# 國立交通大學

# 生物資訊及系統生物研究所

# 碩士論文



靈長類特有之微小核醣核酸叢集的比較基因體學分析 1896 Comparative analysis of C19MC in primate genomes

# 研究生:潘承宗

指導教授:林勇欣

中華民國九十九年六月

# 靈長類特有之微小核醣核酸叢集的比較基因體學分析 Comparative analysis of C19MC in primate genomes

研 究 生:潘承宗

Student: Cheng-Tsung Pan

指導教授:林勇欣 Advisor: Yeong-Shin Lin



A Thesis Submitted to Institute of Bioinformatics and Systems Biology College of Biological Science and Technology National Chiao Tung University in Fulfillment of the Requirements for the Degree of Master in Bioinformatics and Systems Biology

> June 2010 Hisinchu, Taiwan, Republic of China 中華民國九十九年六月

## **Chinese Abstract**

Alu 單元和微小核醣核酸〔microRNA〕是兩個截然不同的基因體序列。在近幾年, 兩者皆被廣泛研究。Alu 單元是專一出現在靈長類動物的短散布重複因子〔SINE〕中佔 多數的因子之一,且它們不轉譯出有功能的蛋白質。在最近的研究中指出 Alu 單元可能 和癌症的產生有所關聯。其他的研究也認為 Alu 單元可能影響基因的功能。基於這些發 現,可以合理的認為,Alu 單元在靈長類的演化中可能扮演某種很重要的角色。在人類 第19號染色體上的微小核醣核酸叢集〔C19MC〕中的 Alu 單元和微小核醣核酸,被發 現有一種很特殊的位置上的靠近。這個不尋常的片段可能是起因於一連串的複製 〔duplication〕事件。在這篇研究中,我們分析了這些可能是複製單元的序列,無論是 在同物種的同源基因 [paralogs] 之間,或是異物種同源基因對 [ortholog pairs] 之間, 希望能辨識出序列中每個位置所遭受的選擇壓力 [selection pressure]。我們發現在恆 河猴 C19MC 中的同物種同源基因序列,相對於人類有較高的變異度。同時我們的全面 性和區域性的分析也指出,無論在人類或恆河猴的 C19MC 中,不同的序列區域遭受到 不同的選擇力量。在表現子〔exon〕區域有最高的變異性,然而在中間區域靠近微小核 醣核酸的週邊,則有較高的保留度。此外,先前的研究指出 Alu 可能會被微小核醣核酸 所抑制,因為Alu單元的擴張可能會使基因體受損。在此篇研究的第二部份,我們假設 降低 Alu 單元的表現程度,可能會使人類基因體有選擇優勢。因此我們認為在靈長類的 演化中,C19MC 可能是主要的防禦機制,用來對抗 Alu 單元的擴張。我們初步的結果指 出,無論在人類或恆河猴的基因體中,C19MC 中的微小核醣核酸傾向針對 AluS 而非 AluI 和 AluY。這個發現支持了我們的假說, C19MC 的出現可能是在 Alu 擴張之後的一種防 禦反應。

## **English Abstract**

Alu elements and microRNAs are two different types of genomic sequences. Each of them has been extensively studied in recent years. Alu is one of the most abundant SINEs (short interspersed nuclear element) discovered specifically in primates and it does not encode a functional protein. In recent studies, Alu was reported to be related to cancer. Some other studies further suggested that Alu may influence gene functions. Based on these findings, it is reasonable to speculate that Alu might play some important roles during primate evolution. An unusual cluster of positional proximity of Alu and microRNA was found in human chromosome 19 microRNA cluster (C19MC). This unusual fragment seems to be derived from a serious of duplication events. In this study, we analyzed the sequences of the possible duplication unit either in paralogs within a species or ortholog pairs between species to identify the selection pressures on each nucleotide site. We found that the sequences of rhesus C19MC are more diverged than their paralogs in human. Meanwhile, our global and local analyses revealed that in both human and rhesus C19MC, different regions are under different selection forces. The exons are more diverged, while the internal flanking regions, which are adjacent to microRNAs, are more conserved. In addition, previous studies suggested that Alu elements in human might be repressed by microRNAs, because Alu expansion might damage human genome. In the second part of our study, we hypothesized that reducing the expression level of Alu might have selectively advantage for human genome. We suggested that this C19MC cluster might be a major defender against Alu expansion during primate evolution. Our preliminary results indicate that microRNAs in C19MC tend to target AluS rather than AluJ and AluY in human and macaque genomes. This finding supports the hypothesis that the appearance of C19MC might follow Alu expansion as a response of defense.

# Acknowledgement

經過兩年非常愉快的碩士班,終於要畢業邁入下一個階段了,在新竹的兩年過得非 常的愉快,除了有很棒的實驗室夥伴和老師之外,還有在新竹認識很多親切的朋友們, 都讓我感覺研究生的生活甚至比大學更快樂。還沒有玩遍新竹的各地名勝,還沒有吃遍 新竹的各色餐廳,不知不覺地就要離開新竹了,除了畢業的喜悅外,更多的是依依不捨, 我想未來有機會我應該會經常回來交大這個令人懷念的地方。

首先當然要先感謝我的父母,這兩年來讓我全心全意在新竹生活,雖然比較少回家, 但還是能體諒我,無論經濟上或生活上,都讓我沒有後顧之憂。

也要感謝應達,你是個很好的夥伴,兩年前暑假我們剛到新竹,還沒找到房子、實 驗室裡什麼設備什麼生活用品都沒有,兩個人一起睡在實驗室,一起動手施工一起討論 要怎麼建立起這個實驗室。整個過程中你都一直很努力,也不斷忍受我的唠叨,為了讓 實驗室運作得更好,從買耗材買儀器,到自己做燈箱自己釘櫃子,雖然消耗了一年多, 但卻很開心也學到很多,這些成就感不是言語可以形容的。每天一起吃飯一起生活,也 讓我們有很多時間討論程式的寫法,學術的問題,甚至生活中的大小事,讓我學習到很 多不同的觀點和想法。還要感謝我們的另一位室友重延,因為我們常不在家,家裡的繳 費打掃等幾乎都由重延一手包辦,感謝他讓我們的家一直維持恰到好處的乾淨。

當然,能夠大膽嘗試這一切,無憂無慮的建立起實驗室,也要感謝林勇欣老師的大 力支持,無論是經費上或是精神上,有他對我們的信任和支援,我們才能在短短一年內 做這麼多事。雖然現在林先生有女朋友比較忙碌,不過這兩年還是跟我們到處玩,去香 山看夕陽,去中南部玩,還有一直想拍沒拍到的太空站,還有很多想吃的餐廳還沒吃, 兩年實在過得很快,每天在實驗室待到清晨的日子一下就過去了。除了學術上的指導之 外,更難得的是跟林老師的友誼,除了傳統的師生關係,更多的是像朋友間的幫忙和鼓 勵。

感謝志欽,雖然你常常被使喚,但為了實驗室的傳承,也只好繼續使喚你了。實驗 室的建立並不容易,希望你能從中學到很多,並把實驗室繼續的維護下去。也謝謝你在

iii

口試當天的幫忙跑腿,讓我可以專心在現場做準備。

感謝勤政,沒有你實驗室就沒有了娛樂,除了你本身就是個笑點之外,也要謝謝你 推薦電影,推薦新產品,還有在實驗室的維護上分憂解勞,協助我很多。感謝羿喬學長, 在人生規畫上和研究上的建議,還有感謝你的座位在我大混亂時暫時堆放東西。感謝家 豪學長在研發替代役申請時的協助,雖然你在實驗室待得時間太短暫,想學釣魚也還沒 學到,但我想之後還有很多機會找你教我釣魚。

感謝黃兆祺老師以及黃老師實驗室的成員們,你們是我兩年來最熟的一間實驗室, 也是在二館以外可以去泡茶聊天的唯一去處。黃老師給了我很多鼓勵和幫忙,和林勇欣 老師的嘴砲式鼓勵不同,您給我了滿多實質上的建議。

感謝熊昭老師,雖然只上過您兩門課,但還是熱心的來幫我口試,也讓我第一次有 機會去參觀一下國衛院,口試上的許多建議和叮嚀對我未來在學術上的思維很有幫助。

還要感謝交大攝影社的朋友們,雖然我只來了兩年,只參加了幾次出遊,但還是讓 我一起跟團拍照,讓我有機會辦成果展;感謝國倫一起討論和場佈,完成巨大的光牆作 品。還要感謝德芬當我的外拍模特兒,在兩年新竹生活中第一次拍出滿意的作品,雖然 1896 妳的新竹生活不太順遂,但還是要加油,一定可以撐過去。

要特別感謝介伶,兩年來一有時間就到新竹陪我,也在不知不覺中似乎要變成實驗 室的成員之一,雖然實質幫助並不多,當作精神食糧也沒有很好吃,但還是盡責地做為 頂級食材,補充海怪能量。兩年來你也過得很辛苦,第一年要重考,第二年的研究生活 不愉快,但終於妳順利地考上新學校,也要展開新生活,希望妳之後兩年的研究生活也 能過得很平安如意。

最後謝謝交大土地公,雖然每次去拜拜我都忘記帶夠零錢,也都只請祢一罐仙草蜜, 但我相信祢不會太介意,除了謝謝祢保佑我兩年平安之外,也謝謝祢保佑全交大師生。

承宗 2010/06/24 於實驗室

iv

# **Table of Contents**

Chinese Abstract i
English Abstract ii
Acknowledgment iii
Table of Contents v
List of Figures vii
List of Tables viii
List of Appendix ix
Abbreviations x
Chapter 1 General Introduction 1.1 Transposable Elements is not Just Junks
Transposable Elements is a Kind of Genome Parasites
Evolution of Alu Subfamilies
Impact of Alu on Gene Regulation
Alu Elements is a Source of Microsatellites
1.2 Primate Specific microRNA Clusters
Biogenesis and Function of microRNAs4
Primate Specific microRNA Cluster5
Consideration of Transposable Elements in Epigenetics
Chapter 2 Substitution Analysis of C19MC
2.1 Backgrounds and Hypothesis7
2.2 Materials and Methods

	2.2.1 Sequence Collection and Annotations	8
	2.2.2 Define C19MC Homologs between Human and Rhesus 10	0
	2.2.3 Nucleotides Substitution Analysis of C19MC1	1
2.3	Results and Discussion	
	2.3.1 MicroRNAs within C19MC Apeear a Characteristic Duplication Unit 1	3
	2.3.2 Homology Definition and Genetic Composition of C19MC 1	3
	2.3.3 Substitution Analysis 1	5
Chapter 3	3 C19MC probably as a defender against Alu Elements	
3.1	Backgrounds and Hypothesis18	8
3.2	Materials and Methods	
	3.2.1 Determine Mature microRNA Expression Profile from	
	Next-Generation Sequencing Data	9
	3.2.2 Whole Genome Scanning (From Seed Candidates' Point of View)	1
	3.2.3 Whole Genome Scanning (From Alu elements' Point of View)	2
3.3	Results and Discussion	
	3.3.1 C19MC has Extremely Different Expression Levels	3
	3.3.2 The Biased Distribution of C19MC Target Ability	
	against Alu Elements 22	3
	3.3.3 Analysis of Alu Elements Targeted by C19MC 2	5
Figures		7
Tables		5
Referenc	es	8
Appendi	ix	3

# List of Figures

Figure 1	The Distribution and Structure of Transposons	27
Figure 2	Distribution of Alu Subfamilies in Primate Genomes	28
Figure 3	Evolution of Alu Subfamilies	29
Figure 4	Biogenesis and Structure of microRNA	30
Figure 5	The microRNA Clusters of Human	31
Figure 6	Remarkable Duplication Unit of C19MC	32
Figure 7	Two Different Models of C19MC Transcription	33
Figure 8	Procedure of Rainbow Plot	34
Figure 9	Determine C19MC Homologs in Rhesus	35
Figure 10	Process of Substitution Calculation	36
Figure 11	Pairwise Similarity Distances of C19MC	37
Figure 12	Duplication Hotspots	38
Figure 13	The Special L1MB7 Sequences on C19MC	39
Figure 14	Overall Scatter Plot of P <sub>Substitution</sub> in Two Species	. 40
Figure 15	Proportion of Substitution in Macroscopic (Global) View	. 41
Figure 16	Substitution Analysis and Multiple Sequence Alignment	42
Figure 17	Hypothetical Model of Co-Evolution	47
Figure 18	Alignment of Sequencing Reads with C19MC Annotated Mature Region	48
Figure 19	C19MC Seed Selection from Human Sequencing Data	. 49
Figure 20	Rhesus Seed Candidates Selection	. 50
Figure 21	Distribution of Alu Target Ability of each Seed	. 51
Figure 22	Targeting Preference of Seed Candidates: Different Alu Subfamilies	. 52
Figure 23	Interchange Seeds to Target Genomes of Each Other	53
Figure 24	Percentages of Alu Subfamilies Not Targeted by C19MC	. 54

# List of Tables

Table 1	Othrlog Pairs Defined by Rainbow Dotplot in This Study 5	55
Table 2	Summary of Human C19MC Features 5	57



# List of Appendix

Appendix 1	Annotation	of	C19MC	 63	,



# Abbreviation

- C19MC <u>Chromosome 19 microRNA cluster</u>
- C14MC <u>Chromosome 14 microRNA cluster</u>
- TE(s) <u>Transposable element(s)</u>
- LTR <u>L</u>ong <u>t</u>erminal <u>r</u>epeat
- LINE <u>L</u>ong <u>i</u>nterspersed <u>e</u>lement
- SINE <u>Short interspersed e</u>lement
- SVA An element made by <u>SINE</u>, <u>v</u>ariable number of tandem repeats or <u>A</u>lu-like region

1896

- L1 LINE1, Long interspersed element 1
- RNA Pol II RNA polymerase II
- RNA Pol III RNA polymerase III
- ORF <u>Open reading frame</u>
- FAM <u>Free Alu monomer</u>
- FRAM <u>Free Alu right monomer</u>
- FLAM <u>Free Alu left monomer</u>
- Myr <u>M</u>illion <u>y</u>ea<u>r</u>s
- RISC <u>RNA-induced silencing complex</u>
- hsa- prefix which means human
- mml- prefix which means rhesus
- pd- prefix which means prediction
- mir- prefix which means microRNA

# Chapter 1 General Introduction

## **1.1** Transposable Elements is not Just Junks

#### **Transposable Elements is a Kind of Genome Parasites**

The genome composition was well understood after the genome project had completed. Our genome seems be occupied not only by the protein coding genes but also by other genomic elements. One element which is found ubiquitously and cosmically in most eukaryotic genomes is the transposable elements (TEs) due to their extent of genome contribution approximately to 45% or even more. (Lander, Linton et al. 2001) These TEs are often separated into two main categories, the DNA transposons and the retrotransposons, by their different biogenesis mechanisms. (Figure 1a) TEs from both categories could duplicate and move, so called "jumping", within the genome by either genomic recombination or insertion of new copies into new positions. This special event could cause either positive results via triggering useful gene expansion or negative results via disrupting normal gene function and inducing diseases. Besides the DNA transposons occupied 3% of human genome but stopped activating currently, the retrotransposons function as well. One type of retrotransposons, the LTR retrotransposon, is slightly inaction within human genomes. By contrast, another type of retrotransposons, the non-LTR retrotransposon, including LINE (long interspersed element), SINE (short interspersed element) and SVA (An element made by SINE, variable number of tandem repeats or Alu-like region) are expressed detectably and found uniformly among primates so that they might be considered as "alive" along with primate evolution. That might be a reason that non-LTR retrotransposons are found approximately one-third of human genome and responsible for many genetic disorders. (Mills, Bennett et al. 2007; Cordaux and Batzer 2009)

LINE1 (usually abbreviated to L1) is one of the abundant non-LTR retrotransposons found in primateswhich contains a RNA polymerase II (RNA Pol II) promoter at its 5' UTR to drive transcription and encodes RNA-binding protein from its ORF1 (open reading frame 1) and other two proteins, the endonuclease and reverse-transcriptase from ORF2.. (Swergold 1990) (Figure 1b) Although many L1 in human genome are truncated (Ostertag and Kazazian 2001), the proteins encoded by other active L1 elements continuously help the primary transposition and duplication mechanism for both LINEs themselves and even Alu elements. (Dewannieux, Esnault et al. 2003)

Alu element is another well-investigated retrotransposon in addition to L1 elements. In contrast to L1 elements which are transcribed by RNA Pol II and encode proteins, Alu elements are transcribed by RNA polymerase III (RNA Pol III) and do not encod proteins. They compose by two similar but not identical monomers with an internal A-rich linker and rely on L1 mechanism for transposition. Interestingly, Alu elements do not follow the RNA Pol III terminator for stopping transcription but extend their transcript into the downstream flanking sequence until a terminator is found or the activity of RNA Pol III is reduced. **(Figure 1b)** (Batzer and Deininger 2002) In the opinion of selfish theory which assumes genetic components or genes devoted to keep themselves alive without nature selection. According to this theory, Alu elements are probably the most successful TEs not only because of their large territory in primate genomes but also their economical transposition mechanism by hijack the L1 machinery. (Weiner 2002; Mills, Bennett et al. 2007; Comeaux, Roy-Engel et al. 2009)

#### **Evolution of Alu Subfamilies**

The oldest Alu-like elements might be the 7SL RNA genederived monomer such as FAM, FRAM and FLAM, and then evolved into recent Alu elements which is a dimeric element made by two very similar but not identical monomers. The evolution of Alu elements expansion can be separated to 3 stages by evolutionary history. The first expansion of Alu elements started from AluJ which is the ancient Alu subfamily appeared about ~60 to 65 million years (Myrs) ago. They might accumulate many mutations and become truncation through the evolutionary time. After AluJ appearance, the AluS burst out its expansion after 15 Myrs later that caused itself rise to the major amount of Alu subfamily recently. **(Figure 2)** In our study, we assumed the AluS burst is the primary crisis of genome damage due to its huge amount of expansion and must be repressed by microRNAs. Finally, the youngest subfamily so far is AluY which continues in transposition and be polymorphic in population but inferior in amount. (Price, Eskin et al. 2004) **(Figure 3)** 

#### **Impact of Alu Elements on Gene Regulation**

More and more studies found evidences that transposable elements have an important role and impact on gene regulatory networks. For example, L1 could facilitate the emersion of new regulatory proteins and transcription factors as well as move the regulated sequences (Belancio, Roy-Engel et al. 2008); Alu elements might donate itself as the target sites of small RNA regulation (microRNAs, piwiRNAs or other kinds of RNAi). (Shankar, Grover et al. 2004) Those investigations suggest extensive functions of transposable elements which involved in gene regulatory network, epigenetics and even the popular topics, the small RNA regulation. (Polak and Domany 2006; Faulkner, Kimura et al. 2009) Most of the non-coding RNAs, such as piwiRNA, are not well described, not to mention their interact function with transposable elements. Studies indicated that not only well-known microRNAs but also the novel piwiRNAs are probably as the defender against selfish transposons. (Brennecke, Malone et al. 2008; Halic and Moazed 2009) Here, we hypothesized that the phenomenon of Alu inserting into genes and changing the original regulatory mechanism might be dangerous without control. A defense mechanism of preventing Alu from over-expansion and a monitoring mechanism must exist, and probably guarded by microRNAs inhibition. (Detail in chapter 3)

#### Alu Elements are Probably a Source of Microsatellites

Non-LTR retrotransposons, especially Alu elements could generate microsatellites by their capacity of homopolymeric tract, which means a DNA sequence made of tandem repeats with same nucleotides. Each new copy of Alu could offer microsatellite source from its middle A-rich linker region or 3' A-rich tail. (Arcot, Wang et al. 1995; Jurka and Pethiyagoda 1995) This concept was stronger while ~20% of all microsatellites shared by human and chimpanzee lie within Alu elements were found. (Kelkar, Tyekucheva et al. 2008) We expected to demonstrate an analysis of microsatellite generated by Alu elements within C19MC duplication units to be an approach to reconstruct C19MC microRNAs duplication in the future. (Not in this study)

## **1.2 Primate Specific microRNA Clusters**

#### **Biogenesis and Function of microRNAs**

Typically, microRNA genes are derived from a precursor microRNAs transcribed by RNA polymerase II as same as normal protein coding genes. These precursors then are processed into 60-70 nt long pre-microRNA by an enzyme called Drosha. After the pre-microRNA had transported into the cytoplasm, another enzyme called Dicer cleaves the pre-microRNAs by cutting

its stem loop structure and generates a hybridized RNAs with two strands. These trimmed hybridized RNAs without stem loop are mature form of microRNAs. Either one strand or both strands could function as mature microRNAs to complementary hybridize with their mRNA target sites. This hybridization (perfect match of full-length mature sequence is not necessary) recruits an enzyme complex called RISC (RNA-induced silencing complex) to bind and induce the degradation of mRNA or inhibition of translation. (Neilson and Sharp 2008; Ghildiyal and Zamore 2009) **(Figure 4)** The mRNA degradation and translation inhibition process typically depend on the match of seed region (The nucleotides from second to eighth sites at 5' end of mature microRNAs. (Kertesz, Iovino et al. 2007; Bartel 2009) This characteristic of microRNA is widely used in target site predication by bioinformatics approch. We also used this fundamental principle in our study to determine whether an Alu element is targeted by C19MC in chapter 3.

