Genes & Genomics (2012) 34: 517-528
DOI 10.1007/s13258-012-0058-0

RESEARCH ARTICLE

Intragenic long interspersed element-1 sequences promote promoter hypermethylation in lung adenocarcinoma, multiple myeloma and prostate cancer

Suphakit Khowutthitham · Chumpol Ngamphiw · Wachiraporn Wanichnopparat · Kulachanya Suwanwongse · Sissades Tongsima · Chatchawit Aporntewan · Apiwat Mutirangura

Received: 26 March 2012 / Accepted: 14 May 2012 / Published online: 31 August 2012

Abstract
In cancers, although the methylation of long interspersed element-1 sequences (LINE-1s) and tumor suppressor gene promoters are modified in the opposite direction, LINE-1 hypomethylation and promoter hypermethylation of some loci are directly associated. During carcinogenesis, the reduction in LINE-1 methylation occurs. Intragenic LINE-1s produces antisense RNA in introns and reduces mRNA transcription levels. Several antisense RNAs have been reported to mediate methylation of the associated CpG islands. Here we compared genome-wide promoter methylation and expression profiles of LINE-1-hypomethylated malignancies, reported in the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo), including lung adenocarcinoma, multiple myeloma and prostate cancer. Then we analysed a microarray experiment if promoters of a set of genes containing LINE-1s or *Alu* are commonly methylated. Finally, the differences in structural characteristics of LINE-1s were compared between LINE-1 groups. Here we found that genes that contained LINE-1s were frequently repressed ($p < 0.01$) and possessed promoter hypermethylation ($p < 1.0E-4$). The expression levels of genes containing LINE-1s with promoter hypermethylation were the lowest. Finally, the genomic distributions of gene-repressing LINE-1s and promoter-hypermethylating LINE-1s were neither co-segregated nor randomly segregated. In conclusion, cancer-associated intragenic LINE-1 epigenetic change promotes promoter hypermethylation and represses gene expression. These two mechanisms are independently influenced by genomic locations but synergistically down-regulate genes.

Keywords Antisense RNA; Cancer epigenomics; Gene promoter hypermethylation; Global hypomethylation; Intragenic LINE-1; Long interspersed element-1; LINE-1 hypomethylation

SK, CN and WW contributed equally to this work.

S. Khowutthitham · C. Ngamphiw
Inter-Department Program of Biomedical Sciences, Faculty of Graduate School, Chulalongkorn University, Bangkok 10330, Thailand

C. Ngamphiw · S. Tongsima
Genome Institute, National Center for Genetic Engineering and Biotechnology, 113 Thailand Science Park, Paholyothin Road, Klong 1, Klong Luang, Pathum Thani, 12120, Thailand

W. Wanichnopparat · K. Suwanwongse
Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand

C. Aporntewan
Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Rama IV, Bangkok, 10330, Thailand

A. Mutirangura (✉)
Department of Anatomy, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand
e-mail: mapiwat@chula.ac.th

S. Tongsima · C. Aporntewan · A. Mutirangura
Center of Excellence in Molecular Genetics of Cancer and Human Diseases, Chulalongkorn University, Bangkok 10330, Thailand

Introduction

Two common DNA methylation aberrations in cancer are promoter hypermethylation of tumor suppressor genes and genome-wide hypomethylation (Feinberg and Tycko, 2004; Brait and Sidransky, 2011). Different mechanisms have been proposed to mediate these phenomena because the modification of DNA methylation in these two events occurs in different directions (Ehrlich, 2002). Interestingly, many studies have evaluated the correlation between these two epigenetic

modifications. Several recent reports have demonstrated a correlation between decreased methylation levels of long interspersed element-1 sequences (LINE-1s or L1s) and hypermethylation of promoters at some loci (Florl et al., 2004; Cho et al., 2007; Choi et al., 2007; Ogino et al., 2008; Yamamoto et al., 2008; Park et al., 2009; Baba et al., 2010; Daskalos et al., 2011; Poage et al., 2011). This correlation suggests the existence of a cause-and-effect relationship between these two epigenomic events. Surprisingly, no direct correlation between global or LINE-1 hypomethylation and promoter hypermethylation at several loci has also been frequently reported (Ehrlich et al., 2002; Florl et al., 2004; Ehrlich et al., 2006; Cho et al., 2007; Iacopetta et al., 2009; Kim et al., 2009; Park et al., 2009; Igarashi et al., 2010; Trankenschuh et al., 2010; Woloszynska-Read et al., 2011). Here, we hypothesised a heterogeneous nature of promoter hypermethylation and proposed an underlying mechanism in which some hypermethylated loci possess a direct relationship with LINE-1 hypomethylation.

