, 5' UTR, ORF1, ORF2 and 3' UTR. We put the overall features, e.g., G-C content and cannot be placed according to a specific location on L1 in a group called “Overall”. Detailed information on finding of each feature can be found from the cited references. The measurement outputs are in two forms, categorical (e.g., conserved/mutated, for chi-square test) and non-categorical (e.g., %A, %T for student’s t-test).

Overall

- ORF StartStop: check the presence of valid methionine start and stop codons in both ORF1 and ORF2 in the form of (M*, M*). This feature is reported as conserved, ORF1 conserved, ORF2 conserved, or mutated.
- Monomer Family : classify mouse L1 families to F, A, TF, and GF, using the last monomer.
- CpG Islands: count the number of annotated CpG islands.
- G-C Content: calculate the percentage of G-C content of the L1 element in a 50nt-window.
- Intactness Score: calculate the overall score of categorical (conserved/mutated) features. Every intact feature (conserved) awards one point.

5' UTR

- SA-154 : check the conservation of this splice site (see note below).
- Number of Monomers : count the number of mouse L1 promoter monomers.
- Number of Monomer Splice Sites: count the number of monomer splice sites.

ORF1

- ORF1 conserved: check for the conservation of ORF1.
- 66/42 Monomers repeat : check the pattern of the monomers in ORF1, such as 66-42-42 monomers.
- REKG235, ARR260, YPAKLS282 : check for the intactness of amino acid residues at these particular loci (235, 260 and 282 positions, respectively, on the mouse L1 ORF1, see note below).
- SA+106, SA+120, SD+29, SD+52, SD+106, SD+288, SD+350 : check for the conservation of these splice-sites at the respective positions (see note below).
- ORF1 gaps, ORF1 frameshifts, ORF1 stops: count the number of gaps, frameshifts and stop codons (TAA, TAG, TGA) in ORF1, respectively.

ORF2

- ORF2 conserved: check for the conservation of ORF2.
- N14, E43, Y115, D145, N147, T192, D205, SDH288, R363, FADD700, HLKK1091, STS1096, I1220, T1259 : check for the intactness of amino acid residues at these particular loci (see note below).
- SA+1930, SA+4117, SD+1881, SD+2036, S_SA+1237, BG_SA+4578, BG_SD+4694, BG_SD+4903 : check for the conservation of these splice-sites (see note below).
- ORF2 gaps, ORF2 frameshifts, ORF2 stops: count the number of gaps, frameshifts and stop codons (TAA, TAG, TGA) in ORF2.

3' UTR

- Poly A Signal [7]: check for the conservation of two poly-A patterns, namely ‘AATAAA’ or ‘AATTAAA’.
- SD+5094, SA+5260, SA+5614 : check for the conservation of these splice-sites (see note below).

Note

- Splice site loci: check for the conservation of mouse L1 splice-site (SD: splice donor, SA: splice acceptor, BG_: sites found in L1 inserted within an intron of the beige gene, S_: splice-site on the sense strand of L1 (without this, denotes the splice site on the antisense strand of L1), '-': represents nucleotide position backward from ORF1 start site, '+' : represents nucleotide position from ORF1 start site. The number followed the '+' or '-' represents the nucleotide position.
- Amino acid residues changes: check for the intactness of amino acid residues on the ORF1 and ORF2. For example, REKG235 in ORF1 refer to checking of the amino acid residues starting at 235 on the ORF1 start codon whether they match the sequence 'R-E-K-G', respectively. N14 in ORF2 refer to checking the residue 'N' at the position 14 on the ORF2 start codon.

VITA

Chumpol Ngamphiw is now Ph.D. student in the interdisciplinary program in Biomedical Sciences, Chulalongkorn University. He was born in Nakhon Nayok, Thailand on March 19, 1975. He received his B.Eng. and M.Eng. in Electrical Engineering from King Mongkut's University of Technology North Bangkok in 1997 and 2004, respectively.

After finishing his Master degree, he had starting his work as assistant researcher of Dr. Sissades Tongsima in the Bioinformatics Laboratory at the National Center for Genetic Engineering and Biotechnology (BIOTEC) on June, 2004. The first project under, he had assigned to create the Thailand SNP database, which is the first single nucleotide polymorphism (SNP) of Thai population. From this first project bring him to more studying on the biology. During the time of his work, he had collaborating with a lot of famous Thai geneticist and the foreigner. On 2009, he had collaborating with Prof. Apiwat Mutirangura to help him in bioinformatics analysis. After that, he began his Ph.D. studying on Biomedical Science program at Chulalongkorn University on 2010.