### C19MC is a Primate Specific microRNA Cluster

Vast surveys of novel microRNA discovery indicated that many microRNA genes prefer to form clusters in a related shorter genomic distance to the normal genes rather than being random distributed throughout the genomes. (Lagos-Quintana, Rauhut et al. 2003) A famous one microRNA cluster located in chromosome 19 is called C19MC which means chromosome 19 microRNAs cluster. This concentration of microRNA genes as a cluster was especially found in placental mammals (Glazov, McWilliam et al. 2008) that implies the importance and uniqueness of C19MC (the main research object in our study) in primates. Furthermore, the rapid evolution and functional diversification of two primate clusters in chromosome 19 and chromosome X were studied as a view of primate evolution. (Zhang, Peng et al. 2007; Zhang, Wang et al. 2008; Li, Liu et al. 2010) Another major character of the C19MC is that a huge amount of Alu

elements insertion into the flanking region of microRNAs. It's totally different while compared with C14MC (Another rare investigated cluster in somatic chromosome) which has plentiful microRNAs but rare transposon insertions. **(Figure 5)** Here, we focused on the C19MC and studied in-depth by more local concepts with comparative sequences analysis in nucleotides level.

#### **Consideration of Transposable Elements in Epigenetics**

The transposons activity and small RNA expression were studied and believed to be connected with epigenetics. (Costa 2008) As we mentioned that human C19MC are primary expressed in the placenta and embryonic stem cell and switched off in somatic cells, thus it makes sense to connect the C19MC into regulatory network in development. We believed that C19MC must play an importance role in primates' development. A further study provided more evidence and confidence since they found C19MC are expressed in human cancer cell by an unusual control via epigenetic mechanism. (Tsai, Kao et al. 2009) According to our speculation, some important nucleotide sites with highly substitution we identified in this study could be the dominant regulatory sites. Although these sites were not reported in transcriptional level in previous C19MC studies, they might have some role in epigenetic level so that they encounter the selection force. In short, we believed these highly substituted sites may be involved in the epigenetic control rather than the mRNA level (such as RNA structure or transcription factor binding sites), so that the selection force on these sites was preserved.

# Chapter 2 Substitution Analysis of C19MC

## 2.1 Background Introduction and Hypothesis

### **Controversial Transcription Mechanism of C19MC**

In recent years, more and more studies discovered the remarkable microRNA clusters arrangement in genome from bacteria to primates. (Lagos-Quintana, Rauhut et al. 2003) Within the primate-specific microRNA clusters called C19MC, microRNA genes are averagely located within the cluster and separated by the proximal Alu elements by approximately 100 bp spacing. (Bentwich, Avniel et al. 2005) This special rearrangement of microRNAs and Alu elements forms a distinctive duplication unit or a kind of tandem repeat. (Figure 6) We thought they are probably a duplication unit, but the mechanism of this unit both explained by molecular evolution or by cellular biology were still unknown even microRNA is a popular research topic. The transcriptional mechanism of mammalian microRNAs by RNA polymerase II is well known. (Cai, Hagedorn et al. 2004; Lee, Kim et al. 2004) However, Borchert reported these proximal Alu inside the duplication unit in C19MC could donate its own promoter for its neighbor microRNA for RNA polymerase III transcription. (Borchert, Lanier et al. 2006) (Figure 7B) This amazing finding not only suggests the new transcription model of microRNAs but also indicates the novel contribution of Alu elements: the upstream Alu elements could express downstream microRNAs. They finally showed and proved the miR-515-1, miR-517a, miR-517c and miR-519a-1 in C19MC are truly expressed using Pol III via Alu promoter *in vitro*.

Interestingly, an extremely opposite finding of C19MC transcription was reported after three years of Borchert's discovery. (Bortolin-Cavaille, Dance et al. 2009) They found a unique exon sequence which is located in the flanking of most C19MC microRNAs and is processed by RNA polymerase II via alternative splicing mechanism *in vivo*. They proved that C19MC microRNAs are the intron-encoded genes of a novel RNA Pol II transcript and their expressions depend on the alternative splicing of the novel RNA Pol II transcript. **(Figure 7A)** 

## 2.2 Materials and Methods

#### 2.2.1 Sequence Collection and Annotations

#### **C19MC**

C19MC sequences (or so called genomic context) including microRNA genes and flanking regions were fetched from UCSC with version hg19 of human genome and version rheMac2 of rhesus (monkey) genome. (Rhead, Karolchik et al. 2010) Sequence lengths of human and rhesus C19MCs were 102301-nt between position 54168188 and 54270488 in human chromosome 19, 105001-nt between position 59771000 and 59876000 in rhesus chromosome 19 respectively. All selected range of C19MC sequences were confirmed by our Rainbow Dotplot program roughly to make sure the full-length of C19MC had been selected.

#### **Alu Elements**

The annotation of Alu elements within C19MC were followed by the Repeat Masker. (Smit 1996-2010) Only the SINE/Alu element category of records was kept in our study and the other categories such as simple repeat, FAM and etc. were removed. The possible ancestor of Alu elements, the FAM families, was eliminated because they are too distant to affirm what their function and activation is in primate lineage. In the section of determining the ability of seed candidates targeting Alu elements, we included all Alu elements scanned and recorded by Repeat Masker. The primary role we analyzed in this section is the seed candidates so that we did not filter Alu element, since all of them could have possibility of targeted by C19MC either currently or in the evolutionary history. Although some of them are partial fragments as relics, they still might be targeted in long time ago.

However, in the section involving the Alu elements insertion within genes, we filtered those Alu elements with length shorter than 200-nt. This criterion implies that only the Alu elements with at least one monomer were kept, because the normal full length of Alu elements is about 300-nt. Although those shorter Alu elements were under the selection force as same as the others, they are incomplete and probably inactive currently. This incompleteness implies that they might be not involved in the gene regulation and not reserved by natural selection. Annotated standard sequences of Alu subfamilies and the standard consensus ancient sequences of Alu and other TEs were fetched from Repbase, GIRI. (Jurka, Kapitonov et al. 2005)

#### **Exons**

The consensus exon sequence of C19MC in primates is defined by previous study (Bortolin-Cavaille, Dance et al. 2009), which indicated a ~123-nt long, spliced exon with strong sequence similarity among primates. All exons defined in our study were searched and found using a local alignment against the consensus exon sequence by our own program and set similarity larger than 80%, i.e. a fragment with more than 99-nt are identical to the 123-nt consensus exon was defined as another exon.

#### 2.2.2 Define C19MC Homologs between Human and Rhesus

#### **Rainbow Dotplot**

For determining the homology of C19MC among primates, we performed an analysis with our homemade visualization program called "Rainbow Plot". The new idea of multiple-color indicators was integrated to this program for intuitively observing the sequence similarity. This dotplot not only represents the relationship among paralogs or orthologs on a diagonal straight line but also the similarity on different color of each line.

As the normal dotplot program, sequences of C19MC from human and rhesus were put on two axes and scanned by 20-nt sliding window. We also tested different sliding window sizes to find the optimal criterion. The resolution of dotplot will decrease if size is larger than 20-nt. Meanwhile, because the multi-color gradient depends on similarity scores, the number of colors will too few to present if sliding window size is smaller than 20-nt. The sliding windows moved along with two axes by one nucleotide shift at each time, the score was counted as one while match pair and zero as mismatch pair. Thus, the possible scores were calculated from zero to twenty in each movement. After counting of full length square (~100K x ~100K), a reduction scoring process was done for reduce picture size. The maximum score of each 10 x 10 square area was selected as the final drawing score for representing the similarity of this square area. **(Figure 8)** 

#### **Homologs Definition**

We used Rainbow Dotplot to rebuild the C19MC annotation in rhesus genome. The fragments of rhesus microRNA homolog genes were selected from dotplot by visualization and then processed pairwise sequence alignment with corresponding human microRNA for confirmation. We also included mature microRNA sustained by Deep Sequencing data as a reference alignment sequence (see section 3.2.1 for detail) and combined MFold program. The MFold could predict RNA structure to double check the possible mature region and to confirm the nucleotides where the rhesus microRNA gene start and end. **(Figure 9)** 

# 2.2.3 Nucleotides Substitution Analysis of C19MC Multiple Sequence Alignment and Aligned Sequences Trimming

Interested sequences region between each exon and microRNA were fetched from C19MC genomic context of human and rhesus. Each selected sequences were started from the first nucleotide of our defined exon and ended at the last nucleotide of microRNA gene annotated by miRBase in human and defined by our homologs definition process in rhesus (see section 2.2.2). Due to some sequences are lacks of internal transposable elements, all internal TEs were completely excised from the selected sequences before alignment. However, the internal flanking sequences except transposons were kept to gain information and improve analysis. Multiple sequence alignment was done by MEGA4 (Tamura, Dudley et al. 2007) and followed by manual check to improve gap reliability. One human sequence with downstream LTR13 flanking the exon rather than microRNA was excised, 4 rhesus sequences without downstream microRNA were also removed. Unreliable aligned regions of each sequence were also eliminated. Finally, a 456-nt length (with gap) alignment of 49 interested paired ortholog sequences were chosen for further analysis. We kept sequences appear only in one species and set a blank sequence to its pair because these sequences still contain information within paralogs even though they do not have paired one.

#### **Site Substitution Calculation**

After multiple sequence alignment and trimming, each nucleotide site of ortholog pairs between two species was calculated by a proportion. The proportion of substitution was formulated by the number of pairs with substitution divided by the total number of pairs. If one sequence of a pair contains a gap in the site, it was deducted from the total number of pairs. For example, we have 49 ortholog pairs, synonymous with 98 sequences, for calculating. A site was found where are 30 pairs be identical, 7 pairs with substitution and 12 pairs have a gap or a blank sequence on one sequence, it was scored as 7 substitution divided by 37. Every sequence within a species, i.e. paralog, was calculated by determining whether the site differ to the nucleotide of majority between species. **(Figure 10)** 

## **MicroRNA Similarity Distance within C19MC**

In order to observe the similarity among microRNAs, we calculated the similarity distance of microRNA with its 6 neighbor micrRNAs, three from upstream and three from downstream. Used fragments of each microRNA were those described above in this section, exon and internal flanking region were included for distance calculation to improve reliability. Pairwise distance calculation was done by MEGA4 with Kimura 2-parameter model and complete deletion, and then plotted the scores along with the length of C19MC. Six dots on a certain x-axis position represent the distance between the local microRNAs and its six adjacent microRNAs; smaller the radius, more distant the neighbor. The least three similar microRNAs, mir-512-1, mir-512-2 and mir-498 of two species were not included in this analysis, because they are too different to compare with the others. **(Figure 11)** 

## 2.3 Results and Discussion

#### 2.3.1 MicroRNAs within C19MC Appear a Characteristic Duplication Unit

The annotation of rhesus C19MC in miRBase is insufficient, so that we re-annotated this region in our study. Only the transposon annotations were followed by Repeat Masker without any change. Both microRNA genes and exons in either human or rhesus genomes were re-defined and manual confirmed in this study. By our visualized annotation **(Appendix 1)**, we found a special unit with a regular patter including exon, Alu elements and microRNAs. Although we did not find evidence and mechanism of gene duplication, we believe this unusual pattern is must as a duplication unit which relates to the expansion of microRNA within C19MC. Moreover, because the transposition of Alu needs proteins offered by L1 element, we speculated the L1MB7 within C19MC also might have a key role in the start of duplication.

## 2.3.2 Homology Definition and Genetic Composition of C19MC

Dotplot is used for determining homologs extensively; we also performed this approach to find the homologs between human and rhesus and further to detect the evolutionary events such as insertion, deletion or recombination. Unlike normal dotplot program defined only one color for presentation, our new program shows straightforward information intuitively, friendly and visually by adding rainbow colors. Each color represents the degree of conservation within a sliding window size: red is the most conserved and followed by orange, yellow, green, and dark blue or purple is the least conserved, black represents no similarity found. **(Figure 8c)** 

#### L1MB7 are Located nearby Two Rearrangement Hot Spots

We found two regions with multiple duplication events, usually called a hotspot, where display as many similar symmetrical diagonal lines within a narrow square region on the dotplot picture. **(Figure 12)** It is noteworthy that L1MB7, a kind of LINE 1 truncated form, are located at the proximal region of the hotspot. We speculated that the locations of L1MB7 correspond to the facilitation of C19MC duplication since L1 offers the primary mechanism of gene duplication, of cause including Alu elements, in most mammalian genomes. (Esnault, Maestre et al. 2000; Dewannieux, Esnault et al. 2003; Han and Boeke 2005) The special L1MB7 pattern was found at 3 regions on rhesus C19MC at 7K, 75K and 103K (related positions start within fetched C19MC sequence); meanwhile, we identified their 4 homologs in human C19MC. Two of them (at 12K and 17.5K) are probably caused by duplication because the Rainbow Dotplot shows a beautiful red line as high similarity. **(Figure 13)** 

#### **MicroRNA Homologs Definition and Annotation**

Because the annotation of C19MC on rhesus genome is insufficient, we need to reconstruct the annotation of rhesus C19MC for further analysis. Forty-seven human C19MC microRNAs were confirmed including one which is not annotated by miRBase (we denoted it by a prefix "hsa-pd-"). On the rhesus genome, 46 orgholog pairs were defined in our study, most of which are not annotated by miRBase, even though they did, many of their annotations are probably wrong. **(Table 1)** For example, some rhesus microRNA position are annotated at the homolog site to human, but their name is inconsistent to human, such as hsa-mir-520b is connected to mml-mir-519a. We kept all the names as is annotated in miRBase but re-annotated them to the right genomic position. The rhesus microRNAs not exist in miRBase but predicted and used in our study is named with prefix "mml-pd-" (means prediction). An undoubted

896

duplication event of human miR-512 from rhesus one was observed. A copy lost of miR-526a on human related to rhesus which has two copies was also indicated.

#### 2.3.3 Substitution Analysis

We paid more attention to the consistent unit sequence which positions from exon to microRNA in C19MC for studying which nucleotide is important and significant to a certain extent of evolution. In the meantime, we expected to observe the fragment detergence of the sequence among different species. Sequence comparison using nucleotide substitution analysis was done by comparing within paralogs within one organism and comparing orgholog pairs between two species. Both line chart for macroscopic tendency of substitution (global view in full length of C19MC) and scatter plot for microscopic variation (local view in a certain region within C19MC) were used for visual presentation in our analysis.

The overall substitution of rhesus is more frequent than human among three different sequence regions. According to the scatter plots, the pattern of three regions in paralogs substitution related to ortholog pairs is the same inside species between human and rhesus: exon is the most divergent, followed by internal flanking region and the microRNA is more conserved. However, when this pattern was compared between two species instead inside species, we found rhesus has less conservation than human. These facts exclude the possible sampling bias caused by pseudogene; meanwhile, the result suggests that stronger selection force occurred in rhesus and caused more divergent in its paralogs. **(Figure 14)** 

#### **The Consistent Fluctuation of Global Tendency**

The fluctuated peak along the line chart shows the difference among three proportions, human almost has lower substitution proportion related to the ortholog pairs' one than rhesus in each site. The overall tendency indicates that human reached more adaptation and reduce its variety. The exon region expresses the unstable fluctuated related to the internal flanking and microRNA region. The internal flanking region, especially the part near to microRNA, shows the least fluctuant difference within paralogs in both human and rhesus. However, few nucleotids of microRNA have high proportion of substitution related to average low proportion of other nucleotides; this phenomenon is in line with the known microRNA conservation rules correspond to its structures.

In addition, the substitution pattern, including the difference and trend, of internal flanking region closed to exon is similar to exon region. It seems as a transitional progress of substitution gradient from varied exon region to the conserved flanking of microRNA but rose again in microRNA sequence. **(Figure** 1896

### Substitution Variety in a Global View

These three proportions of each site were next drawn on an X-Y scatter plot per 20-nt length to investigate the local trend of substitution. The substitution proportion of ortholog pairs was plotted on X-axis as a reference value corresponds to the two paralog proportion on Y-axis. Therefore, this plot could reflect the different degree of substitution in a site within paralogs in one species to which between ortholog pairs in two species.

Most spots centralized at bottom-left area and presented the similar intercept of Y display the site consistency between human and rhesus. However, few sites are biased and located at top-right area suggest the higher substitution occurred both within paralogs and between orthologs. The sites we had interested are those located at top-right and existed higher intercept distance between two paralogs, since these sites have more variation between species. (Figure 16)

The results were separated according to different components again. It is consistent with macroscopic analysis that most interest sites belong to exon region and only few are located in internal flanking. Even though many top-right spots observed in microRNA region, but they are much closed between two species instead of the phenomena found in exon. It implied even the sites within microRNA have high divergence between species, it is conserved within paralogs. In other words, the microRNAs in C19MC may not only evolve specifically and separately after speciation but also be conserved to maintain some important gene regulation within a species.

1<u>896</u>

# Chapter 3 C19MC probably as a defender against Alu Elements

## 3.1 Backgrounds and Hypothesis

#### The duplication of microRNA is related to repeat elements

Mammalian microRNAs derived from two inverted repeat elements via transposition were reported. (Smalheiser and Torvik 2005)

#### The Impact of Gene Expression by Alu Elements

It is found in recent research that Alu elements contain transcription factor binding sites might modulate gene expression (Shankar, Grover et al. 2004; Polak and Domany 2006) even function as promoters in C19MC. (Borchert, Lanier et al. 2006) Moreover, the major rules of Alu or other transposable element involved in the evolution of eukaryotic gene regulatory network were reviewed. (Feschotte 2008) However, this convenient and efficient mechanism of regulatory network evolution could be lethal due to unexpected, uncontrolled and rapid expansion of Alu. We believed a counteracting mechanism must be present to control the expansion of transposable element.

#### MicroRNA as a Guard of Genome against to Transposons

In the two cases we described in 2.1, some novel microRNAs governed by upstream Alu elements and transcribed by RNA polymerase III were indentified. (Gu, Yi et al. 2009) It suggests Alu elements involved in small RNA regulation. However, more and more studies suggested that transcription suppression by small RNAs is probably the means of killing transposons. (Malone and Hannon 2009)

#### The Co-Evolution Model of C19MC Expansion

We also followed this hypothesis that one role of C19MC might be the guard of genome and maintain its completeness. The selfish transposons are devoted in their expansion; at the same time, microRNAs strive to repress their expansion. The microRNA duplication could be caused while transposons expand themselves; TEs facilitate the growth of their enemy while extending unintentionally. More microRNAs were duplicated as the defenders, more transposons were occurred as the occupiers, meanwhile induced more microRNA duplicated again. We expected to discover some evidences supporting this infinitive loop. (Figure 17) If this circle existed, maybe some Alu elements escaped from microRNA repression could be existed and discovered. The youngest AluY was found as this situation that the percentage of targeting AluY is reduced than AluS. It implies that the AluY probably find out

the salvation.