Decreases in LINE-1 methylation occur early, commonly and progressively during multistep carcinogenesis (Chalitchagorn et al., 2004; Kitkumthorn and Mutirangura, 2011). In cancer, direct correlations between the methylation levels of genome-wide LINE-1s and each LINE-1 locus and between different interspersed repetitive sequence classes have been reported (Phokaew et al., 2008; Iramaneerat K et al., 2011). Therefore, a genome-wide hypomethylation mechanism in cancer appears to be a generalised process that reduces genome-wide methylation (Phokaew et al., 2008; Kitkumthorn and Mutirangura, 2011). LINE-1s are interspersed repetitive sequences that are widely distributed across the human genome. LINE-1s played an important role in evolution of human genome (Cordaux and Batzer, 2009) and possessed physiologic function of the cells (Belancio et al., 2006; Aporntewan et al., 2011). Most LINE-1s are truncated. Approximately 10,000 LINE-1s contain a 5' UTR, which include CpG islands and are frequently methylated (Penzkofer et al., 2005). LINE-1s are classified into 2 categories by location, i.e., inside and outside of genes. Interestingly, intragenic LINE-1s are more conserved than intergenic LINE-1s, and CpG dinucleotides and sequence determine transcriptional activity (Aporntewan et al., 2011). These conserved characteristics of intragenic LINE-1s imply the biological functions of the sequences (Aporntewan et al., 2011).

Both promoter hypermethylation and LINE-1 hypomethylation down-regulate gene expression. Promoter hypermethylation inhibits the transcription initiation complex and, consequently, inhibits gene expression (Herman, 2005). Although genomic instability is the most commonly known consequence of global hypomethylation (Chen et al., 1998; Pornthanakasem et al., 2008; Kongruttanachok et al., 2010), LINE-1 hypomethylation also perturbs gene expression (Aporntewan et al., 2011; Kitkumthorn and Mutirangura, 2011). Previously, we discovered the mechanisms for this regulation by performing a correlation genome-wide expression array from cancer samples using Connection Up- and Down-Regulation Expression Analysis of Microarrays (CU-DREAM) and CU-DREAM eXtension (CU-DREAMX) programs (Aporntewan and Mutirangura, 2011) with cancer cells, demethylated normal cells and AGO2 knocked-down cells (Aporntewan et al., 2011). We observed preferential down-regulation of genes that contained LINE-1s in cancer and demethylated normal cells and up-regulation of gene expression of AGO2-depleted cells. We found that in cancer, hypomethylated LINE-1s are transcribed into RNAs that are compliments to intron sequences. This antisense RNA forms a complex with pre-mRNA and AGO2. Therefore, the mRNA transcripts of genes containing LINE-1s are depleted (Aporntewan et al., 2011). The LINE-1 methylation level of each locus in cancer varies (Phokaew et al., 2008). Therefore, differential gene repression depends on the intragenic LINE-1 hypomethylation levels. Antisense RNA transcripts that mediate methylation of downstream CpG islands have been discovered in several epigenetic regulation events. Well-known examples of downstream CpG islands are promoters of genomic imprinting genes, such as *Igf2r* (Wutz et al., 1997) and the *Xist* gene in X-inactivation (Navarro et al., 2005). Moreover, a transgenic study in a form of alpha-thalassemia has demonstrated that methylation also occurs at non-imprinted autosomal loci in differentiating embryonic stem cells (Tufarelli et al., 2003). In the current study, we performed a genome-wide analysis using the CU-DREAM and CU-DREAMX programs (Aporntewan and Mutirangura, 2011) and found interesting correlations between intragenic LINE-1s and promoter hypermethylation in cancer.

Materials and Methods

Lists of genes containing LINE-1s, genes containing *Alu* and microarrays

LINE-1s that were reported in the L1base (http://l1base.molgen.mpg.de) (Penzkofer et al., 2005) were categorised according to their genomic locations as “intragenic” or “intergenic” based on the NCBI Reference Sequence (RefSeq) annotation (Aporntewan et al., 2011). The list of genes containing LINE-1s was the same as previously reported (Aporntewan et al., 2011). Genes containing *Alu* were available from transpogene database (http://transpogene.tau.ac.il/) (Levy et al., 2008). All microarray data were taken from Gene Expression Omnibus (GEO) (Barrett et al., 2005; Barrett et al., 2009). The GEO series (GSE) methylation arrays were GSE16559 (Christensen et al., 2009), GSE21304 (Walker et al., 2011), and GSE26126 (Kobayashi et al., 2011) and the expression libraries were GSE18842 (Sanchez-Palencia et al., 2011), GSE6691 (Gutierrez et al., 2007), and GSE12378 (Jhavar et al., 2009) for lung adenocarcinoma, multiple myelo-

ma, and prostate cancer, respectively. The GEO datasets for methylation and expression are from separate patients. However, GSE18842 (the lung adenocarcinoma expression library) consisted of both adenocarcinoma and squamous-cell carcinoma samples. Only adenocarcinoma samples were selected in this study. GEO Sample (GSMs) representing test and control samples were listed in Supporting Table S2, and the clinical data for each GSE were reported in Supporting Table S1.

CU-DREAM and CU-DREAMX

The CU-DREAM and CU-DREAMX programs, which are available at http://pioneer.netserv.chula.ac.th/~achatcha/cu-dream/, were executed by testing the gene symbols for each probe, and the data were evaluated using the Student's *t*-test and the Chi-square test. Up- and down-regulation and promoter hypermethylation were determined by comparing intensity of probes by Student's *t*-test. CU-DREAM was used to study the association between two microarray analyses (Fig. 1A). For example, the association between gene expression and promoter methylation is displayed in Table 1. The gene expression microarray libraries and the methylation microarray libraries were selected and analysed by the CU-DREAM program. The results were divided depending on the LINE-1 status using a 2x2 contingency table, and the Chi-square test and Student's *t*-test were used for statistical analyses. The CU-DREAMX program was used to analyse between microarray data and a gene set of interest, such as genes containing LINE-1s (Fig. 1B). The mi-