## 3.2 Materials and Methods

# 3.2.1 Determine Mature microRNA Expression Profile from Next-Generation Sequencing

896

It's difficult to determine the expression of C19MC since many studies found the C19MC is only expressed in few cell lines including placenta cell, germ line cell and embryonic stem cell. Fortunately, we found a study might imply the data we want because they used the next-generation sequencing (also called deep sequence) to discover novel microRNAs in human embryonic stem cell. The microRNA expression profiles of human embryonic stem cell were fetched from the Solexa sequencing data published by Morin (Morin, O'Connor et al. 2008), who collected the appropriate length of mRNA by electrophoresis and then processed for sequencing. We collected all the sequences with longer than 10-nt length and counting their number of reads as the expression level. Because the next-generation sequencing technique is powerful enough to detect mRNA even it has only one sequence, we could use this data to overcome the erratic mature sequence of microRNAs, which may occur nucleotide shift and not be identical in each time of expression. The residues of microRNA generation such as the loop cut by Dicer could be found in our alignment indicate the high-resolution of this sequencing technique. **(Figure 18)** 

#### **Human Seed Candidates Selection**

After redundant reads were removed, we aligned each selected reads, the possible mature microRNA, with annotated C19MC genes. We next confirmed the reads aligned with high possibility rather than random match. The sequencing reads which match to the mature region of microRNA (the arm structure of a microRNA) were kept, but the sequencing reads match to the loop region or other flanking region of microRNA rather than mature form were eliminated. In addition, the low number of sequenced reads still remained in our selection because we are not sure whether it functions or not. (Figure 19) Generally, it is believed that microRNA seeds are located on 2-8 nucleotides counted from the 5' end of mature sequence. Although it is not the golden rule, the seed are still the most important sites for determining target site interaction. (Bartel and Chen 2004; Kertesz, Iovino et al. 2007; Bartel 2009) Moreover, many studies involved in microRNA target sites prediction suggested the match of seed region is the major parameter while scanning. (Brennecke, Stark et al. 2005; Krek, Grun et al. 2005; Nakamoto, Jin et al. 2005; Xie, Lu et al. 2005) We also followed this consensus pattern by selecting seed from 2-8 nucleotides of 5' end mature sequence as our query candidates against genomes. Finally, 64 seed candidates were selected and subjected to further analysis.

#### **Rhesus Seed Candidates Selection**

Unfortunately, there was no C19MC expression data of rhesus reported so far so that we cannot use the same approach to determine real mature sequence of rhesus C19MC. To solve this problem, we combined the defined homologs described in 2.3.2 and the human mature sequences from real expression data to select rhesus seed candidates. Because the mature regions of each microRNA are not conserved among human C19MC and the mature sequence could be varied even within a microRNA, we relaxed our standards while choosing rhesus seeds. The standard mature sequences of rhesus were followed as the position aligned to human and considered 3-nt shift of mature region. **(Figure 20)** Due to the relation of standards, 231 seed candidates were selected more than the human.

## 3.2.2 Whole Genome Scanning (From Seed Candidates' Point of View)

The genome position perfectly matched with seed candidates were recorded after genomic scanning. Next, these positions were cross-matched with the collated annotation of Alu elements described in the section 2.2.1. (Figure 21)

#### Ability of Hitting Alu Elements of each Seed

Because the number of Alu targeted by microRNAs could be varied depend their specific seeds, the different amount of hit number may also reflect the repression ability of the seeds. To determine whether the numbers of Alu targeted by each seed of C19MC have a bias against normal distribution of random event, we generated a random dataset of seven-nucleotide-length seed candidates to simulate the 7-nt length seed candidates we selected from the real datasets of human C19MC and rhesus C19MC. This generated random dataset contains 5000 seeds which are not included those seeds selected from real C19MC in two species. All Alu elements annotated by Repeat Masker were used as target pool against by three seed datasets, seeds of human C19MC, seeds of rhesus C19MC and 5000 random generated seeds. Whole genome scan was done by our own program.

#### **Interchange of Seed Candidates**

To determine whether the seeds have specialization between two species and as a reference dataset, we next interchanged the seed candidates of one species to against the genome of another species. In other words, we used human seeds as queries to scan rhesus genome and vice versa.

# 3.2.3 Whole Genome Scanning (From Alu elements' Point of View) Targeted Proportion among Different Alu Subfamilies

Alu elements are raised in different evolutionary stages from the ancient AluJ to the most abundant AluS then the youngest AluY. Presumably, the AluJ is inactive now due to its long evolution age, but recent studies suggest the AluS could still maintain its activity. (Bennett, Keller et al. 2008) According to our original hypothesis, the C19MC we observe now probably have ability to repress AluS and AluY expansion which might damage the genomes, but lost function to inhibit the inactive AluJ which has no harm to the genome. Meanwhile, because AluS appear early and more amount than AluY, C19MC could evolve to against AluS primarily. Even if we cannot observe the expectation, the proportion of Alu targeted by C19MC might be different among 3 Alu subfamilies. Here, all Alu elements were considered in the analysis because all of them probably had been targeted in the evolutionary history no matter which positions or strands they are located recently.

#### **Classification of Alu Elements by Inserted Position**

Alu elements annotated by Repeat Masker were scaned and separated into 4 categories including insertion of 5'UTR, ORF, 3'UTR and non-coding genes, by the reference gene annotation fetched from UCSC. We assumed this gene regulation involving by Alu element might depend on microRNA mechanism. Therefore, only the Alu elements inserted into genes in the same direction of mRNA transcription were selected. Only the Alu elements transcribed with genes into mRNA have opportunity to be targeted by microRNAs.

### 3.3 **Results and Discussion**

## 3.3.1 C19MC has Extremely Different Expression Levels

The expression or microRNAs in C19MC ranges from extremely low level with less than 10 reads to high level with hundreds reads. We expected to find out the relationship between expression level and Alu elements, but achieved nothing. It may because the regulatory mechanism of C19MC could via epigenetics instead of mRNA level. The expression profiles of each mature microRNAs were summarized on the Table **(Table 2)**.

#### 3.3.2 The Biased Distribution of C19MC Target Ability against Alu Elements

Basically, although the nucleotides variation of a local genomic region could be different, the stochastically generated seeds with short length in the extreme should hit genomes in random and show a normal distribution. We used the 5000 random seeds to against genomes as a control to determine whether seeds of C19MC have different targeting ability to Alu elements. (Figure 22, 23) The number of Alu targeted by each seed were plotted on a logarithmic scale to reduce their extreme variation and shown on X-axis; the Y-axis represents how many seeds against Alu elements in the same amount. As we expected, the trend of random seed dataset displayed a smooth normal distribution which indicates the seeds of microRNA, if they generated randomly and not encountered selection force, might target genomes fortuitously. Both human and rhesus seeds have biased distribution located on the high Alu element hits region (right side of x-axis). We found that seeds belong to this distribution are complementary to the potential microRNA targeting sequence of Alu elements reported by previous study. This coincident complement between our biased seeds with the core target site of Alu reported by them indicates that the defense mechanism probably exists and selection force might occur.

## **Biased Seed Candidates have Conserved Sequences**

These biased seeds of human and rhesus are the same mostly. It suggests that the defense mechanism against Alu elements maybe conserved in primate lineage. Based on our hypothesis, the C19MC are primary to defense the AluS **1896** expansion, thus they also duplicated themselves to achieve the enough ability of repression. If this hypothesis is real, we could find the conservation of seeds between human and rhesus due to their functional preservation.

#### **Specialization of Human C19MC Seeds**

While we interchanged C19 seed candidates of each species to against with the genome of another species, we found that some seed candidates present a biased pattern. **(Figure 24)** Almost seeds of two resources were lined up with a diagonal as equal Alu hits between two species. However, the human seeds seem be specialized against to human genome because they deviated from the diagonal line. This pattern indicates that the specialization occurred and selection force existed in human C19MC seed during primate evolution, thus the human C19MC provides specific seeds against most human Alu elements but
useless while against rhesus Ale elements. Also, it indicates the seeds might be functional in the past although they are not highly expressed now. Only they functioned during the primate evolution, the selection force could be existent. Interestingly, both human and rhesus seeds targeting AluY are deviated from the patterns of targeting AluJ or AluS and tend toward targeting more rhesus Alu elements than human ones. Rhesus also has two specific biased seeds not found in human and they deviated only in targeting AluY.

#### 3.3.3 Analysis of Alu Elements Targeted by C19MC

#### Preference of Alu Targeted by C19MC: Different Alu Subfamilies

The slight different microRNA abilities of targeting Alu against three Alu subfamilies could be observed if the variation occurred in conserved target regions of AluY. In addition, because AluY was born after AluS, if C19MC originally recognizes the conserved sequence of AluS elements, the biased targeting number of AluY is still detectable. We demonstrate extended analysis that separated the number of targeted Alu elements into three different subfamilies in order to detect the biased target preference of C19MC among Alu subfamilies. (Figure 25) As our expectation, the AluJ shows the least hit by human C19MC. This result supported our two reasons: one is that many AluJ are truncated in sequences and accumulated mutations during evolution so that cannot be targeted by C19MC recently; another is that C19MC might appear while AluS expansion and originally evolve to against AluS. Although the results seems not be significant in rhesus, the trend among three subfamilies is similar. This is probably because of the overestimation in rhesus seeds selection which mixed many pseudo-seeds and caused false positive.

Interestingly, AluY shows slightly reduction of targeting as we expected even we did not know the mechanism of its escapes. The few amount of AluY escaped from C19MC targeting could be induced by the target site mutation as same as AluJ. These mutated AluY might have more opportunities to survive and to duplicate under the C19MC defense force so that we observed this slightly decrease of targeting. The related low difference between AluS and AluY may because the evolution time is not enough for AluY to accumulate mutation and to encounter selection.

#### Preference of Alu Targeted by C19MC: Different Alu Inserted Position

Many genes are inserted by transposons during evolution process, especially Alu elements. Depends on the inserted position, we classified Alu elements into 4 categories, harbored on 5'UTR, ORF (open reading frame) and 3'UTR of genes and those inserted to the non-coding genes. In our opinion, if the inserted Alu elements become the component of gene regulation by microRNA, the Alu harbored on 3'UTR of genes are more important and beneficial. While the inserted Alu provides the biological benefits and involved in regulatory network, the repression and inhibition of them are not necessary anymore. We examined this concept by analyzing the targeting proportion of Alu inserted into different position of genes in human and rhesus.

## Figure 1 The Distribution and Structure of Transposons

(a) Percentage of transposable elements in human genome. Alu elements approximately occupy the one-third of genome content. Adapted from (Cordaux and Batzer 2009).
(b) Genomic structure of L1 and Alu element. The main difference between them is whether protein encoded or not. Promoters are marked as dark blue.



## (a) Transposable Element of Human Gneome

(b) Genomic Content of LINE1 and Alu

#### LINE1

Appear in **~150 Myr** ago



### ΑΙυ

Appear in **~65 Myr** ago

→	Left Monomer	A-rich	Right Monomer	polyA →
RNA	Pol III Promoter		~300 bp	

#### Figure 2 Distributions of Alu Subfamilies in Primate Genomes

The current distributions of Alu subfamilies in primate genomes are shown. Genome versions are fetched from UCSC by hg19 for human, ponAbe2 for orangutan, rheMac2 for rhesus and panTro2 for chimpanzee. The other category represents the non-Alu SINEs.



#### Figure 3 Evolution of Alu Subfamilies

Evolutionary tree of 213 Alu subfamilies throughout the primate evolution were identified in previous study. Adapted from (Price, Eskin et al. 2004).



# Figure 4 Biogenesis and Structure of microRNA

(a) Biogenesis of microRNA. Adapt from (Abeloff 2008).



(b) Structure of microRNA stem-loop. Mature microRNAs could be produced from both arms of pre-microRNA. If the amount of mature sequences formed from two arms are relatively equal, they are denoted by -5p and -3p to indicate their origins. However, if one of them is expressed as extremely low level, it is denoted by \*.



Nomenclature of mature microRNAs Identified mature products from two arms are equal: → mir-xxx-5p and mir-xxx-3p Identified mature product from one arm is much lower: → mir-xxx (higher one) and mir-xxx\* (lower one) Only one type of mature product is identified: → mir-xxx (both 5' or 3' is possible)

### Figure 5 The microRNA Clusters of Human

The C19MC composition was shown in a real scale of length  $\sim$ 100Kb. Transposable elements located on the same direction with microRNA were marked as green, on the opposite direction were marked as blue. MicroRNAs and exons were represented as red and gray. Chromosome positions and two nearest flanking genes were also indicated.



Green Bar: TEs, the same direction with microRNA genes; Blue Bar: TEs, the opposite direction with microRNA genes; Red Bar: microRNA genes. All elements within clusters were drawn in actual scale. Positions of flanking genes were shown as relative locations, not in scale.

#### Figure 6 Remarkable Duplication Unit of C19MC

(a) Flanking genomic components of microRNA. Some cases are shown here as example in real scale and position.(b) General composition within a unit. Most microRNAs of C19MC could be found as this rearrangement with one exon, one or two L1M relics and few flanking Alu elements.



#### Figure 7 Two Different Models of C19MC Transcription

Two extremely different mechanisms were shown. Both models were confirmed by biological experiments. a) C19MC microRNAs are transcribed by RNA Pol II and processed by alternative splicing. b) C19MC microRNAs are co-transcribed by RNA Pol III using promoter of Alu elements located on upstream.



Ribosom

mRNA located on the same direction. The effect of opposite Alu to microRNA is unkwon.

#### Figure 8 Procedure of Rainbow Plot

(a) Diagram of the scanning process of "Rainbow Plot". (b) Resize process of scanned pictures. (c) An output example of Rainbow Dotplot showed the visualization of indel and similar regions in a simple way. Colors represent the similarity of nucleotides a sliding window size, high conservation as red, low conservation as dark blue or purple and no similarity as black.



#### Figure 9 Determine C19MC Homologs in Rhesus

The process of determining C19MC Homologs in rhesus was shown. First, the position and similarity of homolog pairs were found by the Rainbow Dotplot. Second, fetched rhesus fragments were aligned with annotated human C19MC genes and the mature sequences confirmed by deep sequencing data. Finally, the structures of homologs were predicted by MFold as a further verification.



(2) Aligned and confirmed with expression profile



#### Figure 10 Process of Substitution Calculation

Calculation of Substitutions between Orthologs



- (1) Determine whether substitution occurred in each pair.
- (2) Calculate the score by the function :

n –	Number of Pair with Substitution
$P_{\text{substitution}}$ –	Number of Ortholog Pairs

(3) One sequence in a pair with a gap is eliminated from total number of pairs.

Calculation of Substitutions within Paralogs



- (1) Find the nucleotide of majority in all sequences.
- (2) Calculate the score by the function :

### Number of Sequence with Substitution

# $p_{\text{substitution}} = \frac{1}{\text{Number of Sequence within Paralogs}}$

(3) If the site contains a gap, it is deducted from the total number of sequence within paralogs.

#### Figure 11 Pairwise Similarity Distances of C19MC

Similarity distances between microRNAs and its 6 microRNA neighbors from two sides were plotted along with length of C19MC. The flanking region of microRNA used in substitution analysis was included in calculation to enhance scoring reliability. Each dot with the larger radius means the distance calculated with the nearer neighbor. A similar trend closed to first L1MB7 was found both in human and rhesus (arrow indicated). Mir-512-1, mir-512-2 and mir-498 were not included because they are too distant to calculate the reliable scores. The first and last microRNA have only three microRNA neighbors located on one side.

<u>Human</u>



**Rhesus** 



# Figure 12 Duplication Hotspots

An example of duplication hotspot was shown. It is worth to mention that the L1MB7 are located nearby the hotspots.



## Figure 13 The Special L1MB7 Sequences on C19MC

a) The positions of 4 L1MB7 defined by human-human Rainbow Dotplot were identified.

b) The relative positions of L1MB7 in human and rhesus.



## Figure 14Overall Scatter Plot of P<sub>Substitution</sub> in Two Species

The proportion of ortholog pairs are plotted as X-axis against to Y-axis with paralog proportion, dot were classified to 3 categories by genomic components. Overall plots of another species were drawn on the same figure as a control.



#### Figure 15 Proportion of Substitution in Macroscopic View

A macroscopic view of  $P_{\text{substitution}}$  was presented by a line chart along the full length of analyzed sequence with red line as rhesus, blue line as human and black line as orgholog pairs. Three different genomic regions, exon, internal flanking and microRNA are labeled. The difference of paralogs substitution between two species was highlighted by a gradient color.





The  $P_{\text{substitution}}$  are drawn on the line chart along the sequence. Scatter plots show the relationship between two paralogs and orghotlog pairs within a local region of sequence. The alignment of the site with difference we interested is shown.



43

#### **Multiple Sequence Alignment**

	1	5	4	3	
	10 20	10 20	40 56	40 50	
	· · · · l · · · · <mark>l</mark> · · · · l ·	· · · · [ · · · · [ · · · · ] · ·	· [ · · · · ] · · · • ] · · · · ]	· L · · · · L · · · · · · · · · · · · ·	
# hsa-mir-512-1					* mml-mir-512-1
# hsa-mir-512-2	<b>-</b>	<mark>-</mark>			* mml-mir-512-1
# hsa-mir-1323	<b>-</b>				mml-pd-1323
hsa-mir-498					mml-mir-498
hsa-mir-520e	CATTTTGGATCCTCCC		GGTGTACTGGAAC	<mark>-</mark>	Blank
hsa-mir-515-1	CATTTTGGATGTTCCC	CATTTTGAATGCTCCC	TGTTTCTGTGCTGGAAC	TGTCCCTGTGCTGGAAC	mml-pd-Up1
hsa-mir-519e	CATTTTGGTTGCTCCC	CATTTTGGATGCTCCC	TGTCCCTGTGCTGGAAC	TGTCCCTGTCCTGGAAC	mml-pd-Up2
hsa-mir-520f	CATTTTGGATGCTCCC	CATTTTGGATGCTCCC	TGTCCCTGTGCTGGAAC	TGTCCCTGTCCTGGAAC	mml-pd-Up3
hsa-mir-515-2	CATTTTGGATGCTTGC	CATTTTGGAAGCTCCC	TGTCCCTGTACTGGGAC	TGTCCCTGGGCTGGAAC	mml-pd-Up4
hsa-mir-519c	CATTTTGGATGCCCCC	CATTTTGGATGCCCCC	TGTCACTGTACTGGAAC	TGTCCCTGTCCTGGAAC	mml-mir-519c
hsa-mir-1283-1	CATTTTGGATGTTTCC	CATTTTGGACGTTTCC	TGTCTCCGCACTGGAAC	TGTCTCCACACTGCAAC	mml-pd-1283-1
# hsa-mir-520a	CATTTTGGATGCCCCC		TGTGCCTATACTGGAAC		mml-mir-520a
hsa-mir-526b	CATTTTGGATGCTCCC		GGTCC-TGTACTGGAAC		mml-mir-526b
hsa-mir-519b	CATTTTGACTGCTCTA	CATTTTGACTGCTCAG	TGTCCCTGTGCTGGAAA	GGTCCCTGTGCTGGAAC	mml-pd-519b
hsa-mir-525	- CTTTTAAGGCCTCGC	- TTTTTATGGCCTCGC	TGTCCCTGTCCTGGAAC	TGTCCCTGTCCTGGAAC	mml-pd-525
nsa-mir-523	CATTTTGGATGTTCCT	CATTTTGGATGCTCCT	TGTCCCTGTGCTGGAAC	TGTCCCTGTGCTGGAAC	mml-mir-523a
hsa-mir-5181	CATTTTGGATGCTCCC	CATTTTGGAGGCTCCC	TGTCCCTGTGCTGGAAC	TGCTCCTGTCCTGGAAC	mml-mir-518f
hoo-mir-519b	CATTTTGGATGCTTGC	CATTTTGGATGCCTCC	TGTCCCTGTACTGAAAC	TGTCCCTGTCCTGGAAC	mml-mlr-519a
hea-mir-526a-1	CATTTTGGATGCTCCC	CATTTTGGATGCTCCC	TGTCCCTATGTTGGAAC	TGTCCCTAG <sup>G</sup> CTGGAAC	mml nd Mdl
hsa-mir-520a-1	AGTTTTGAA I GCTCCC	CATTTTGGATGCTCGC	TGTCCCTGTGCTGGAAC	TGTCCCTGTGCTGGAAC	Rinni-pa-Mai
hsa-mir-518c	CATTTCGGATGCCCCC		CGCTCCTGTACTGGAAC		mml_mir_518c
hsa-mir-524	CATTTTGGATGCTCCC	CATTTTGGATGCTCCC	TGTCCCTGTGCTGGAAC	TATCCCTGTGCTGGAAC	mml-mir-524
# hsa-mir-517a					mml-mir-517a
hsa-mir-519d					mml-mir-519d
hsa-mir-521-2		CATTTTGGATGCTCCC	TATCCCTGTGCTGGAAC	TGTCCCTGGGCTGGAAC	Blank
hsa-mir-520d	GATTTTGGGTGCTCCC		TGTACCTGTCCTGGAAC		mml-mir-518a
# hsa-mir-517b		CATTTTACA   GCTCCC		TGTCCCTGTCCTGGAAC	Blank
hsa-mir-520g					mml-mir-520g
# hsa-mir-516b-2					mml-pd-516b-2
* hsa-mir-526a-2	GATGTTGGATGCTCCC		TGTCCCTGTACTGGAAC	TGTCCCTGTCCTGGAAC	mml-pd-Md3
* hsa-mir-526a-2	CATTTTGGATGCTCCC	CATTTTGGATGCTCCC	TGTTCCTATATAGGAAC	TTTCCCTGTCCTGGAAC	mml-pd-Md4
hsa-mir-518e	CATTITIGGATGCTCCC		TGTTCCTATATAGGAAC	TGCTCCTGTCCTGGAAC	mml-mir-523b
hsa-mir-518a-1	CATTTIGGATGCTCCC	CATTTTGGATGCTCCC	TGTCCCTGTGCTGGAAC	TGTCCCTGTCCTGGAAC	mml-pd-Down1
hsa-mir-518d	CATTITICGATCCTCCC		TGTCTCTGT	TGCTCATTTCCTGGAAC	mml-pd-Down2
# hsa-mir-516b-1	TATTTTGGAGGCTCCC	CATTTTGGAGGCTCCC	TGTCCCTGTGCTGGAAC	TGTCCCTGTGCTGGAAC	mml-pd-Down3
hsa-mir-518a-2	CATTITICGATCCCC		TGCCTATACTGGAAC	TGTCCCTGTGCTGGAAC	mml-pd-Down4
hsa-mir-517c	CACTTTEGATECTCCT		TGTCCCTGTGCTGGAAC	GGTCCCTGTCCTGGAAC	mml-pd-Down5
# hsa-mir-520h	CACIFICOATOCICCI	CACTTTEGATECTECC		TGCTCCTGGGCTGGAAC	Blank
Blank	CATTTEGATCCTTCC		TGCCCCTGTCCTGGAAC		mml-pd-Down7
hsa-mir-521-1	CATTITICCATCOTCCC	CATTITICAT	TGTCCCTGTSCTGGAAC	TGTCCCTGGGCTGGAAC	mml-mir-519b
Blank	CARTITICGATOCICCO	CATTTCGGAUACCCCC	TGTCCCTGTGCTGGGAC	TGTCCCTGTCCTGGAAC	mml-mir-521
nsa-mir-522	CAUTTITECATOCICCO	CATTITIGGACGCTCCC	TGTCCCTGTACTGGAAC	TGTCCCTGTGCTGGAAC	mml-mir-518e
nsa-mir-519a-1	CATTTTEGATGCCCTC	CATTTTGGALGCCCCC	- TCCCTGTACTGGAAC	TGTCCCTGTCCTGGAAC	mm1-mir-519a-1
nsd=mir=527	CATTITICATTICCTTCC	CATTITIGGAT	TGTCCTTGTSCTGGAAC	TCCCTGTACTGGAAC	mmi-mir-52/
hsa-mir-1283 2	CATTT CONTECTOO	CATTTIGGT	TGCCCCTGTACTGGAAC	TGTCCCTGTGCTGGAAC	mml-mlr-516a-1
hea-mir-516a-2	CATTTTCCCTCCC	CATTTTGAATGCTCCC	TGTCCCTGTACTGGAAC	TGTCCCTGTCCTGGAAC	nunt-pa-1283-2 Plank
hsa-mir-519a-2	CATTTTEGGTGCTCC		TGTCCCTGTACTGGAAC		Didlik mml-pd-5195-2
nsa miri Sisa-z	CATTITIOOUTOCCTCC		TOTOCCTOTACTOGAAC		nuur-pu-sraa-z