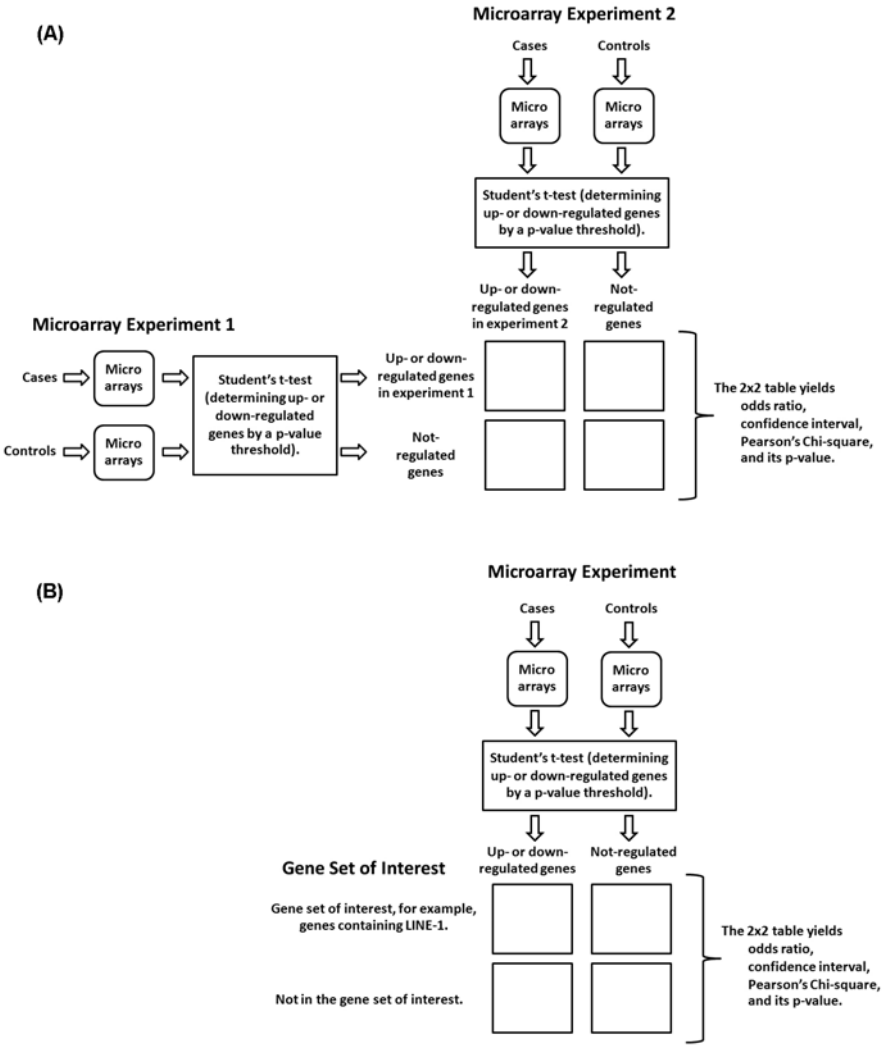


Figure 1. The overall framework of the CU-DREAM and CU-DREAMX programs. (A) The CU-DREAM program is used to statistically identify up and down regulated genes from case/control microarray datasets using Student's t-test. (B) The CU-DREAMX program investigates the correlation between the genes from microarray libraries and some specific feature of these genes. The microarray data were intersected with the gene set with a specific feature, such as the set of genes containing LINE-1s. CU-DREAMX calculates Pearson's Chi-square from the intersected 2×2 contingency table and reports the corresponding odds ratio and p-value.

Table 1. The correlation between promoter methylation and gene down-regulation in lung, prostate and plasma malignancies.

Cancer types	Gene regulation	Met + (%)	Met – (%)	Odds ratio	95% CI	p-value
Lung adenocarcinoma	Down	127 (17.28)	147 (20.00)	1.66	Lower : 1.22	1.16E-3
	Not down	158 (21.50)	303 (41.22)		Upper : 2.25	
	Up	61 (8.30)	144 (19.59)	0.58	Lower : 0.41	1.80E-3
	Not up	224 (30.48)	306 (41.63)		Upper : 0.82	
Multiple myeloma	Down	1,110 (11.39)	1,805 (18.53)	1.05	Lower : 0.96	3.07E-1
	Not down	2,525 (25.92)	4,302 (44.16)		Upper : 1.15	
	Up	908 (9.32)	1,646 (16.90)	0.90	Lower : 0.82	3.22E-2
	Not up	2,727 (27.99)	4,461 (45.79)		Upper : 0.99	
Prostate cancer	Down	179 (1.59)	389 (3.45)	0.87	Lower : 0.73	1.45E-1
	Not down	3,696 (32.75)	7,020 (62.21)		Upper : 1.05	
	Up	287 (2.54)	298 (2.64)	1.91	Lower : 1.62	1.36E-14
	Not up	3,588 (31.80)	7,111 (63.02)		Upper : 2.26	

Met+ represents “hypermethylated”, and Met- represents “unchanged or hypomethylated”.
Down represents “down-regulated genes”, whereas Not down represents “unchanged” or “up-regulated genes”.
Up represents “up-regulated genes”, whereas Not up represents “unchanged” or “down-regulated genes”.

Table 2. Gene expression incidences of the genes containing LINE-1s.