\* Duplication occurred in one of two species

# Highly expressed by count of deep sequencing read larger than 300

70 – 72

The  $P_{\text{substitution}}$  are drawn on the line chart along the sequence. Scatter plots show the relationship between two paralogs and orghotlog pairs within a local region of sequence. The alignment of the site with difference we interested is shown.



101, 104

#### **Multiple Sequence Alignment**

	70 86	70 86	100 110	100 110	
		ertered exections t	[]	terrest stars becaute	
# has min 510 1					+ E10 1
# hsa-mir-512-1					* mm1-m1r-512-1
# hsa-mir 1222					* mm1-m1r-512-1
# IISa-IIII-1525					mm1-pa-1323
has min 520s					mmi-mir-498
haa min 515 1	TIGGCTCAAAATTCATTG	TOOPTCALLATOO	TCTAGC-CAGTTCTCTT		Blank
haa min 510a	- IGGCTCAACATCC	- IGGCICAAAAICC	TCCAGCUCAATTCTCTT	IGCAGCCCAATICICIT	mml-pd-Upl
has min 520f	AIGGCICAAAAICCAIIG	TIGGCICAAAATCCATIG	TCCAGCCCAATTCTCTT	ICGAGCCCAG ITCICGT	nm1-pa-op2
has min 515 2	TTGGCTCAAAATCCATTG	TIGGCICAAAATCCATIG	TCCAGCI CAATTCTCAT	ICCAGGUCAATICICIT	nm1-pa-up3
has min 510a	TTGGCTTCAAATCCATTG	TTGGCTCAAAATCCATTG	TCCAGCCCAATTCTCAT	TGGAGCCCAGTTCTCTT	nml-pd-0p4
has min 1292 1	TTAGCTCAAAATCTATTG	TTGGCTCAAAATCCATGG	TCCAGCCCAATTCTCAT	TCCAGCCCAATTCTCAT	mm1-m1r-519C
has min 520s	TTGGCTCAGAATCCATTG	TTGGCTCAGAATCCATTG	TCCAGTCCAATTCTCTT	TCCAGTCTAATTCTCTT	mm1-pa-1283-1
# 115a-1111-520a	TTGGCTCAAAATCCATTG		TTCAGTTCGATTCTCTT		mml min 520a
hsa-mir-519b	TTGGCTTAAAATCCATTG		TCCTGCCCAATTTTCTC		mml nd 510b
hea-mir-525	TTGGCTGAAAATCCACTG	TTGGCTGAAAATCCATTG	TCCAGCCCAATTCTCTT	TCCAGCCCAGTTCTCAT	mml-pd-525
haa-mir-523	TTGGTTTAAAATTGATTG	TTGGCTTAAAATCCATTG	TCCAGCCCAATTCTCTT	TCTAGCCCA <sup>A</sup> TTCTTTT	mml min 5225
haa-mir-519f	TTGGCTCAAAATCCATTG	TTGGTTCAAAATCCATTG	TCCAGCCAAATTCTATT	TCCAGCCAAATTCTATT	mml min E10f
haa-mir-520b	CTGGCTCAATATCCATT-	TTGGCTCTCCATCCATTG	TCCAGCCCAGTTCTCTT	TCCAGCCCGGGTCTCTT	mm1 min 5100
hca_mir_518b	TTGGCTCTCCGTCCATTG	TTGGCTCAAAATCCATTG	TCCAGCCCAGTTCCCTT	TCCAGCCCAATTCTCTT	mml min E19b
hea-mir-526a-1	TTGGCTCAAAATCCATTG	TTGGCTCAAAATCCATTG	TCCTGCCTAATTCTCTT	TCCAGCCCA <sup>A</sup> TTCTCTT	mml nd Mdl
hsa-mir-520a-i	TTGGCTCAAAATCCATTG	TTGGCTCTCCACCCATTG	TCCAGCCCAATTCTCTT	TCCAGCCCGGTTCTCTT	Blank
hsa mir 5200	TTGGCTCAAAATCCATTG		TCCAGCCCAATTCTCTT		mml_mir_519g
hsa-mir-524	-TGTGACTCAATCCATTG	- TGACTCAAAATCCATTG	TCCACCTCAATTCTCTT	TCCAGCTCAATTCTCTG	mml_mir_524
# hsa-mir-517a					mml_mir_5175
hsa-mir-519d					mml_mir_510d
hsa-mir-521-2	TCAGCTCAAAGTCCATTG	TTGACTCAAACTCCATTG	TCCAGCCCASTTCTCTT	TTCAGTGCANTTCTCTT	Blank
hsa-mir-520d	TICCCTCATAATCCATTC	THOACTO BE CECATIO	TOTTACC ADTTOTOCT	TCACIOCATITCICIT	mml_mir_518a
# hsa-mir-517b	HIGGETCATAATCCATTG	TTECCTCAAAATCCATTE	ACTOCT	TCCTCCCTANTTCTCTT	Blank
hsa-mir-520g		TIGOCICONATCCATIO	ACTOCI	TCCTGCCTASTICTCTT	mml-mir-520g
# hsa-mir-516b-2		TTOPOTOTOT CARACTE		TO ACCOMPTOTOT	mml-nd-516b-2
* hsa-mir-526a-2	TTOCOTOTALCTTCATTC	TIGCGTCTCT-GACAATT	TC-AGCGCAATTCTCTT		mml-nd-Md3
* hsa-mir-526a-2	TIGGCTCTAAGTTCATTG	TIGGCICICCATCCITIG	TCCAGCCCAATTCTCTT	TCCAGCTCAATTCTCTT	mml-nd-Md4
hsa-mir-518e	TIGGCTTAAAATCCATTG	TIGGCICTCCATCCATIG	TTAGCCCAGTTATCTT	ICCAGCICAATICICIT	mml-mir-523b
hsa-mir-518a-1	TIGGCTTAAAATCCATTG	TIGGCICICCATCCATIG		ICCAGCICAAIICICII	mml-nd-Down1
hsa-mir-518d	IIGGCICAAAAICCAIIG	TTGGCTCTCCATCCATTG	ICCAGCCCAAIICICII	TCCAGCICAATTCTCTT	mml-pd-Down2
# hsa-mir-516b-1	CTGGCTCICCATCCATTG	TTGGCTCICCATCCATTG	TCCAGCCCAATTCTCTT	TCCAGCCCACTTCTCT-	mml-pd-Down3
hsa-mir-518a-2	TTGGCTGTACATCCATTG	TTGGCACTCCATCCATTG	TTCAGCCCAATTCTCTT	TCCAGCCCGG TTCTCTT	mml-pd-Down4
hsa-mir-517c	TTGGCTCAAAATCCATTG	TTGGCTCTCCATCCATTG	TCCAGCCCAATTCTCTT	TCTAGCCTGGTTCTCTT	mml-pd-Down5
# hsa-mir-520h	TTGGCTGAAAATCCATTG	TTGGCTC <sup>TCC</sup> ATCCATTG	TCCAGCCCAAGTCTCTT	TCCAGCTCAATTCTCTT	Blank
Blank		TTGGCTGAAAATCCATTG		TCCTGCCCANTTCTCTT	mml-pd-Down7
hsa-mir-521-1	TTGGCTCAAAATCCATTG		TTCTGCCCTAATCTCTT		mml-mir-519b
Blank	TTGGCTCAAAATCCATTG	TTGGCTCAAAATCCATTG	TCCTGCCCTATTCTCTT	TCCTGCCCA <sup>1</sup> TTCTCTT	mml-mir-521
hsa-mir-522	TGGGCTCTCCATCCATTG	TTGGCTCAAAATCCATTG	TCCAGC <sup>C</sup> CA <sup>A</sup> TTCTCTT	TCCAGCTCAGTTCTCTT	mml-mir-518e
hsa-mir-519a-1	TTGGCTCAAAATCCATTG	TTGGCTCTCCGTCCATTG	TCCAGCCTAATTCTCTT	TCCTGCCCGATTCTCTT	mml-mir-519a-1
hsa-mir-527	TTGGTTCAAAATCCGTCG	TTGGCTCAAAATCCGTGG	TCCAGCTCAGTTCTC	TTCAGC <sup>C</sup> CA <sup>A</sup> TTCTCGT	mml-mir-527
hsa-mir-516a-1	CTGGCTCAATATCCATTG	TTGGCTCAAAATCCATTG	TCCAGCCCAGTTCTCTT	TCCAGCCCAGTTCTCTT	mml-mir-516a-1
hsa-mir-1283-2	TTGGCTCAACATCCATTG	CTGGCTCAGT ATCCATCG	TCCAGGCCAATTGTCTT	TCCAGCCCA <sup>A</sup> TTCTCTT	mml-pd-1283-2
hsa-mir-516a-2	TTGGCTCATAATCCATTG	TTGGCTCAATGTCCATT-	TCCAGCCCAATTCTCTT	TCCAGGCCA <sup>A</sup> TTATCTT	Blank
hsa-mir-519a-2	TTGGCTCAGAGTCCATTG		TCCAGCCTAATCCTCTT		mml-pd-519a-2
					-

\* Duplication occurred in one of two species

# Highly expressed by count of deep sequencing read larger than 300

The  $P_{\text{substitution}}$  are drawn on the line chart along the sequence. Scatter plots show the relationship between two paralogs and orghotlog pairs within a local region of sequence. The alignment of the site with difference we interested is shown.



#### **Multiple Sequence Alignment**

	134	. 136	1:	56	
	130 140	130 140	150 160	150 160	
	· · · · I · · · · I · · · I · ·			••••••••••••••••••••	
# hsa-mir-512-1			<mark>-</mark>		* mml-mir-512-1
# hsa-mir-512-2				<mark>-</mark>	* mml-mir-512-1
# hsa-mir-1323					mml-pd-1323
hsa-mir-498				<mark>-</mark>	mml-mir-498
hsa-mir-520e	TATGTATASTAGTCGCT		GCTTTTTCTTTCGGGGGAC		Blank
hsa-mir-515-1	TATGTGTG <sup>G</sup> TAGGCATT	CATGCATAGTAGGCATT	GCTTTTTCTCTTTGGGGAC	GCTTTTTCTCTTTGGAGAC	mml-pd-Up1
hsa-mir-519e	TATGTGTG <sup>G</sup> TAGGCTTT	TATGTGTAGTGGGCTTT	GCTTTTTCTCTTTGGAGAC	GGTTTTTCTGTTTTGAGAC	mml-pd-Up2
hsa-mir-520f	TATATGTG5TAGGCATT	TATGTGTÄGTGGGCTTT	GCTTTTTCTCTTTGGGGAC	GGTTTTTCTGTTTTGAGAC	mml-pd-Up3
hsa-mir-515-2	TATGTGTGGTAGGCATT	TATGAGTAGTAGGCATT	GCTTTTTCTCTTTGGGGAC	GCGTTTTCTCTTTGGAGAC	mml-pd-Up4
hsa-mir-519c	TATGTATA <sup>G</sup> TCGGCATT	TATATATAGCAGGCATT	GCTTTTT	GCTTTTT	mml-mir-519c
hsa-mir-1283-1	TATGTCTASTAGGCATT	TATGTATGTCAGGCATT	GCTTTTTCTGTTTGAGAAC	<b>GCTTTTTCTCTTTG</b> GGAAC	mml-pd-1283-1
# hsa-mir-520a	TATGAGTAGTAGGCATT		GCTTTTTCTCTTTGGGGGC		mml-mir-520a
hsa-mir-526b	TGT_AGTAGGCATT		GCTTTTTCTCTTTGGGGAC		mml-mir-526b
hsa-mir-519b	TATECATASTAGECATT	TATGTGTAGTAGGCATT	CTTTTTCTTTCT	SCTTTTTCTTTCTA GAGAC	mml-pd-519b
hsa-mir-525	TATCTCTASTACCCACT	TATCCATACTACCCATT	COTTENTOTECACCO	COTTECTOTECACAT	mml-pd-525
hsa-mir-523	TATOTOTASTAGGCACT	TATCTCTACTACCATT	CONTRACTOR	SCITICICITICIOGAGAT	mml-mir-523a
hsa-mir-518f	TATGICIASTAGICATI	TATGTGTAGTAGTATT	GCTTTTTCTCTTTGGAGAC	GETTTTTCTCTTTGGAGGC	mml-mir-518f
hsa-mir-520b	TACATATGLCAGGCATT	TATGTGTATTGGGCATT	GCTTTTGCTCTTTGAGAAC	GCTTTTTCTCTTTGGAGAC	mml-mir-519a
hsa-mir-518b	TATGICIASCAGGIATI	TACATCTAGCAGGTACT	GCTCTTTCTCTT-GCAGAC	GCTTTTTCTCTT-GGAGAC	mml-mir-518b
hsa-mir-526a-1	TATATGTASCAGGCATT	TATATGTAGTAGGCATT	GCTTTTTCTCTTTGGGGAC	GCTTTTTCTCTTTAGGGAT	mml-pd-Md1
hsa-mir-520c	TATGTGTA <sup>®</sup> TAGGCAAT	TATGTGTATTGGGCATT	GCTTTTTCT <sup>G</sup> GAGAC	GCTTTTTGTCTTTGGGGAC	Blank
hsa-mir-518c	TATGTGTA <sup>G</sup> TAGGCATT		GCCTTTTCTCTTTGGGAAC		mml-mir-518c
hsa-mir-524	TATGTGTA <sup>G</sup> TGGGCATT	TATGTGTAGTGGGCATT	GCTTTTTCTCCTTGGAGAC	GCTTTTTCTCCTTGGAGAC	mml-mir-524
# hsa-mir-517a					mml-mir-517a
hsa-mir-519d					mml-mir-519d
hsa-mir-521-2	TATGTGTAGTAGGCATT	TATGTGTATTGGGCATT	GCTTTTTCTCTTTAGAGGC	GCTTTTTCTCTTTGGGAAC	Blank
hsa-mir-520d	TACGTGTA <sup>G</sup> TAGGCATT		GTTTTTTCTTAGCAGATAC		mml-mir-518a
# hsa-mir-517b	TATATGTAGTAGGCGTT	TATATGTAGTAGCGTT	GCTTTTTCTCTTT-GAAGA	GCTTTTTCTCTTTGAAGGC	Blank
hsa-mir-520g					mml-mir-520g
# hsa-mir-516b-2	TATGAGTASTAGGTATT	TATGAGTAGTAGGTATT	CCTTTTTCTCTTTGGGGAC	CCTTTTTCTCTTTAGGGAC	mml-pd-516b-2
* hsa-mir-526a-2	TGCACGTASTAGGCATT	TATGTGTACTGGGCATT	GCTTTTTCTTTCTGGAGAC	GCTTTTTCTTTCTAGAGAC	mml-pd-Md3
* hsa-mir-526a-2	TGTGTCTGTTAGGCATT	TATGTGTACTGGGCATT	GATTTTTCTTTCCGGAGAA	GCTTCTTCTTTCTAGAGAC	mml-pd-Md4
hsa-mir-518e	TGTGTCTGTTAGGCATT	TATGTGTACTGGGCATT	GATTTTTCTTTCCGGAGAA	GCTTTTTCTTTCTAGAGAC	mml-mir-523b
hsa-mir-518a-1	TATGTGTASTAGGCATT	TATGTGTACTGGGCATT	SCTCTTTCTCTTTA SAGAC	GCTTCTTCTTTCTAGAGAC	mml-pd-Down1
hsa-mir-518d	TACCTCTASTACCCATT	TATGTGTATTGGGTATT	CONTRACTOR	COTTUTICTOTUTIC	mml-pd-Down2
# hsa-mir-516b-1	TATCTCTASTAGCCATT	TATEAGTAGTAGCATT	COTTENT	COTTITICTOTITICGAGAC	mml-pd-Down3
hsa-mir-518a-2	TATOTOTASTAGGCATT	TATCACTACTACCOATT	CONTINUE	SCTTTTCTCTTTCGGGAC	mml-pd-Down4
hsa-mir-517c	TATGTGTASTAGCATT	TATCTCTATTGCCCATT	GCTTTTTCTTTCTAGAGAC	GETTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT	mml-pd-Down5
# hsa-mir-520h	TATGTCTAS IGGGCATT	TATGTGTATIOGGCATT	GCTTTTTCTTTTTGGGGAC	GCTTTTTCTTTCCAGAC	Blank
Blank		TATGTCTAGCOGGCACT		GCTTTTTCTCTTTGGGAAC	mml-pd-Down7
hsa-mir-521-1	TATGIGIASTAGGCITT		GCTTTTTCTCTTTGGGGAC	-CTTCTTTCCAGAGAC	mml-mir-519b
Blank	IAIGTCTA®TCGGCATT	TATGICIAGICGGCATT	GATITTCTTCTGGAGAC	GATITTTCTTTCTGGAGAC	mml-mir-521
hsa-mir-522	IAIGTGTA <sup>®</sup> TAGGCATT	CATGIGIAGIAGACATT	GCTTTTTCTTTCTGGAGAC	GCGTTTTCTCTTTGGCGAC	mml-mir-518e
hsa-mir-519a-1	TATGAGTA <sup>®</sup> TAAGCATT	TATGTGTAGTCAGCATT	GCTTTTTCTCTTTGGGAAC	GCTTTTTCTTTCTGGAGAC	mml-mir-519a-1
hsa-mir-527	TATGTGTA <sup>S</sup> TAGGCATT	TACGTCTAGTAGACATT	GCTTTTTCCCTTTAGAGAC	GCTTTTTCTCTTTGGGAAC	mml-mir-527
hsa-mir-516a-1	TATGTGTA <sup>G</sup> CAGGCATT	TATGTGTAGTAGGCATT	GCTTTTTCTTTCCAGAGAC	GCTTTTTCTCTTTAGAGAC	mml-mir-516a-1
hsa-mir-1283-2	TATGTGG- <sup>G</sup> C <sup>G</sup> GGC	TATGTAT-GT <sup>A</sup> GGCATT	TGGTGGCT <sup>C</sup> ATGCC	GCTTTTTCTCTTT <mark>G</mark> GGAAC	mml-pd-1283-2
hsa-mir-516a-2	TATGTGTAGCAGGCATT	TATGT <mark>G</mark> GGC-T-	GCTTTTTCCTTCCAGAGAC	GTCTAATGTCACTGCTTTT	Blank
hsa-mir-519a-2	TATATCTA <sup>G</sup> TAGGCATT	T-	GCTTTTTCTTTCTGGAGAC	GCTTTTTCCTTCTGGAGAC	mml-pd-519a-2

\* Duplication occurred in one of two species

# Highly expressed by count of deep sequencing read larger than 300

The  $P_{\text{substitution}}$  are drawn on the line chart along the sequence. Scatter plots show the relationship between two paralogs and orghotlog pairs within a local region of sequence. The alignment of the site with difference we interested is shown.



#### **Multiple Sequence Alignment**

	13	79	208	208 – 209								
	180	180	200 210	200 210								
# bsa-mir-512-1					* mml_mir_512_1							
# hsa-mir-512-2					* mml-mir-512-1							
# hsa-mir-1323					mml-nd-1323							
hsa-mir-498					mml-mir-498							
hsa-mir-520e	AGGATTGCTTCTTG		AACCTGTTGAGATTTTC		Blank							
hsa-mir-515-1	AGGATTGCCCCTTG	AGGATTGCCCCTTG	AACCTGCTAAGCCTTTG	AACCTGCTGAGCCTTTG	lau-ba-lmm							
hsa-mir-519e	AGGGTTGCCCCTTC	AGGATTGTCCCTTG	AACCTGCTGAGCCTTTG	AACCTGCTGAGCCTTTG	2qU-bq-lmm							
hsa-mir-520f	AGGATTGCCCCTTG	AGGGTTGTCCCTTG	AACCTGCTGATTCTTTG		EqU-bq-lmm							
hsa-mir-515-2	AGGATTGTCTCTTG		AACCTGCTGASCCTTTG	AACCTGCTGAGCCTTTG	mml-pd-Up4							
hsa-mir-519c	AGGATTOTETETTO	AUGATIOTECETTO	AACCTOCTOASCCTTTO	AACCTOCTOAGCCTTTG	mml-mir-519c							
hsa-mir-1283-1		ACCATTCTCCCTTC	AACCTCCTCASCCTTTC	AACCTECTEAECCTTTE	mml-pd-1283-1							
# hsa-mir-520a	AGGATTGCCCCTTG	AGGATIGICCCTIG	AACCTGCTGASCCTTTG	AACCIGCIGAUCCITIG	mml-mir-520a							
hsa-mir-526b	AGGATTGCCCCTTG		AACCTGCTGAGCCTTTG		mml-mir-526b							
hsa-mir-519b	AGGATTGCCCCTTG	ACCATTCACCCTTC	AACCTGCTGATTCTTTG		mml-pd-519b							
hsa-mir-525	AGGATTTCCCCTTG	AGGATTGACCCTTG	AACCTGCTGAGCCTTTG	AACCTGCTGATTCTTTG	mml-pd-525							
hsa-mir-523	AGGATIGCCCCTIG	AGGATEGTEEETEA	AACCIGCIGAGCCITIG	AACCTGCTGAGCCTTTC	mml-mir-523a							
hsa-mir-518f	AGGATIGCCCCTIG	AGGATTGGCCCTTG	AACCIGCIGA	AACCIGCIGAGCCIIIG.	mml-mir-518f							
hsa-mir-520b	AGGATTG	AGGATIGCCCCTIG	AACCTGCTGAICCTTTG	AACCIGCIGAGCCIIIG.	mml-mir-519a							
hsa-mir-518b	AGGATTGCCCCTTG	AGGATTGUTCCTCG	AACCTGCTGAL ICTTTG	AACCTGCTGAI ICTTTG	mml-mir-518b							
hsa-mir-526a-1	AGGATTGCGCCTTG	AGGATTG I CCGTTG	AACCTGCTGAGCCTTTG	AAGCTGCTGAGCCTTTG	mml-pd-Md1							
hsa-mir-520c	AGGATTGTCCCTCG	AGGGTTG I CCCTTG	AACCTGCTGA <mark>GA</mark> CTTTG	AACCTGCTGAGCCTTTG.	Blank							
hsa-mir-518c	AGGATTGCCCGTTG		AACCTGCTGATTCTTTG		mml-mir-518c							
hsa-mir-524	AGGATT-CCCCTTG	AGGATT - <sup>C</sup> CCCTTG	AATCTGCTGA <sup>GC</sup>	AACCTGCTGAGC	mml-mir-524							
# hsa-mir-517a		<b>-</b> ATTA	AACCTGCTTATTCTTTG	AACCTCCTTATTCTTTG.	mml-mir-517a							
hsa-mir-519d					mml-mir-519d							
hsa-mir-521-2	AGGGTTGCCCTTG	AGGATCGTCCCTTG	AACCTGCAGAGCCTTTG	AACCTGCTGAGCCTTTG	Blank							
hsa-mir-520d	AAGATTG <sup>T</sup> CCTATG		AACCTGCTGATTGTTTG		mml-mir-518a							
# hsa-mir-517b	ACAGTTGTCCCTCG	AAGATTGCCCCTTG	AACCTGCTGAACGTTTG	AACCTGCTGAATGTTCG	Blank							
hsa-mir-520g					mml-mir-520g							
# hsa-mir-516b-2	ACGATCGCCCTTG	AGGATTGCCCCTTG	AACCTGCTGAGCCTTTG	AACCTGCTGAGCCTTTG	mml-pd-516b-2							
* hsa-mir-526a-2	AGGATTGCCCCTTG	AGGATTGGCCCTTG	AACCTGCTGAGACTTTG	AACCTGCTGATTCTTTG	mml-pd-Md3							
* hsa-mir-526a-2	AGGATTTCCTCTTG	AGGATTGTCCCTTG	AGCCTGCTGAGCCTTTG	AACCTGCTGATTCTTTG	mml-pd-Md4							
hsa-mir-518e	AGGATTTECTCTTG	AGGATTGCCCCGTC	AGCCTGCTGAGCCTTTG	AACCTGCTGATTCTTTG	mml-mir-523b							
hsa-mir-518a-1	-GGATTGCCCTTG	AGGATTGTCCCTTG	AACCTGCTGATTCTTTG	AACCTGCTGATTGTTTG	mml-pd-Down1							
hsa-mir-518d	AGGATTGCCCTTG	AGGATTGCCCCTTG	AACCTGCTGATTCTTTG	AACCTGCTGAGCCTTTG	mml-pd-Down2							
# hsa-mir-516b-1	AGGATTGCCCCTTG	AGGATCGTCCCTTG		AACCTGCTGAGCCTTTG	mml-pd-Down3							
hsa-mir-518a-2	AGGATTGCCCTTG	AGGATCGTCCCTTG		AACCTGCTGAACCTTTG	mml-pd-Down4							
nsa-mir-51/C	AGGATTGCCCTTG	AGGATTGCCCCTTG		AACCTGCTGATTCTTTG	mm1-pd-Down5							
# nsa-mir-520n	AGGATTO-LCCTTG		AACCTOCTOA	AACCTGCTGAGCCTTTG	Blank							
Blank	ACCETTO	AGGATTGTCCTGTG	AACCTCCTCASCCTTTC		mml-pd-Down/							
nsa-mir-ozi-l Dloph	ACCONTOPLECTIC	AGGATTGETCCTCC	AACCTGCTGASCTTTG	AACCTGCTGATTCTTTC	mml-mir-519b							
biank bes min 522	AGGGTTGECCCTGC	AGGATTGACCCTTG	AACCTGCTGASCCTTTG	AACATGCTGAGCGTTTG	mml-mir-521							
has min 510a 1	AGGATTGPLLLTIG	ACCATTCETTCETTC	AACCTGCTGA	CACCTOCTCATTC	mmi-mir-Sive							
hsa-mir-527	AGGATTGECCCTTG	AGGATTCCCCCTTC	AACCIGIIGA	AACCTGCTGATTCTTG	mmi-mir-519a-1							
hsa=mir=527	AGGATIGECCCTIG	AGGATTGECCCTTG	AACT I GCAGA <sup>L I</sup> C ITTG	AACCTOCTOCTOCTOCT	mm1 - m1r - 52/							
hea_mir_1202_2	AGGATTGECCCTTG	AGGATTGCCCCTTG	AACCTGCTGA	AACCIGCIGATICITIG	mml-mir-siba-l							
hea-mir-516a-2		GGGATTGTCCCTTG	AACCTGCTGA	AACATGCGGABCCTTTG	nunt-pa-1283-2 Black							
hearmin-510a 2	AGGATTG CCCTTG	AGGATTGUCCCATG	AACCTGCTGA	AACCTGCTGAGCCTTTG	Bidlik mml nd Ellos 2							
iisa-mir-pryd-z	AGGATTGCCCTTG	AGAATTG <mark>CCCC</mark> TTG	AACCTGCTGA	AACCTGCTGATIGTTTG	mmi-pa-siya-2							

\* Duplication occurred in one of two species

# Highly expressed by count of deep sequencing read larger than 300

## Figure 17 Hypothetical Model of Co-Evolution

We made an assumption which implies the co-evolution of microRNA and Alu elements. Details of our model are described in the context.



# Figure 18 The Alignment of Reads from Deep Sequencing to C19MC Annotated Mature Region

All microRNA genes are lined as black on the top and highlighted their mature region as light blue. We aligned each sequencing reads (list on the left) with the microRNA genes to find whether the read is matched to the mature region. Arrow indicates the sequencing reads which aligned to the non-mature region or stem loop of microRNAs.

Alignment of Reads to C19MC genes

C19MC Genes are labeled as black, annotated mature regions are highlighted as light blue.



#### Figure 19 C19MC Seed Selection from Human Sequencing Data

Each sequencing reads were aligned to the microRNA genes to determine the possible mature sequences. As the figure shown, the mature microRNA sequences could be shifted and not identical in each time of expression. Due to this reason, we selected human seed candidates depend on real expression profile instead of mirbase annotation.



#### Figure 20 Rhesus Seed Candidates Selection

a) An example of our alignment of sequencing reads with microRNA gene. The real expression regions of mature microRNA are not conserved. b) Rhesus seed candidate selection was referenced with annotated human microRNA genes and sequenced human mature microRNAs, and selected by  $\pm 3$ -nt shift.



#### Rhesus Seed Candidates are Selected by 3-nt Shift of Human Seed Candidates

[Example] Human micro Human Seed Candid	RNA ate	CATTCTCCAAAAGAAA TCTCCAA	From Annoatation From Deep Sequencing
Rhesus microRNA Homo	log	CATTATCCAAAAGATA	From Dotplot Definition
	0	TATCCAA	
	+1	ATCCAAA	
	+2	TCCAAAA	
Selected Rhesus Seed	+3	CCAAAAG	
	-1	TTATCCA	
	-2	ATTATCC	
	-3	CATTATC	





Figure 22 **Targeting Preference of Seed Candidates: Different Alu Subfamilies** The diagram presented the number of Alu subfamilies targeted by seed candidates within species.



LN Transformation



Targeting Preference of Rhesus Seed Candidates

#### Figure 23 Interchange Seeds to Target Genomes of Each Other





53

## Figure 24 Percentages of Alu Subfamilies Not Targeted by C19MC

All Alu elements were considered because they might be targeted in the past no matter what situation they are now.



Type of Alu Subfamilies

#### Table 1 Ortholog Pairs Defined by Rainbow Dotplot in this Study

The inconsistent rhesus microRNAs of our annotation with miRBase is marked as green. That means our sequence analysis supports the ortholog pairs but which are annotated in different name by miRBase. The microRNAs which not annotated by miRBase are named and marked as red. If a microRNA is annotated by miRBase, we kept the name originated from miRBase in this study even its name could be doubtful.

Name	Posi	tion	Name	Position	
Human microRNA	Relat 5416 Start	ed to 8188 End	Rhesus microRNA	Related to 59771000 Start End	Note
hsa-mir-512-1 <sup>1</sup> hsa-mir-512-2 <sup>1</sup>	1745 4223	1828 4320	mml-mir-512-1		Two orthologs are found in human
hsa-mir-1323	7034	7106	mml-pd-1323		
hsa-mir-498	9263	9386	mml-mir-498	•	
hsa-mir-520e	10777	10863	mml-mir-520e <sup>2</sup>		Human only
hsa-mir-515-1	14069	14151	mml-pd-Upl	E	
hsa-mir-519e	15006	15089	mml-pd-Up2	<b>X</b> E	
hsa-mir-520f	17225	17311	mml-pd-Up3	° IE	Duplication hotspot
hsa-mir-515-2	20075	20157	mml-pd-Up496	IE .	
hsa-mir-519c	21535	21621	mml-mir-519c		
hsa-mir-1283-1	23547	23633	mml-pd-1283-1		
hsa-mir-520a	25947	26031	mml-mir-520a		
hsa-mir-526b	29459	29541	mml-mir-526b		
hsa-mir-519b <sup>3</sup>	30279	30359	mml-pd-519b		
hsa-mir-525	32599	32683	mml-mir-525		
hsa-mir-523	33451	33537	mml-mir-523a		
hsa-mir-518f	35081	35167	mml-mir-518f		
hsa-mir-520b	36293	36353	mml-mir-519a $^4$		
hsa-mir-518b	37803	37885	mml-mir-518b		
hsa-mir-526a-1	41318	41402	mml-pd-Md1		
hsa-mir-520c	42519	42605	mml-pd-Md2	No Sequence	
hsa-mir-518c	43801	43901	mml-mir-518c		
hsa-mir-524	46068	46154	mml-mir-524		

<sup>&</sup>lt;sup>1</sup> Duplication of mml-mir-512-1

<sup>&</sup>lt;sup>2</sup> This microRNA appears in miRBase without any position annotation.

<sup>&</sup>lt;sup>3</sup> Do not confuse with mml-mir-519b annotated by miRBase.

<sup>&</sup>lt;sup>4</sup> This rhesus microRNA sequence is similar to the human one and probably is an ortholog, but it is annotated in a doubtable name by miRBase.

Name	Posi	tion	Name	Posi	cion	
Human microRNA	Relat 5416	ed to 8188	Phesus microRNA	Relate 5977	ed to 1000	Note
	Start	End	KIESUS MICIONNA	Start	End	Note
hsa-mir-517a	47334	47420	mml-mir-517a			
hsa-mir-519d	48413	48500	mml-mir-519d			
hsa-mir-521-2	51660	51746	mml-pd-521-2	No Seq	uence	
hsa-mir-520d	55162	55248	mml-mir-518a <sup>5</sup>			
hsa-mir-517b	56142	56208	mml-mir-517b	No Seq	uence	
hsa-mir-520g	57232	57321	mml-mir-520g			
hsa-mir-516b-2	60508	60592	mml-pd-516b-2			
hsa-mir-526a-2	61988	62052	mml-pd-Md3 mml-pd-Md4			Two orthologs are found in rhesus
hsa-mir-518e	64904	64991	mml-mir-523b <sup>4</sup>			
hsa-mir-518a-1	66072	66156	mml-pd-Down1			
hsa-mir-518d	69943	70029	mml-pd-Down2			
hsa-mir-516b-1	71911	72000	mml-pd-Down3	8		
hsa-mir-518a-2	74399	74485	mml-pd-Down4	ĬE		Duplication hotspot
hsa-mir-517c	76379	76473	mml-pd-Down56	IE .		
hsa-mir-520h	77578	77665	mml-pd-Down6	No Seq	uence	
hsa-pd-Down7			mml-pd-Down7 6	*		
hsa-mir-521-1	83702	83788	mml-mir-519b <sup>4</sup>			
			mml-mir-521			Rhesus Only
hsa-mir-522	86277	86363	mml-mir-518e <sup>4</sup>			
hsa-mir-519a-1	87463	87547	mml-mir-519a-1			
hsa-mir-527	89084	89168	mml-mir-527			
hsa-mir-516a-1	91807	91896	mml-mir-516a-1			
hsa-mir-1283-2	93298	93384	mml-pd-1283-2 7			
hsa-mir-516a-2	96199	96288	mml-mir-516a-2			
hsa-mir-519a-2	97410	97496	mml-pd-519a-2			

## Table 1 Ortholog Pairs Defined by Rainbow Dotplot in this Study (Continuous)

<sup>&</sup>lt;sup>5</sup> This rhesus microRNA sequence is similar to the human one and probably is an ortholog, but it is annotated in a doubtable name by miRBase.

 $<sup>^{\</sup>rm 6}~$  LTR13 inserted into the upstream of this microRNA in both human and rhesus.

<sup>&</sup>lt;sup>7</sup> MicroRNA harbored on Alu elements with the same genomic direction in rhesus.