Cancer types	Gene regulation	L1 (%)	No L1 (%)	Odds ratio	95% CI	p-value
Lung adenocarcinoma	Down	470 (2.31)	4,948 (24.29)	1.53	Lower : 1.37	4.78E-13
	Not down	872 (4.28)	14,084 (69.13)		Upper : 1.72	
	Up	296 (1.45)	5,026 (24.67)	0.79	Lower : 0.69	4.53E-4
	Not up	1,046 (5.13)	14,006 (68.74)		Upper : 0.90	
Multiple myeloma	Down	316 (2.42)	3,543 (27.13)	1.27	Lower : 1.10	1.03E-3
	Not down	605 (4.63)	8,594 (65.81)		Upper : 1.46	
	Up	155 (1.19)	766 (5.87)	0.59	Lower : 0.49	4.16E-9
	Not up	3,097 (23.72)	9,040 (69.23)		Upper : 0.71	
Prostate cancer	Down	97 (0.60)	725 (4.48)	1.77	Lower : 1.42	3.14E-7
	Not down	1,082 (6.68)	14,283 (88.24)		Upper : 2.20	
	Up	43 (0.27)	837 (5.17)	0.64	Lower : 0.47	4.89E-3
	Not up	1,136 (7.02)	14,171 (87.55)		Upper : 0.88	

Down represents “down-regulated genes”, whereas Not down represents “unchanged” or “up-regulated genes”.
Up represents “up-regulated genes”, whereas Not up represents “unchanged” or “down-regulated genes”.

croarray data were intersected with the gene set of interest. For example, to evaluate whether intragenic LINE-1s influenced the host gene expression or promoter methylation, the GSEs from the methylation array data and GSEs from the up-/down-regulated gene expression array data were intersected with genes containing LINE-1s using the CU-DREAM program to identify regulated genes at $p<0.05$ (Table 2).

Association studies by Chi-square and Student’s *t*-test

The genes from CU-DREAM and CU-DREAMX data were selected to be analysed for their numbers and expression levels under conditions of interest. To study the regulation patterns of promoter-methylated genes containing LINE-1s compared to genes without LINE-1s, the methylation microarray libraries and the expression microarray libraries were tested using the CU-DREAM program. The genes were then classified by expression, methylation status, and the presence of intragenic LINE-1s and performed Chi-square test. To evaluate the expression level, the genes were classified by CU-DREAMX. The expression values were expressed as Mean1 and Mean2 from the expression library table of each cancer type and analysed by Student’s *t*-test.

Analysis of LINE-1 characteristics

L1base (http://l1base.molgen.mpg.de) (Penzkofer et al., 2005) described structural characteristics of each LINE-1 in detail, including orientation, locations, sequence variations, CpG islands, and subfamilies. Here, we related these 127 characteristics with promoter methylation and down regulation using the Chi-square test and Student’s *t*-test for categorical and non-categorical functionally important features, respectively. The frequency of features was counted according to the number of LINE-1s that contained the tested features. Yates’s continuity correction was also applied to the Chi-square test, which has the expected count of less than 5.

Results

Datasets from Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo) were screened for expression and promoter methylation genome-wide arrays until April 2011; arrays from lung adenocarcinoma, prostate cancer and malignant plasma cells were selected (Barrett et al., 2005; Barrett et al., 2009). The methylation arrays from GEO were GSE16559 (Christensen et al., 2009), GSE21304 (Walker et al., 2011) and GSE26126 (Kobayashi et al., 2011), and the expression libraries were GSE18842 (Sanchez-Palencia et al., 2011), GSE6691 (Gutierrez et al., 2007) and GSE12378 (Jhavar et al., 2009) for lung adenocarcinoma, multiple myeloma and prostate cancer, respectively. These datasets were chosen because lung, prostate and plasma cell malignancies are known to possess LINE-1 hypomethylation (Chalitchagorn et al., 2004; Bollati et al., 2009). Details of the available clinical information are described in the Supporting Table S1, and the GEO samples analysed in the CU-DREAM program are listed in the Supporting Table S2.