	DAAO'I"I'	L OTT	-			-	-		-	+	+-			_	+	+	_	-				_	H	_	-	$\vdash$	-		_	+	_	_	+			54643	912 STETOE	8005
	DTADDT	ε									Ħ																			Ħ						4382	2029	552
	DEAGGG AAADDT	S T	┝			_	-							-	+	+	-	-		_	$\left  \right $		H	_	-		+		_	$\left  \right $	-	_	-			1547 468	568T 9TS	735 73
	ADTTOT	29																								Ħ				Ц						12328	595 <i>L</i> T	LL₽T
	ADDTDT ADDTDT	059 22	┝				+		-		+				+	+	-									$\mathbb{H}$				+	+					8063 5206	Z6∳6I T606S	908E 605E
	ADDIT	58 T (									Ц																									5087	15310	TSZ
	ADATDT	2																																		3336 1922	08TT	345
	ADACOT DDTDDT	ττ₩ /6	L		-	_	-		_	-		_			_		_	_						_			_			+	_	_	_		_	56 <i>L</i> Z	9576 57##T	5537 5888
	AAADDT	τ																																		2848	14500 14510	2872
	ADAADT DDAADT	Т 9	⊢				-		-+	+	┢	+	+	_	+	+	_	_	$\vdash$	_	+	+	+	_	-		+	+	_	┢	-	-	+			12020 15020	15730 1956	849T 00E
	AAAADT	τ																												П						1217	627	ST
	TTODTD AAADAT	89 T	⊢		+	-									+	+				-			+			$\left  \right $	+			+	-		+			258 603	795 2696	174 98
	GTGCTG	15				_			_		П				_		_	_		_									_	П			_		_	560S	907-8	6₩8I 77
	CAAAGC	τ																																		124	EST.	ST
	OTCTGG DADDTD	τ 0τ	H		+		-				+			_	+	+	_						$\square$		-	$\mathbb{H}$				+	_	-	+			€365	8453 12035	ESST SOTE
ds)	<b>BATDTD</b>	50					L				Ħ																									7025	₽₽0T	τ8
rea	DADDTD DATDTD	τ	⊢		+	-					+				+	+		-		-			+			$\left  \right $	+			+	-					542 142	288 655	59 59
+ of	AADDTD	9 70				_			_		$\square$				_	$\square$	_	_					$\square$	_			_		_			_	_			80T9	5828	9₽6 ₩96/7T
E .	DADATO DDADATO	τ																																		26523	2201	222
Leve	ADDTDD AADATD	13 18	$\vdash$			_			-					_	+	+	_	_						_		$\vdash$	+		_	++	+	_	+		_	1582 923	260T 85€T	18T 61
sion	CGAGGA	τ									H																			H						2248	0988	5190
ores	9A9ADD ATDTDD	4 3																												$\mathbf{H}$						2040 1228	580T	121 671
EX1	OODAAO DTOTAO	6 <i>L</i> T					-		_		$\left  \right $			_	+	+	_									$\square$	_			$\left  \right $	-	_	+			2029 8/281	60T8 TT#0Z	007 7813
and	TAĐAAD	τ									Ħ															Ħ										10010	8001	2655
Seed	DAAAAD TDAAAD	300 080	⊢			_	-				+		$\left  \right $	-	+	+	-	-		_						$\mathbb{H}$	+				+	-	+			0TSLTT 08ST	6T6ETÞ 88ST	66∉7∠ 787
	APCTO	528									П															П				П						5251	1855	69T
	TODTDA	S														. 1																				2233	4750 9717	962 61⊅
	CODTDA TODTDA	965 67									$\left  \right $												4			$\square$	+			$\left  \right $	-	_	+			T05542 5083	387221 3975	98925 354
	AGTGCA	τ									Ħ				4															Ħ						3295	₩12₩	56₽
	TOODDA DODDDA	8	⊢		+	-					+			/	4	E					Ì									+	-		+			TOSC	30230 STZ6	579 7975
	DACI'DA DDDADA	Т т				_					$\square$	1			_		_			1						$\left\{ \right\}$			_	$\square$			_			714 5756	τ6 <i>L</i> ₽ /ε6Ζτ	4TTT 20TT
	ACGCGC	95																				1		-1												185	2172	29
	ACGCAC	L	_		+	-	-	_	-			-	-	-	+			-					H				-		_	+-		_	+		_	437	05L 785	48 55
	ACAAAG	7																																		426	000	
	ODTDAA ODTDAA DAAADA	7 799 7																ſ			2					4				L						450 110580 3831	₽TTE0₽ /₩Z6	55686 T#8T
	DƏTƏAA DƏTƏAA DƏTƏAA DƏTƏAA ƏAAADA	۲ 995 5 89																			5	916	5													436 110386 3831 11366	403114 403114 12814	SS686 T₱ET ₱T6T
	2020244 0202044 2070544 2070544 2070544 044420	₹ 995 89 55 55 122																			5															436 3837 3837 343 543 5430	403114 6344 71814 632 2041	55686 T≯€T ∳T6T 667 008
	AAGGGG AAGGGG AAGGGG AAGTGC AAGTGC AAGTGC AAGTGC	₹ 995 98 55 55 172 50																			3															450 110580 3837 543 5430 153 153	403114 63742 12814 632 829 829	58622 13∉1 761∉ 560 800 917
	APAGTG BAPAGG AACGCG AACGCG AACGCG AACGCG AACTGC AACTGC AACTGC AACTGC	4 989 55 55 724 724 50 70 759 759																			5															430 110380 3831 5430 133 5430 133 1080 1080 1044	<pre>&gt; &gt; &gt;</pre>	689522 1341 500 500 517 517 517 518 538 538 538 538 538 538 538 538 538 53
	872545 872555 872555 8725555 8725555 8755555 8755555 8755555 8755555 8755555 8755555 8755555 8755555 8755555 8755555 8755555 87555555 87555555 87555555 87555555 87555555 8755555555	4 98 55 55 52 76 76 76 750 750 750 750 750 750 750 750																																		430 170380 3837 5430 5430 7530 7080 7080 7080 7080 7080 7080 7080 70	40317 40317 40317 4032 4047 526 1697 4023 4023 5047 5047 5047 5047 5047 5047 5047 5047 5047 5047 5047 5047 5047 5047 5047 5047 5047 505 505 505 505 505 505 505 50	55686 1341 560 500 800 517 518 8183 124 62 24 25 26 26 26
	T7AAAA DT7AAA D7DAAA D7DAAA D7DAAA D7DAAA D7DAAA D5DDAA D5DDAA D5DDAA D5DDAA D5DTAA D2DTAAA D2DTAAA DAAADA	4 98 55 86 72 72 72 72 73 73 75 75 75 75 35																			5															426 3430 3431 5430 5430 5430 102144 102144 5550 2250 2350 200	2000 200 2000 2	68622 1341 1314 500 800 800 8178 154 82783 154 43 1580 43 5180
	a bit	4 (Kpb) 35 442 35 442 35 442 35 442 35																				Auta														426 3831 110586 3831 1336 5430 5430 101144 5530 2250 5001 2250	403114 403114 556 71821 526 538 638 638 7693 7693 7693 7693 7693 7693 7693 7693	68822 1341 535 536 800 536 800 5378 43 5380 43 5380 5380 43 5380
	hind be and be and and and and and and and and and and	<pre>&lt;1.5Kbp) 21.5Kbp) 32.662 447 32.66 447 32.66 447 32.66</pre>																				Sx1, AluY							×							426 3426 3431 3431 3430 343 3430 343 3430 3430 3	403114 5342 5342 638 7460 807 4603 800 4603 800 4603 800 800 800 800 800 800 800 800 800 8	<pre>4 4 5 6 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8</pre>
	thin think the second s	on (<1.5Kbp) 32 647 662 32 2447 56 2 32 2447 56 2 32 2447 56 2 32 56 2 447 56 2 32 56 2 32 56 7 32 56 7 35 7 35 7 35 7 35 7 35 7 35 7 35 7 35							AluSz													AluSx1, AluY							AluSx	luSx						426 426 3831 11396 243 2539 25430 2540000000000	403114 6347 6347 638 804 803 804 803 803 804 803 804 804 804 804 804 804 804 804 804 804	Target Alur 1240 200 43 724 800 43 724 734 734 734 734 734 734 734 734 734 73
	is strain trie and trie and tr	iection (<1.5Kbp) 32 26 32 26 447 32 26	Sg Sg	Sx1 v ninse					Sz, AluSz	Sz	Sx3	ext							Sg	SX1		SP Yg6, AluSx1, AluY							Sx, AluSx	Y, AluSx				BS BS	Sx	of Target A26 120286 247 247 247 2529 274 267 274 272 2831 2831 2831 2832 2832 2832 2832 283	6 1347 6 1347 6 1491 6 1491 7	of Target Alur 7340 299 299 299 216 2180 43 797 2180 43 797 2180
	AcAAAC AcAAAC Athen Acacco Acaco Acacco Acaco	Direction (<1.5Kbp) 2662 44756 662 44157 2662 44156 662 44156 665 44156 665 44156 665 44156 665 665 665 665 665 665 665 665 665	AluSg	AluSx1 AluSx1	The converse of a second				AluSz, AluSz	AluSz	AluSx3	XSNTY							AluSg	TXSNTE		AluSp AluYg6, AluSx1, AluY							AluSx, AluSx	AluY, AluSx Alus				AluSg	AluSx	# cf Target 426 Target 710286 2831 2743 2832 2009 277 2832 2832 2833 2833 2833 2833 2833	+03114 3247 934 17814 935 17814 935 4693 397774 4693 397774 4693 397774 4693 397774 4693 397774 4693 397774 4693 397774 4693 397774 4693 397774 4693 4693 4693 4693 4693 4693 4693 469	# of Target Alux 797 799 198955 799 198955 799 799 799 799 799 799 799 799 799
uo	ALTERM ALTERM	Up Direction (<1.5Kbp) 26 662 76 447 01757 22 26 747 22 26 7447 22 26 7467 22 7467 22 26 7467 22 26 7467 22 26 7467 22 26 7467 22 26 7467 22 26 7467 22 26 7467 22 26 7467 22 26 7467 22 26 7467 22 7467 22 7567 22 26 7467 22 7567 22 26 7467 22 7567 22 7567 22 7567 22 75777 22 75777 22 75777 22 75777 22 75777 22 75777 22 75777 22 75777 22 75777 22 75777 22 75777 22 75777 22 75777 22 75777 22 75777 22 757777 22 757777 22 7577777777	AluSg	Alusx1	T. colores / solars				* AluSz, AluSz	* Alusz	* AluSx3	* XSULA		*					AluSg	* TXSXL	*	AluSP AluYg6, AluSx1, AluY				*	* *		* AluSx, AluSx *	* AluY, AluSx		*	* **	* AluSg	* Alusx	# 0f Target Alo	+03114 603114 71414 826 82041 801 801 801 801 801 801 801 801 801 80	# of Target Alur 797 797 797 797 798 799 734 734 734 734 734 734 734 734 734 734
u	antinent stream ting Exc properties the shing Exc process anting the shing the shing the shing antin antin anting anting anting	3'D Up Direction (<1.5Kbp) 2662 32647 32756 32756 32757 32756 327577 32757 32757 327577 327577 327577 327577 327577 3275777 3275777 3275777 32757777 32757777777777	AluSg	Alusx1			*		* AluSz, AluSz	* AluSz	* AluSx3	* XSULA *		*					Alusy	* * * ALUSXI		* AluY96, AluSx1, AluY		* *			* *		* AluSx, AluSx + *	* AluY, AluSx			K *	* AluSg	* & AluSx * *	426 436 3831 3831 1238 3831 1238 2430 2430 2539 2430 25400 254000 2540000000000	# of Tarree Alues # 17814 #	# of Target AluY 809 55 1240 299 1240 299 1240 299 1240 299 2180 43
ц¢	riven riven riven riven riven riven riven riven riven restrem	And d 1 Direction (<1.5Kbp) 1 22 6447 22 6462 23 6622 23 6622 23 6622 24 7 25 6652 25 6652 26 7477 27 6652 28 6 28 6 28 6 28 6 28 6 29 7 20 7 2	AluSg	Alusx1			*	*	* AluSz, AluSz	* Alusz	* AluSx3	XSULA *		* *					AluSg	TXSNTY * *		* AluSp * AluYg6, AluSx1, AluY		* *		*			* * AluSx, AluSx	a * AluY, AluSx			K *	* AluSg	* AluSx * *	426 436 112386 3831 23430 2430 2500 2500 2742 2500 2500 2500 2500 2500 2500 2500 25	# of Target Alus 4693114 469314 4693 8041 4693 807 807 807 807 807 807 807 807 807 807	# of Target AluY 1998955 7998155 7998185 7998185 7998185 7998185 79945 799577 79957 79957 79957 799577 799577 799577 799577 799577 7995777 799777 799777 799777 799777 799777 799777777
u(	Kine of the second seco	H RNA d 3 Dprection (<1.5Kbp) 32 26 4126 32 26 662 32 26 662 412 22 86 662	3p AluSg	3p Alusx1 Alusx1 Alusx2			*	£ * * *	3p * AluSz, AluSz	- dc.	3p * AluSx3	37 * Alusx	32 * 1	* *			* *		3p AluSg AluSg	a * * Atusx1		b * AluSp AluSx1, AluY		* *		3p *	* dg	3p *	c * * AluSx, AluSx	sgene * AluY, AluSx			3 <sup>2</sup>	* AluSg	3p * AluSx a * AluSx	426 12120286 3831 2229 2740 277144 2009 5273 2229 2229 2229 2229 2229 2229 2229 2	+03114 +03114 5542 71814 832 804 807 4633 801 4633 801 801 801	# of 13rdef Alux 13d1 12rdef 200 3299 43 124 43 43 200 43 43 124 43 124 43
u( :	And the second s	3.arm RNA d. D Up Direction (<1.5Kbp) 3.2677 2662 3.2787 2662 3.2847 2662 3.28	512-3p AluSg	512-3p AluSx1 AluSx1	L on and L on and L on and	520e * *	519e * *	520f  *  *	515-3p * AluSz, AluSz AluSz * A	* AluSz * AluSz	* AluSx3 *	519b-3p * Alusx	525-3p * *	523 * *	520b * * *	518b * *	* *	518c * * 5	524-3p AluSg	519d * * AtuSXI 519d * *		52/0d-3p Alusp Alusy Alusy 517b * Alusy1, Aluy	5209 *	516b* * *	518e * *	518-3p * *	516b* * 516b* *	518-3p * *	517c * * AluSx, AluSx AluSx 520h * *	Psudogene * AluY, AluSx	521         *	519a * * * *	16a-3p *	* AluSg	519a 3p * AluSx 519a * * 519a	426 110286 3831 3831 2430 2535 2530 2525 2525 2525 2525 2525 25	+03114 60314 63547 17814 635 1491 804 801 801 801 801 8770 8770 8770 8770 877	# of Target Alux 1241 2396 43 737 737 737 737 737 737 737 737 737
ца :	n in in in in in in in in in i	a 3'arm RNMA d 1'D Uptection (<1.5Kbp) 3'arm RNMA d 2'D 0'2'S 2'S 4'12'S 2'S 4'12'S 2'S 3'2'S 4'12'S 3'S 5'S 5'S 5'S 5'S 5'S 5'S 5'S 5'S 5'S 5	5p 512-3p Alusg	5p     512-3p     AluSx1       a     a     a		520e * * *	* * 5196 * *	520f * *	5P 515-3P * AluSz, AluSz / P / P	ap arac-ap * AluSz * AluSz	5p 520a-3p * AluSx3 [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [ ] [	5 5260* * Alusx 59 519b-3p * Alusx	5p 525-3p *	* 523 * *	* 520b * *	518b * *	* *		5p 524-3p AluSg AluSg	519d * * Alusx1		bp 520d-3p AluSp AluSx1, AluY 517b * AluSx1, AluY	520g *	516b*         *           *         *	* 518e * *	5p 518-3p * *	5p 518d-3p * *	5P 518-3P *	* 517c * * AluSx, AluSx 520h * *	ble Psudogene * AluY, AluSx	521         *	* 519a * * *	5p 516a-3p *	3 * AluSg	5p 516a-3p * AluSx 519a * A	#36 3831 3831 3831 1238 3831 1238 2430 2430 2500 2430 2500 2430 2500 2430 2500 2430 2500 2500 2430 2500 2430 2500 2500 2500 2500 2500 2500 2500 25	#03114 0547 9247 9247 9247 1491 801 4693 39777 801 4697 801 4697 801 4697 801 4697 801 4697 8770	# of Target Alux 1240 239 991855 797 797 797 797 797 797 797
цс :	it the physical states of ACAAAG ACAAAG ACAAAG ACAAAG AAAAAGGG AAAAAGGG AAAAGGG AAAAGGG AAAAGGGG AAAAGGGG AAAAGGGG AAAAGGGG AAAAGGGG AAAAGGGG AAAAGGGG AAAAGGGG AAAAGGGGG AAAAGGGGG AAAAGGGGG AAAAGGGGGG	5. arm 3. arm RMA d Direction (<1.5Kbp) 3. arm 2. 2557 3. 2677 6. 4 3. 3. arm 3. arm 2. 2557 5. 255757 5. 255757 5. 255757 5. 255757 5. 255757 5. 255757 5. 255757 5. 255757 5. 25575	512-5p 512-3p Alusg	1225 512-3p Alusx1 Alusx1 1223	498	515-5p 515-3p * *	519e* 519e * *	520f   *   *   *	315-5p 515-3p * AluSz, AluSz	1283 * AluSz * AluSz	20a-5p 520a-3p * AluSx3 * *	526D 526D* * Alusx 19b-5p 519b-3p * Alusx	525-5p 525-3p * *	523* 523 * *	5181* 5181 * * * 5205 * *	518b * *	526a * * * * * * * * * * * * * * * * * * *	20C-3P 320C-3P * * *	524-5p 524-3p AluSg	2T/x 2T/x x 2D/TG x x 2D/TG 2T/x 2D/TG 2D/		20d-5p 520d-5p AluSp AluSy 517* 517b * AluYe6, AluSx1, AluY	520g *	516b 516b* * * 526a * *	518e* 518e * *	318-5p 518-3p * *	18d-5p 518d-3p * 516b* * 516b* *	518-5p 518-3p *	517* 517c * * AluSx, AluSx 520h * *	robable Psudogene * AluY, AluSx	522* 522 * * * 522 * * *	519a* 519a * * * 	52/ * 16a-5p 516a-3p * 16a-5p 516a-3p * 16a-5p 516a-3p	1283 * AluSg	16a-5p 516a-3p * AluSX 519a * *	426 110286 3831 2337 2430 2747 22909 2707 2707 2707 2707 2837 2837 2837 2837 2837 2837 2837 283	#0114 93247 17814 1491 633 935 1491 663 801 801 807 807 807 807 807 807 807 807 807 807	of Target Alux 1996555 1914 1914 1914 1914 1914 1914 1914 1
ца :	Pavel Mattice	Definition of the section (<1.5Kbp) 3.arm 3.arm and a Direction (<1.5Kbp) 32 26472 86 262 37 26471 252 86 262 37 26471 252 86 2664 1252 88 26644 12528 1252 88 2664 1252 88 2664 1252 88 2664 1252 88 26	34 512-5p 512-3p AluSg	34 512-5p 512-3p Alusx1	91 498 91 498 91 91 91 91 91 91 91 91 91 91 91 91 91	0 515-50 * * * 95 515-50 515-30 * *	0 519e* 519e * *	37 520£ * * *	95 515-5p 515-3p * Alusz, Alusz	03 219C-5P 219C-5P * Alusz * Alusz	52 520a-5p 520a-3p * AluSx3 *	53 526b 526b* * Alusx 95 519b-3p * Alusx	45 525-5p 525-3p * *	17 523* 523 * *	27 5181* 5181 * * * 0 5205 * *	18 518b * *	23 526a * *	87 518c* 518c * *	24 524-5p 524-3p AluSg	10 91/* 519d * * AluSX1 10 91/* 519d * *		49 520G-5p 520G-5p AluSp AluSp 10 517* 517b * AluYg6, AluSx1, AluY 10 517* 517b * AluYg6, AluSx1, AluY	76 520g *	226 516b 516b* * * 337 526a *	15 518e* 518e * *	82 518-5p 518-3p * *	0 518d-5p 518d-3p * * 88 516b 516b* *	82 518-5p 518-3p * *	0 517* 517c * * AluSx, AluSx       22 520h * *	Probable Psudogene * AluY, AluSx	0 522* 522 * * 6 69 522* 522 * *	66 519a* 519a * * * *	10 52/ 37 516a-5p 516a-3p *	12 1283 * AluSg	0 516a-5p 516a-3p * AluSx 12 519a * *	426 110286 3831 3831 2430 2540 2530 2530 2530 2530 2530 2530 2530 253	#05114 0547 17814 935 1491 803 801 4693 801 801 801 8770	# of Target Alur 1998955 998955 43 43 124 43 43 797
u	Appendix and a series of the s	Expr. Le 1 D D Direction (<1.5KDp) 3. arm 3. arm A 3 Direction (<1.5KDp) 32 6 447 28 6 6 6 2 32 26 7 1125 88 6 6 6 6 2 447 78 78 78 78 78 78 78 78 78 78 78 78 78	3 634 512-5p 512-3p AluSg	7 634 512-5p 512-3p Alusx1 5 5312 1323 612-3p Alusx1		6 0 55.515-55 515-35 * * *	3 0 519e* 519e * *	6 37 520f * * *	2 95 515-5p 515-3p * AluSz, AluSz / 1 / 1	6 12 1283 ************************************	4 552 520a-5p 520a-3p * AluSx3 * *	2 53 526D 526D* * AluSX 0 95 519D-5p 519D-3p *	4 45 525-5p 525-3p * *	6 117 523* 523 * *	6 12/ 5181* 5181 * * * 0 0 0 520b * *	2 218 518b * *	4 23 526a * *		6 24 524-5p 524-3p Alusg	6 410 517* 517a * * Atusx1 1 1 2 2 3 4 1 2 2 4 1 2 2 4 2 2 2 2 2 2 2 2 2 2 2		6 410 517* 517b * AluSAL AluSXI, AluY	9 176 520g *	4 626 516b 516b* * * * * * * * * * * * * * * * * * *	7 115 518e* 518e * *	4 82 518-5p 518-3p * *	6 0 518d-5p 518d-3p * * * * * * * * * * * * * * * * * * *	6 82 518-5p 518-3p *	4 0 517* 517c * * AluSx, AluSx	6 NA Probable Psudogene * AluY, AluSx	6 0 22* 522 * * * 6 69 522* 522 * *	4 66 519a* 519a * * * *	4 10 527 9 37 516a-5p 516a-3p *	6 12 1283 * AluSg	9 0 516a-5p 516a-3p * AluSx 519a * AluSx 5 12 519a * *	426 12120286 3831 2229 2229 2229 2229 2229 2229 2229 22	403114 6747 9247 17814 935 1491 633 1491 801 4693 801 4693 801 4693 801 4693 801 4693 801 4693 801 801 801 801 801 801 801 801	# of Target Alur 2399555 43 43 797 797 797 797 797 797 797 797 797 79
ца :	Path Second Seco	Lee rie 1 Direction (<1.5Kbp) 3 arm RNA 3 Direction (<1.5Kbp) 3 26 7 4426 3 3 26 7 3 26 7 4426 3 3 26 7 4426 5 2 2 6 4 4 5 6 6 4 4 5 6 6 4 4 5 6 6 6 6 6 6	5 83 634 512-5p 512-3p Alusg	3 97 634 512-5p 512-3p AluSx1	4 123 91 498	1 86 0 520e * * 3 82 95 515-55 515-35 * *		9 86 37 520£ * * *	5 82 95 515-5p 515-3p * AluSz, AluSz	7 00 03 219C-5P 219C-5P * AluSz * AluSz	9 84 552 520a-5p 520a-3p * AluSx3 / 2000 - 2	7 80 95 519b-5p 519b-3p * AluSX	1 84 45 525-5p 525-3p * *	5 86 117 523* 523 * *	5 86 127 3181* 5181 * * 1 60 0 520b * *	3 82 218 518b * *	3         526a         *           0         0         0         *		2 86 24 524-5p 524-3p AluSg AluSg	3 86 410 51/* 51/a * * Alusx1 August 1 1 1	4 86 0 521 * * *	5 66 410 517* 517b * AluSp AluSx1, AluY	3 89 176 520g *	31         84         626         516b*         *           10         64         37         526a         *	9 87 115 518e* 518e * *	4 84 82 518-5p 518-3p * *	7 86 0 518d-5p 518d-3p * *	3 86 82 518-5p 518-3p *	1 94 0 517* 517c * * AluSx, AluSx 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 106 NA Probable Psudogene * AluY, AluSx / / /	5     8     0     521     *     *       1     86     69     522*     522     *     *	5 84 66 519a* 519a * * * *	2 84 10 52/ 4 89 37 516a-5p 516a-3p *	2 86 12 1283 * AluSg	5 89 0 516a-5p 516a-3p * AluSx 1 86 12 519a * *	426 110286 3831 2229 2229 2229 2229 2229 2229 2229 22	#03114 63247 93247 17814 635 1491 801 801 801 801 8770 8770 8770 8770 87	# of Target Alux 1214 2299 43 797 797 797 797 797 797 797 797
ц.	ACAAAC ACAAAC AACAAAC AACACAAACAAACAAAC	Ice         Ice <td>70016 83 634 512-5p 512-3p AluSg</td> <td>72508 97 634 512-5p 512-3p AluSx1 55064 72 5313 1323 Alive Alive</td> <td>77574 123 91 498</td> <td>79051 86 0 520e * * 32339 82 95 515-55 515-35 * * *</td> <td>83277 83 0 519e* 519e * *</td> <td>85499 86 37 520£ * * *</td> <td>88345 82 95 515-5p 515-3p * AluSz, AluSz * AluSz</td> <td>02009 00 03 D19C-3P 219C-3P * AluSz * 12 1283 * * AluSz</td> <td>94219 84 552 520a-5p 520a-3p * AluSx3</td> <td>91/29 82 53 526b 5266* * Alusx 98547 80 95 5195-5p 5195-3p *</td> <td>00871 84 45 525-5p 525-3p *</td> <td>01725 86 117 523* 523 * * 22255 66 217 523* 523 * *</td> <td>04355 86 127/ 3181* 5181 * * * 04181* 5181 * * * 04181* 5000 * * * * * 04181* 5000 * * * * * * 04181* 5000 * * * * * * 04181* 5000 * * * * * * 04181* 5000 * * * * * * 5000* * * * * * * * 04181* 5000* * * * * * * * * 04181* 5000* * * * * * * * * * * * * * * * *</td> <td>06073 82 218 518b * *</td> <td>09590 84 23 526a * *</td> <td>12089 100 87 518c* 518c * *</td> <td>14342 86 24 524-5p 524-3p AluSg</td> <td>15608 86 410 51/* 51/a * * AluSXI ALUSXI 519d * *</td> <td>19934 86 0 521 * * *</td> <td>23436 86 49 520G-5p 520G-5p 4 AluSp 4 AluSp 23396 66 410 517* 517b * AluYg6, AluSx1, AluY 1 24396 66 410 517* 517b *</td> <td>25509 89 176 5209 *</td> <td>28780 84 626 516b* * * 330240 64 37 526a * *</td> <td>33179 87 115 518e* 518e * * *</td> <td>34344 84 82 518-5p 518-3p *</td> <td>38217 86 0 518d-5p 518d-3p * * * * * * * * * * * * * * * * * * *</td> <td>42673 86 82 518-5p 518-3p *</td> <td>44661 94 0 517* 517c * * AluSx, AluSx AluSx 22 520h * *</td> <td>50471 106 NA Probable Psudogene * AluY, AluSx</td> <td>54976 86 0 0 521 * * 551 551 551 551 551 551 551 551 55</td> <td>55735 84 66 519a* 519a * * * *</td> <td>5/35b 84 10 52/ 60084 89 37 516a-5p 516a-3p *</td> <td>61572 86 12 1283 * AluSg</td> <td>64476 89 0 516a-5p 516a-3p * AluSx 55684 86 12 519a * *</td> <td>426 110286 3831 2347 2430 2747 22909 27900 279000 279000 279000 279000 279000 279000 279000 279000 2790000000000</td> <td>403114 6247 9247 934 1491 633 1491 633 30477 4603 801 4603 801 4603 801 4603 801 4603 801 4603 801 801 801 801 801 801 801 801</td> <td># of Target Alux 1341 2796 239 998185 797 797 797 797 797 797</td>	70016 83 634 512-5p 512-3p AluSg	72508 97 634 512-5p 512-3p AluSx1 55064 72 5313 1323 Alive Alive	77574 123 91 498	79051 86 0 520e * * 32339 82 95 515-55 515-35 * * *	83277 83 0 519e* 519e * *	85499 86 37 520£ * * *	88345 82 95 515-5p 515-3p * AluSz, AluSz * AluSz	02009 00 03 D19C-3P 219C-3P * AluSz * 12 1283 * * AluSz	94219 84 552 520a-5p 520a-3p * AluSx3	91/29 82 53 526b 5266* * Alusx 98547 80 95 5195-5p 5195-3p *	00871 84 45 525-5p 525-3p *	01725 86 117 523* 523 * * 22255 66 217 523* 523 * *	04355 86 127/ 3181* 5181 * * * 04181* 5181 * * * 04181* 5000 * * * * * 04181* 5000 * * * * * * 04181* 5000 * * * * * * 04181* 5000 * * * * * * 04181* 5000 * * * * * * 5000* * * * * * * * 04181* 5000* * * * * * * * * 04181* 5000* * * * * * * * * * * * * * * * *	06073 82 218 518b * *	09590 84 23 526a * *	12089 100 87 518c* 518c * *	14342 86 24 524-5p 524-3p AluSg	15608 86 410 51/* 51/a * * AluSXI ALUSXI 519d * *	19934 86 0 521 * * *	23436 86 49 520G-5p 520G-5p 4 AluSp 4 AluSp 23396 66 410 517* 517b * AluYg6, AluSx1, AluY 1 24396 66 410 517* 517b *	25509 89 176 5209 *	28780 84 626 516b* * * 330240 64 37 526a * *	33179 87 115 518e* 518e * * *	34344 84 82 518-5p 518-3p *	38217 86 0 518d-5p 518d-3p * * * * * * * * * * * * * * * * * * *	42673 86 82 518-5p 518-3p *	44661 94 0 517* 517c * * AluSx, AluSx AluSx 22 520h * *	50471 106 NA Probable Psudogene * AluY, AluSx	54976 86 0 0 521 * * 551 551 551 551 551 551 551 551 55	55735 84 66 519a* 519a * * * *	5/35b 84 10 52/ 60084 89 37 516a-5p 516a-3p *	61572 86 12 1283 * AluSg	64476 89 0 516a-5p 516a-3p * AluSx 55684 86 12 519a * *	426 110286 3831 2347 2430 2747 22909 27900 279000 279000 279000 279000 279000 279000 279000 279000 2790000000000	403114 6247 9247 934 1491 633 1491 633 30477 4603 801 4603 801 4603 801 4603 801 4603 801 4603 801 801 801 801 801 801 801 801	# of Target Alux 1341 2796 239 998185 797 797 797 797 797 797
ца :	ttion ngth argent argen	Item         Item         3         0         D </td <td>54170016 83 634 512-5p 512-3p AluSg</td> <td>54172508 97 634 512-5p 512-3p Alusx1 [4175264 72 532 1223</td> <td>54177574 123 91 498</td> <td>541790511 86 0 520e * * 515-30 520e 520e 520e 8 5515-50 515-30 520e 8 5515-50 515-30 5515-3005-3005-30 5005-3005-5515-3005-5005-5005-5005-5005-</td> <td>54183277 83 0 519e* 519e * *</td> <td>54185499 86 37 520£ * *</td> <td>54188345 82 95 515-5p 515-3p * AluSz, AluSz / P / P</td> <td>24109002 60 03 219C-3P 219C-3P * AluSz 54191821 86 12 1283 * AluSz</td> <td>54194219 84 552 520a-5p 520a-3p * AluSx3</td> <td>5419/729 82 53 526b 226b<sup>*</sup> * AluSX 5419877 80 95 519b-5p 519b-3p * AluSX</td> <td>54200871 84 45 525-5p 525-3p *</td> <td>54201725 86 117 523* 523 * *</td> <td>54203555 86 12/ 5181* 5181 * * * 54204541 60 0 520b * *</td> <td>54206073 82 218 518b * *</td> <td>54209590 84 23 526a * *</td> <td>J#11/33         60         20         30/2005         30/00-35         5           54212089         100         87         518c*         518c         *         *</td> <td>54214342 86 24 524-5p 524-3p AluSg</td> <td>54215608 86 410 51/* 51/a * * AluSX1</td> <td>54219934 86 0 521 * * *</td> <td>54223436 66 410 517* 517b * AluSp AluSx1, AluY 7</td> <td>54225509 89 176 5209 *</td> <td>542287800 84 626 516b* 516b* * * 242387800 84 626 516b* * * 242302401 64 37 526a * 242302401 64 37 526a</td> <td>54233179 87 115 518e* 518e * *</td> <td>54234344 84 82 518-5p 518-3p *</td> <td>542382177 86 0 518d-5p 518d-3p * 54240188 89 388 516b* * *</td> <td>54242673 86 82 518-5p 518-3p *</td> <td>54244661 94 0 517* 517c * * AluSx, AluSx                                      </td> <td>54250471 106 NA Probable Psudogene * AluY, AluSx</td> <td>542519'6 86 0 522* 521 * * 522 * * 522 * * 522 * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * *</td> <td>54255735 84 66 519a* 519a * * * *</td> <td>2425/358 84 10 52/ 54260084 89 37 516a-5p 516a-3p *</td> <td>54261572 86 12 1283 * AluSg</td> <td>54264476 89 0 516a-5p 516a-3p * AluSx 5126a-5p 516a-3p * AluSx 54265684 86 12 519a * *</td> <td>426 110286 3831 2340 2430 2540 2520 2520 2520 2520 2520 2520 252</td> <td>#03114 9247 934 17814 935 1491 804 801 4693 807 801 801 8770 8770</td> <td>0 1349555 1341 1314 1314 1314 1314 1314 1314 1</td>	54170016 83 634 512-5p 512-3p AluSg	54172508 97 634 512-5p 512-3p Alusx1 [4175264 72 532 1223	54177574 123 91 498	541790511 86 0 520e * * 515-30 520e 520e 520e 8 5515-50 515-30 520e 8 5515-50 515-30 5515-3005-3005-30 5005-3005-5515-3005-5005-5005-5005-5005-	54183277 83 0 519e* 519e * *	54185499 86 37 520£ * *	54188345 82 95 515-5p 515-3p * AluSz, AluSz / P / P	24109002 60 03 219C-3P 219C-3P * AluSz 54191821 86 12 1283 * AluSz	54194219 84 552 520a-5p 520a-3p * AluSx3	5419/729 82 53 526b 226b <sup>*</sup> * AluSX 5419877 80 95 519b-5p 519b-3p * AluSX	54200871 84 45 525-5p 525-3p *	54201725 86 117 523* 523 * *	54203555 86 12/ 5181* 5181 * * * 54204541 60 0 520b * *	54206073 82 218 518b * *	54209590 84 23 526a * *	J#11/33         60         20         30/2005         30/00-35         5           54212089         100         87         518c*         518c         *         *	54214342 86 24 524-5p 524-3p AluSg	54215608 86 410 51/* 51/a * * AluSX1	54219934 86 0 521 * * *	54223436 66 410 517* 517b * AluSp AluSx1, AluY 7	54225509 89 176 5209 *	542287800 84 626 516b* 516b* * * 242387800 84 626 516b* * * 242302401 64 37 526a * 242302401 64 37 526a	54233179 87 115 518e* 518e * *	54234344 84 82 518-5p 518-3p *	542382177 86 0 518d-5p 518d-3p * 54240188 89 388 516b* * *	54242673 86 82 518-5p 518-3p *	54244661 94 0 517* 517c * * AluSx, AluSx	54250471 106 NA Probable Psudogene * AluY, AluSx	542519'6 86 0 522* 521 * * 522 * * 522 * * 522 * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * 522 * * * *	54255735 84 66 519a* 519a * * * *	2425/358 84 10 52/ 54260084 89 37 516a-5p 516a-3p *	54261572 86 12 1283 * AluSg	54264476 89 0 516a-5p 516a-3p * AluSx 5126a-5p 516a-3p * AluSx 54265684 86 12 519a * *	426 110286 3831 2340 2430 2540 2520 2520 2520 2520 2520 2520 252	#03114 9247 934 17814 935 1491 804 801 4693 807 801 801 8770 8770	0 1349555 1341 1314 1314 1314 1314 1314 1314 1
u.	Position 54168188 - mgth 54168188 - mgth 54168188 - mgth 54270488 - mature 54270488 - mature 54280 - mature 54280 - mature 54280 - mature 54280 - mature 54280 - 54280 - 542800 - 54280 - 54280 - 54280 - 542800 - 54280 - 542800 -	art Brd Le Stram 3'arm RM d Direction (<1.5KDp) 32 6 7 42 6 4 5 2 2 6 4 1 1 2 2 5 6 7 3 3 2 6 7 1 1 2 2 5 6 4 4 5 6 6 4 4 5 6 7 1 1 2 5 7 5 6 7 1 2 5 7 1 2 5	:9933 54170016 83 634 512-5p 512-3p 4 Aluse	22411 [54172508 97 634 512-5p 512-3p AluSx1 [7] [7] [7] [7] [7] [7] [7] [7] [7] [7]	7451 54177574 123 91 498	8965 54179051 86 0 520e * * * 2257 54182339 82 95 515-5b 515-3b * * *	31314 45183277 83 0 519e* 519e * *	5541354185499 86 37 520£ * *	88263 54188345 82 95 515-5p 515-3p * Alusz, Alusz / / / / / / / / / / / / / / / / / / /	29/23 24109009 00 03 219C-5P 219C-5P * AluSz * 1735 54191821 86 12 1283	14135 54194219 84 552 520a-5p 520a-3p * AluSx3 * AluSx3	7/64/154197729 82 53 5200 5260 7 ALUSX 8467 54198547 80 95 5195-55 5195-35 * ALUSX	0787 54200871 84 45 525-5p 525-3p *	10505 54201725 86 117 523* 523 * *	13269 54203355 86 127 5181* 5181 * * * 4481 54204541 60 0 520b * * *	15991 54206073 82 218 518b * *	91506124203590 84 23 5268 * * * * * * * * * * * * * * * * * * *		.4256 54214342 86 24 524-5p 524-3p AluSg	.5522 54215608 86 410 51/* 51/a * * Alusxi 6601 54216688 87 57 519d * *	3848 54219934 86 0 521 * * *	23501 54223436 86 49 5200-5p 5200-5p AluSp AluSp 4330 54224396 66 410 517* 517b * AluY AluY AluY	5420         54225509         89         176         520g         *	2:8596 54228780 84 626 516b 516b* * * 0176 5422240 64 37 526a * *	3092 54233179 87 115 518e* 518e * *	4260 54234344 84 82 518-5p 518-3p *	88131 54238217 86 0 518d-5p 518d-3p * 0099 54240188 89 388 516b 516b* *	2587 54242673 86 82 518-5p 518-3p *	(4567)54244661 94 0 517* 517c * * AluSx, AluSx, AluSx 75661542445853 87 22 520h * *	0365555556471 106 NA Probable Psudogene * AluY, AluSx	0.1800 5425.1976 86 0 0 521 * * * * * * * * * * * * * * * * * * *	55631 54257735 84 66 519a* 519a * * * *	0.1212 54257356 84 10 521 9995 54260084 89 37 516a-5p 516a-3p *	1486 54261572 86 12 1283 * AluSg	i4387154264476 89 0 516a-5p 516a-3p * AluSx 5598 54265684 86 12 519a * *	426 1120286 3831 2229 2740 2747 2229 2229 2229 2229 2229 2229 2229	403114 3247 9247 93247 633 5041 4693 801 4693 801 4693 801 801 #	# of Target Alux 12107 1214 1214 43 1207 43 12180 43 12180 43 12180 43
	Possition 54168188- Matture alicorobust 54168188- Matture alicorobust pol III riven stream stream abarded stream abarded stream abarded ababarded abar	Start End I Direction (<1.5Kbp) 3267426664 572787 332679286664 3228712228656664	5416933354170016 83 634 512-5p 512-3p Alusg	54172411 54172508 97 634 512-5p 512-3p Alusx1	54177451 54177574 123 91 498	54178965         54178965         54178965         54178051         86         0         520e         *	54183134 54183277 83 0 519e* 519e * *	54185413 54185499 86 37 520£ * *	54188303 54188345 82 05 575-559 515-329 * Alusz, Alusz	7419125 51419126 86 12 1282 5125-55 * AluSz / 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	54194135 54194219 84 552 520a-5p 520a-3p * AluSx3	54196467 54198547 80 95 5195-55 5195-35 * Aluex	54200787 54200871 84 45 525-5p 525-3p *	5420165 54201725 86 117 523* 523 * * *	5420346815420454160 0 520b * *	54205991 54206073 82 218 518b * *	54200506 54209590 84 23 5528 * * *	54211969 54212089 100 87 518c* 518c * 518c * *	54214256 54214342 86 24 524-5p 524-3p AluSg	54215522 54215608 86 410 51/* 51/a * * Alusxi 54216601 54216688 87 57 519d * *	54219848         54219934         86         0         521         *	54223550 54223436 56 410 517* 517b * AluSp AluSx1, AluY 6 410 517* 517b * AluSy	54225420         54225509         89         176         520g         *	54228696154228780 84 626 516b 516b* * * 54228696154228240 64 37 526a * *	54233092 54233179 87 115 518e* 518e * *	54234260 54234344 84 82 518-5p 518-3p *	54238131 54238217 86 0 518d-5p 518d-3p * 54240099 54240198 89 388 516b 516b * *	54242587 54242673 86 82 518-5p 518-3p *	54244567 54244661 94 0 517* 517c * * AluSx, AluSx, AluSx 7 22 5424566 54245853 87 22 520h * *	54250365 54250471 106 NA Probable Psudogene * AluY, AluSx	54251899 54251976 86 0 521 * * 531 54254551 86 69 522* 522 * *	54255605 54255735 84 66 519a* 519a * * * *	2425/2/2/2 5425/356 84 10 52/ 54259995 54260084 89 37 516a-5p 516a-3p *	54261486 54261572 86 12 1283 * AluS9	54264387 54264476 89 0 516a-5p 516a-3p * AluSx 54265684 86 12 519a * *	426 426 110286 2831 2832 26430 27430 2223 22230 22330 223000 223000 223000 223000 223000 223000 223000 223000 223000 223000 223000 223000 223000 223000 223000 223000 223000 223000 22300000000	#03114 63247 93247 93247 93547 60416 801 80176 801 801 801 801 801 801 801 801 801 801	# of Target Alux 1996555 209 124 124 124 124 1917 797 797 797
ша :	Position Position 54168188- Matture 54168188- Matture 54270488 PACT FAPA	ame Start End Lervis 3 Dph Direction (<1.5Kbp) 267769664 ame Start End 1.072876769664	2-1 5416993354170016 83 634 512-5p 512-3p Aluse	2-2 54172411 54172508 97 634 512-5p 512-3p AluSX1	3 54177451 54177574 123 91 498	De 54178965 54179051 86 0 520e * * :-1 54182257 54182339 82 95 515-55 515-35 * *	1 54183194 54183277 83 0 519e* 519e * *	♪		ec attors/zjartasous do asjarta-apjare-	Date         54194135         54194219         84         552         520a-5p         520a-3p         *         AluSx3	20 5419/641 5419/729 82 53 5265 5265* * AluSX 30 54198467 54198547 80 95 5195-5p 5195-3p * AluSX	5 54200787 54200871 84 45 525-5p 525-3p * *	3 52201639 54201725 86 117 523* 523 * * *	21 54204289154204359 86 12/ 3181* 3181 * *	3b 54205991 54206073 82 218 518b * * *	a - 1 54209505 54209550 84 23 528a * * * * * * * * * * * * * * * * * * *		4 54214256 54214342 86 24 524-5p 524-5p 524-5p AluSa	/a 54215522 54215608 86 410 51/* 51/a * * AluSXI Autor	-2 54219848 55219934 86 0 521 * * 521	0d 5#223359 [5#2234366 86 49 52003-5p 52003-5p Alusp Alusp Alusp 10 54224330 [54224336 66 410 517* 517b * Alusg6, Alusx1, Alur	9         54225420         59         176         5209         89         176	5b-2 54228696 54228780  84  626 516b 516b* * * ia-2 54230176 54230240 64 37 526a * *	3e 54233092 54233179 87 115 518e* 518e * *	3a-1 54234260 54234344 84 82 518-5p 518-3p *	ad 54238131542428177 86 015184-59 5518-59 *	13a-2 54242587 54242673 86 82 518-5p 518-3p *	7c 5424456754244661 94 0 517* 517c * * Alusx, Alusx, Alusx	17 54250365 54250471 106 NA Probable Psudogene * AluY, AluSx	L-1 54251800 54251976 86 0 22* 521 * * * * * * * * * * * * * * * * * * *	a-1 54256561 54255735 84 66 519a* 519a * * * *	/ 2425/2/2/2425/3258 84 10 52/ ta=1 54259995 5426084 89 37 516a-5p 516a-3p *	33-2 54261486 54261572 86 12 1283 * Alusg	5a-2 54264387 54264476 89 0 516a-5p 516a-3p * AluSx 1a-2 54265598 54265684 86 12 519a * *	426 110286 3831 2347 2430 2747 22909 2743 27909 279000 27900 279000 279000 279000 279000 279000 279000 279000 2790000000000	403114 9247 9247 9247 17814 633 5041 801 4491 801 4491 801 8770 8770 8770 8770 8770 8770 8770	# of Target Alux 1240 239 991855 797 797 797 797 797 797
υα	Position Bosition 54168188- 54168188- 54168188- 54168188- 54270488 54270488 54270488 54270488 54270488 54270488 54270488 542666 54270488 7AAACTC 7AAAACTC 7AAACTC 7AAACTC 7AAACTC 7AAAACTC 7AAACTC 7AAACTC 7AAACTC 7AAACTC 7AAACTC 7AAACTC 7AAACTC 7AAACTC 7AAACTC 7AAACTC 7AAACTC 7AAACTC 7AAACTC 7AAACTC 7AAACTC 7AAACTC 7AAAACTC 7AAAACTC 7AAAACTC 7AAAACTC 7AAAACTC 7AAAAACTC 7AAAAACTC 7AAAACTC 7AAAAACTC 7AAAAAAAAAA	NA Name Start End Levin 3'arm NA d D Up Direction (41.5KDp) 32 647 69 664 37 NA Name Start End Lizer 3'arm 3'arm NA 66 664	~-512-1 54169933 54170016 83 634 512-5p 512-3p Alusg	r-512-2 54172411 54172508 97 634 512-5p 512-3p AluSx1	~498 54177451 54177574 123 91 498	505 5417905 5417901 86 0 1	519e 541812777 83 0 519e* 519e * *	c-520f 54185413 54185499 86 37 520f * * *	515-2 541882534188345 82 95 515-59 5.515-30 * AluSz, AluSz AluSz 54188253 5.515-59 5.515-30 5.555-5.515-30 5.515-30 5.515-30 5.515-30 5.515-30 5.515-30 5.515-30 5.515-30 5.515-30 5.515-30 5.515-30 5.515-30 5.515-30 5.515-5.515-30 5.555-30 5.5555-30 5.5555-30 5.5555-50-5555-30 5.5555-50-50	モラコンロ フリモッアメニシリン のも のうコンワーンシア - 10122 - 1122	c-520a 54194135 54194219 84 552 520a-5p 520a-3p * AluSx3 * AluSx3	5.519 541196467 54195477 80 95 5195-50 5200* * Alusx	<-525 5420078754200871 84 45 525-5p 525-3p +	523 54201639 54201725 86 117 523* 523 * * *	5-DE 5420/25915420/5535 86 12/1 518E 518E * * *	<-518b 54205991 54206073 82 218 518b * * *	526a-154209506 54209590 84 23 556a *	でつるとし、フォェルマリアオエルバンラ、ロッ・ローコンロンビーンジョンビーンジョン・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・・	e-524 54214256 54214342 86 24 524-55 524-39 AluSq	r-5116 54216601 54216688 87 57 5194 * * AluSAL	e-521-2 54219648 54219934 86 0 521 * * *	r=5.547b 54224330 54224396 66 410 574 517b * Atus5 Atus5 Atus7, Atu		e-516b-2 54228696 54228780 84 626 516b 516b* * * *556a-2 542280240 64 37 526a 516b * *	518e 54233092 54233179 87 115 518e* 518e * * *    518e 54233092 54233179 87 115 518e* 518e * *    518e * 518e	e-518a-1 54234260 54234344 84 82 518-5p 518-3p *	5-5180 1542/38131 86 0 5184-55 1544-35 * * * * * * * * * * * * * * * * * * *	~~518a-2 54242587 54242673 86 82 518-5p 518-3p * *	e-517c 54244567 54244661 94 0 517* 517c * * * AluSx, AluSx AluSx -2-520h 54245766 54245853 87 22 520h * *	-Down7 5425036554250471 106 NA Probable Psudogene * AluY, AluSx	e-521-1 54251899 54251896 86 0 0 522* 522 * * * * * * * * * * * * * * *	×-519a-1 54256551 54257735 84 66 519a* 519a * 4 * *	c-52/1 25425/2/2 5425/358 84 1U 52/ -516a-1 54259995 54260084 89 37 516a-5p 516a-3p *	C-1283-2 54261486 54261572 86 12 1283 * AluS9	e-516a-2 54264387 54264476 89 0 516a-5p 516a-3p * AluSx 2-512a-2 54265598 54265684 86 12 519a * * *	436 110286 3831 2340 2430 2540 2520 2520 2520 2520 2520 2520 252	403114 6247 9247 93247 1491 638 1491 4693 801 4693 801 4693 801 4693 801 4693 801 4693 801 4693 801 801 801 801 801 801 801 801	m of Target Alux 1341 2180 43 7180 43 7180 43 7180 43 7180 739 8800 739 8800 739 7184 72180 739
	Possition 54169189 - Matture 54169188 - Matture 54168188 - Matture 54168188 - Matture 54169188 - Matture 54169188 - Matture 54169188 - Matture 54169188 - Matture 54169188 - Matture 54169188 - Matture 5416918 -	croRNA Name Start End Let 3'D D' D'rection (<1.5KDp) 3267426 64 3 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	a-mir-512-1 54169933 54170016 83 634 512-5p 512-3p Alusg	a-mir-512-2 5124115412508 97 634 512-55 512-35 Allocation Allocation and Allocati	a-mir-498 54177451 54177574 123 91 498	a-mir-5120e 5419965154119501 86 0 520e * * *	a mir - 519e 5418312070 83 0 519e* 519e * *	a-mir-520f 5418541354185499 86 37 520f * * *		a-mir-2:283-1 54191735 54191821 86 12 1282-29 219C-29 ** AluSz	a-mir-520a 54194135 54194219 84 552 520a-59 520a-39 * AluSk3	a-mir-2260 941964(7184197/29) a2 54 2260 32609 8 A Aluss	a-mir-525 54200871 84 45 525-5p 525-3p *	a-mir-533 54201639 54201725 86 117 533* 523 * *	a-mirz-2181 542014281 54201435 80 1271 518F* 3181 * *	a-mir-518b 54205991 54206073 82 218 518b + + +	a-mir=558a-1 54209506 84 23 558a * * * * * * * * * * * * * * * * * * *	a-mint-2300 2942010 09 24212099 100 249 2400-29 * * *	a-mir-524 54214256 54214342 86 24 524-5p 524-5p Alusg	a-mir-1914 54:1552/154215608 66 410 51/* 51/a * * Aluški p.mir-1916 154216601 154216688 87 57	a-mir-521-2 54219948 54219934 86 0 521 * * *	a-mir-2104 3-221330194231345 96 4915-040-59 51246-59 1 Aubs j.mir-2175 64224330 66 410 517* 6214* Autvg6, Aluxi, Auv	a-mir-520g 54225420 54225509 89 176 520g *	= mir-516b-2154220606154287870 84 626 516b 516b* * * * * * * * * * * * * * * * * * *	a-mir-518e 54233092 54233179 87 115 518e* 518e * *	a-mir-518a-1 54234260 54234344 84 82 518-5p 518-3p *	- mirr-516-1 542340313 542340217 86 0 5184-55 5184-32 *	a-mir-518a-2 54242587 54242673 86 82 518-5p 518-3p *	a-mir-517c 54244567154244661 94 0 517* 517c * * Alusx, Alusx Alusx Alusx	a-pd-Down7 61250355154250471 106/M Probable Psudogene * AluY, AluSx 75250355154250471 016/M Part 2000 100 100 100 100 100 100 100 100 10	a-mir-521-1 54251990 54251976 86 0 0 521 * * * 1 * * * * * * * * * * * * * * *	a-mir-5j9a-1 54255651 54255735 84 66 519a* 519a * * *	a-mirr-162-27 99-25/12/21/21/21/21/21/21/21/21/21/21/21/21/	a-mir-1283-2 54261486 54261572 86 12 1283 * AluSg	a-mir-516a-2 54264387 54264476 89 0 516a-5p 516a-3p * AluSx i-mir-519a-2 54265598 54265684 86 12 519a * *	426 426 1120286 3831 23430 5279 5279 5279 5279 5279 5279 5279 5279	403114 6247 9247 17814 1491 635 1491 605 801 4633 801 4633 8770 801 # 13129 # 13129 # 13129 # 13219 # # 13219 # 1 1 1 1 1 1 1 1 1 1 1 1 1	# of Target Alux 1340 239 43 7180 43 7180 43 7180 43 7180