The promoter hypermethylation and gene expression status

Promoter hypermethylation is one of the major mechanisms for inhibition of gene expression in cancer. To evaluate the genome-wide contribution of this epigenetic mechanism, genes were classified based on promoter hypermethylation and expression status and counted by the CU-DREAM program (Fig. 1) (Table 1) (Aporntewan and Mutirangura, 2011). Although many reports of an association between promoter hypermethylation and gene repression exist, Table 1 shows that promoter hypermethylation and down-regulation of the gene are not exclusive. Many promoter-hypermethylated genes were not significantly down-regulated (Table 1). Moreover, many down-regulated genes lacked promoter hypermethylation (Table 1). Nevertheless, the prevalence of gene down-regulation was higher when the promoters were hypermethylated in lung adenocarcinoma (odds ratio (OR) = 1.66, 95% confidence interval (CI) = (1.22-2.55)). The prevention of up-regulation was also found in promoter-hypermethylated lung adenocarcinoma genes (OR (95% CI) = 0.58 (0.41-0.82) and in multiple myeloma (OR (95% CI) = 0.90 (0.82-0.99)). Surprisingly, in prostate cancer, many promoter-methylated genes were not down-regulated but were up-regulated (OR (95% CI) = 1.91 (1.62-2.26)).

Intragenic LINE-1s repress genes in lung adenocarcinoma, multiple myeloma and prostate cancer

Our previous study reported that intragenic LINE-1s repress many genes in cancers, including lung and prostate cancers (Aporntewan et al., 2011). Expression arrays were tested using the CU-DREAMX program (Fig. 1) to evaluate the incidence of the expression of genes containing LINE-1s. LINE-1-classified genes were separated into four groups depending on the genetic regulation and the presence of intragenic LINE-1s. The influence of intragenic LINE-1s on gene expression was confirmed in lung and prostate cancers and evaluated for the first time in multiple myeloma. We found that intragenic LINE-1s repressed genes in all three cancers. First, a higher prevalence of down-regulated genes containing LINE-1s was observed in lung adenocarcinoma, multiple myeloma and prostate cancer (OR (95% CI) = 1.53 (1.37-1.72), 1.27 (1.10-1.46), and 1.77 (1.42-2.20), respectively) (Table 2). Moreover, LINE-1s prevented the up-regulation of genes containing LINE-1s in lung adenocarcinoma, multiple myeloma and prostate cancer (OR (95% CI) = 0.79 (0.69-0.90), 0.59 (0.49-0.71) and 0.67 (0.47-0.88), respectively) (Table 2).

Intragenic LINE-1s and methylated genes in lung adenocarcinoma, multiple myeloma and prostate cancer

To evaluate the association between promoter hypermethylation and intragenic LINE-1s, the CU-DREAMX program was used to evaluate GSE-reported methylation arrays for LINE-1s. The prevalence of promoter hypermethylation in lung adenocarcinoma, multiple myeloma and prostate cancer was significantly higher in genes containing LINE-1s (OR (95% CI) = 2.05 (1.31-3.45), 1.29 (1.13-1.48) and 1.28 (1.12-1.48), respectively) (Table 3). These results are the first evidence of a connection between cancer-associated intragenic LINE-1s and promoter hypermethylation.

LINE-1s and promoter hypermethylation synergistically repress genes

Table 1 showed that genes with promoter hypermethylation were regulated in many directions. Here, we compared the regulation patterns of promoter-hypermethylated genes with and without LINE-1s in number (Table 4) and in level (Fig. 2).

Table 3. Promoter methylation statuses of the genes containing LINE-1s in cancers.

Cancer types	L1	Met + (%)	Met – (%)	Odds ratio	95% CI	p-value
Lung adenocarcinoma	L1	34 (4.43)	28 (3.65)	2.05	Lower : 1.21	6.41E-3
	No L1	263 (34.24)	443 (57.68)		Upper : 3.45	
Multiple myeloma	L1	409 (2.83)	563 (3.89)	1.29	Lower : 1.13	1.36E-4
	No L1	4,859 (33.57)	8,645 (59.72)		Upper : 1.48	
Prostate cancer	L1	376 (2.60)	596 (4.12)	1.28	Lower : 1.12	2.67E-4
	No L1	4,453 (30.76)	9,051 (62.52)		Upper : 1.47	

Met+ represents “hypermethylated”, and Met- represents “unchanged” or hypomethylated”.

Table 4. The role of LINE-1s in gene promoter methylation and gene expression.