Table 2 Summary of Human C19MC Features

# References

- Arcot, S. S., Z. Wang, et al. (1995). "Alu repeats: a source for the genesis of primate microsatellites." <u>Genomics</u> **29**(1): 136-144.
- Bartel, D. P. (2009). "MicroRNAs: target recognition and regulatory functions." <u>Cell</u> **136**(2): 215-233.
- Bartel, D. P. and C. Z. Chen (2004). "Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs." <u>Nat Rev Genet</u> **5**(5): 396-400.
- Batzer, M. A. and P. L. Deininger (2002). "Alu repeats and human genomic diversity." <u>Nat</u> <u>Rev Genet</u> **3**(5): 370-379.
- Belancio, V. P., A. M. Roy-Engel, et al. (2008). "The impact of multiple splice sites in human L1 elements." <u>Gene</u> **411**(1-2): 38-45.
- Bennett, E. A., H. Keller, et al. (2008). "Active Alu retrotransposons in the human genome." <u>Genome Res</u> **18**(12): 1875-1883.
- Bentwich, I., A. Avniel, et al. (2005). "Identification of hundreds of conserved and nonconserved human microRNAs." <u>Nat Genet</u> **37**(7): 766-770.
- Borchert, G. M., W. Lanier, et al. (2006). "RNA polymerase III transcribes human microRNAs." <u>Nat Struct Mol Biol</u> **13**(12): 1097-1101.
- Bortolin-Cavaille, M. L., M. Dance, et al. (2009). "C19MC microRNAs are processed from introns of large Pol-II, non-protein-coding transcripts." <u>Nucleic Acids Res</u> 37(10): 3464-3473.
- Brennecke, J., C. D. Malone, et al. (2008). "An epigenetic role for maternally inherited piRNAs in transposon silencing." <u>Science</u> **322**(5906): 1387-1392.
- Brennecke, J., A. Stark, et al. (2005). "Principles of microRNA-target recognition." <u>PLoS</u> <u>Biol</u> **3**(3): e85.
- Cai, X., C. H. Hagedorn, et al. (2004). "Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs." <u>RNA</u> 10(12): 1957-1966.

- Comeaux, M. S., A. M. Roy-Engel, et al. (2009). "Diverse cis factors controlling Alu retrotransposition: what causes Alu elements to die?" <u>Genome Res</u> **19**(4): 545-555.
- Cordaux, R. and M. A. Batzer (2009). "The impact of retrotransposons on human genome evolution." <u>Nat Rev Genet</u> **10**(10): 691-703.
- Costa, F. F. (2008). "Non-coding RNAs, epigenetics and complexity." Gene 410(1): 9-17.
- Dewannieux, M., C. Esnault, et al. (2003). "LINE-mediated retrotransposition of marked Alu sequences." <u>Nat Genet</u> **35**(1): 41-48.
- Esnault, C., J. Maestre, et al. (2000). "Human LINE retrotransposons generate processed pseudogenes." <u>Nat Genet</u> **24**(4): 363-367.
- Faulkner, G. J., Y. Kimura, et al. (2009). "The regulated retrotransposon transcriptome of mammalian cells." <u>Nat Genet</u> **41**(5): 563-571.
- Feschotte, C. (2008). "Transposable elements and the evolution of regulatory networks." <u>Nat Rev Genet</u> **9**(5): 397-405.
- Ghildiyal, M. and P. D. Zamore (2009). "Small silencing RNAs: an expanding universe." <u>Nat</u> <u>Rev Genet</u> **10**(2): 94-108.
- Glazov, E. A., S. McWilliam, et al. (2008). "Origin, evolution, and biological role of miRNA cluster in DLK-DIO3 genomic region in placental mammals." <u>Mol Biol Evol</u> 25(5): 939-948.
- Gu, T. J., X. Yi, et al. (2009). "Alu-directed transcriptional regulation of some novel miRNAs." <u>BMC Genomics</u> 10: 563.
- Halic, M. and D. Moazed (2009). "Transposon silencing by piRNAs." <u>Cell</u> **138**(6): 1058-1060.
- Han, J. S. and J. D. Boeke (2005). "LINE-1 retrotransposons: modulators of quantity and quality of mammalian gene expression?" <u>Bioessays</u> **27**(8): 775-784.
- Jurka, J., V. V. Kapitonov, et al. (2005). "Repbase Update, a database of eukaryotic repetitive elements." <u>Cytogenet Genome Res</u> **110**(1-4): 462-467.
- Jurka, J. and C. Pethiyagoda (1995). "Simple repetitive DNA sequences from primates:

compilation and analysis." <u>I Mol Evol</u> **40**(2): 120-126.

- Kelkar, Y. D., S. Tyekucheva, et al. (2008). "The genome-wide determinants of human and chimpanzee microsatellite evolution." <u>Genome Res</u> **18**(1): 30-38.
- Kertesz, M., N. Iovino, et al. (2007). "The role of site accessibility in microRNA target recognition." <u>Nat Genet</u> **39**(10): 1278-1284.
- Krek, A., D. Grun, et al. (2005). "Combinatorial microRNA target predictions." <u>Nat Genet</u> **37**(5): 495-500.
- Lagos-Quintana, M., R. Rauhut, et al. (2003). "New microRNAs from mouse and human." <u>RNA</u> **9**(2): 175-179.
- Lander, E. S., L. M. Linton, et al. (2001). "Initial sequencing and analysis of the human genome." <u>Nature</u> **409**(6822): 860-921.
- Lee, Y., M. Kim, et al. (2004). "MicroRNA genes are transcribed by RNA polymerase II." <u>EMBO J</u> 23(20): 4051-4060.
- Li, J., Y. Liu, et al. (2010). "Evolution of an X-linked primate-specific micro RNA cluster." <u>Mol Biol Evol</u> **27**(3): 671-683.

896

- Malone, C. D. and G. J. Hannon (2009). "Small RNAs as guardians of the genome." <u>Cell</u> **136**(4): 656-668.
- Mills, R. E., E. A. Bennett, et al. (2007). "Which transposable elements are active in the human genome?" <u>Trends Genet</u> 23(4): 183-191.
- Morin, R. D., M. D. O'Connor, et al. (2008). "Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells." <u>Genome</u> <u>Res</u> 18(4): 610-621.
- Nakamoto, M., P. Jin, et al. (2005). "Physiological identification of human transcripts translationally regulated by a specific microRNA." <u>Hum Mol Genet</u> **14**(24): 3813-3821.
- Neilson, J. R. and P. A. Sharp (2008). "Small RNA regulators of gene expression." <u>Cell</u> **134**(6): 899-902.

Ostertag, E. M. and H. H. Kazazian, Jr. (2001). "Biology of mammalian L1
retrotransposons." <u>Annu Rev Genet</u> **35**: 501-538.

- Polak, P. and E. Domany (2006). "Alu elements contain many binding sites for transcription factors and may play a role in regulation of developmental processes." <u>BMC Genomics</u> **7**: 133.
- Price, A. L., E. Eskin, et al. (2004). "Whole-genome analysis of Alu repeat elements reveals complex evolutionary history." <u>Genome Res</u> **14**(11): 2245-2252.
- Rhead, B., D. Karolchik, et al. (2010). "The UCSC Genome Browser database: update 2010." <u>Nucleic Acids Res</u> **38**(Database issue): D613-619.
- Shankar, R., D. Grover, et al. (2004). "Evolution and distribution of RNA polymerase II regulatory sites from RNA polymerase III dependant mobile Alu elements." <u>BMC Evol Biol</u> 4(1): 37.
- Smalheiser, N. R. and V. I. Torvik (2005). "Mammalian microRNAs derived from genomic repeats." <u>Trends Genet</u> **21**(6): 322-326.
- Smit, A., Hubley, R & Green, P. (1996-2010). "RepeatMasker Open-3.0."
- Swergold, G. D. (1990). "Identification, characterization, and cell specificity of a human LINE-1 promoter." <u>Mol Cell Biol</u> **10**(12): 6718-6729.
- Tamura, K., J. Dudley, et al. (2007). "MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0." <u>Mol Biol Evol</u> **24**(8): 1596-1599.
- Tsai, K. W., H. W. Kao, et al. (2009). "Epigenetic control of the expression of a primate-specific microRNA cluster in human cancer cells." <u>Epigenetics</u> **4**(8): 587-592.
- Weiner, A. M. (2002). "SINEs and LINEs: the art of biting the hand that feeds you." <u>Curr</u> <u>Opin Cell Biol</u> **14**(3): 343-350.
- Xie, X., J. Lu, et al. (2005). "Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals." <u>Nature</u> 434(7031): 338-345.
- Zhang, R., Y. Peng, et al. (2007). "Rapid evolution of an X-linked microRNA cluster in primates." <u>Genome Res</u> **17**(5): 612-617.

Zhang, R., Y. Q. Wang, et al. (2008). "Molecular evolution of a primate-specific microRNA family." <u>Mol Biol Evol</u> **25**(7): 1493-1502.

## **Figures which are Adapted from References**

- Abeloff, M. D. (2008). <u>Abeloff's clinical oncology</u>. Philadelphia, PA, Churchill Livingstone/Elsevier.
- Cordaux, R. and M. A. Batzer (2009). "The impact of retrotransposons on human genome evolution." <u>Nat Rev Genet</u> **10**(10): 691-703.
- Price, A. L., E. Eskin, et al. (2004). "Whole-genome analysis of Alu repeat elements reveals complex evolutionary history." <u>Genome Res</u> **14**(11): 2245-2252.



## Appendix 1 Annotation of Human C19MC



Green: Alu elements which are located on the same direction with microRNAs; Red: microRNAs; Blue: Alu elements which are located on the opposite direction with microRNAs; Gray: exons. Exons are predicted in this study; microRNAs are followed the annotation of miRBase; Alu elements are followed the annotation of RepeatMasker.





Green: Alu elements which are located on the same direction with microRNAs; Red: microRNAs; Blue: Alu elements which are located on the opposite direction with microRNAs; Gray: exons. Exons are predicted and defined in this study; microRNAs are defined in this study; Alu elements are followed the annotation of RepeatMasker.