Cancer types	Condition	L1 (%)	No L1 (%)	Odds ratio	95% CI	p-value
Lung	Met+/Up	6 (0.82)	55 (7.48)	1.20	Lower : 0.50	6.81E-1
adenocarcinoma	The rest	56 (7.62)	618 (84.08)		Upper : 2.92	
	Met+/Down	15 (2.04)	112 (15.24)	1.60	Lower : 0.86	1.32E-1
	The rest	47 (6.39)	561 (76.33)		Upper : 2.96	
Multiple	Met+/Up	50 (0.51)	858 (8.81)	0.71	Lower : 0.52	2.02E-2
myeloma	The rest	674 (6.92)	8,160 (83.76)		Upper : 0.95	
	Met+/Down	123 (1.27)	987 (10.13)	1.67	Lower : 1.36	8.45E-7
	The rest	601 (6.17)	8,031 (82.44)		Upper : 2.04	
Prostate	Met+/Up	17 (0.15)	270 (2.39)	0.76	Lower : 0.46	2.79E-1
cancer	The rest	840 (7.44)	10,157 (90.01)		Upper : 1.25	
	Met+/Down	29 (0.26)	828 (7.34)	2.40	Lower : 1.60	1.18E-5
	The rest	150 (1.33)	10,277 (91.08)		Upper : 3.59	

Met+ represents "hypermethylated".
Down represents "down-regulated genes", and Up represents "up-regulated genes".

The results showed that many repressed genes containing LINE-1s were present in multiple myeloma and prostate cancer. A significant prevention of up-regulation (OR (95% CI) = 0.67 (0.49-0.91)) was found when genes contained LINE-1s in multiple myeloma. In addition, a higher proportion of genes containing LINE-1s in prostate cancer were down-regulated compared with genes without LINE-1s (OR (95% CI) = 2.40 (1.60-3.59)).When the expression levels of genes with promoter hypermethylation and also containing LINE-1s were compared to those genes without LINE-1s, the expression levels of genes containing LINE-1s in cancer were also lower ($p = 0.0002$, <0.0001, and $= 0.0324$ for lung adenocarcinoma, multiple myeloma, and prostate cancer, respectively) (Fig. 2). Among the down-regulated genes containing LINE-1s, genes with promoter hypermethylation were repressed to a greater extent than those without methylation in lung and prostate cancers ($p=0.0121$ and 0.0064, respectively) (Fig. 3). The genes are listed in the Supporting Table S3.

Characteristics of intragenic LINE-1s in repressed or promoter-hypermethylated genes

Using characteristics of LINE-1s listed in L1base (http://l1base.molgen.mpg.de) (Penzkofer et al., 2005), we then evaluated the orientation, locations, sequences, CpG islands, and subfamilies of intragenic LINE-1s that repress genes and promote promoter hypermethylation in the three cancer types. The objective was to explore LINE-1 features that influence gene repression and promoter methylation. The comparison between the two sets of LINE-1 characteristics will imply the relation in mechanisms of gene repression and promoter methylation. Analysis of orientation implied direction of LINE-1 promoters and RNA. Locations may reflect regional mechanism of human genome. Sequence variations usually represent *cis* elements determining transcriptional and retrotransposition activities. CpG islands of LINE-1s may represent potential sequences regulated by DNA methylation. Finally, classification of LINE-1s represents overview of evolution period. One hundred and twenty-seven characteristics of intragenic LINE-1s were analysed. Eighty-five characteristics and 24 chromosome locations were analysed by the Chi-square test, and 18 characteristics were analysed by Student's t-test. All of the data are listed in the Supporting Table S4. Table 5 lists the significantly different characteristics of intragenic LINE-1s in down-regulated genes or genes with promoter hypermethylation compared with the control. Among the 127 characteristics, a few (<7) were significant in each subgroup, and no characteristics were significant in the same direction (promote vs. protective) in more than one subgroup. For gene repression, 5, 7, and 4 characteristics of cancers of lung, plasma cells, and prostate, respectively, were significant, and of these characteristics, 4, 6, and 3 characteristics, respectively, were chromosome locations. For methylation, 2, 2, and 3 characteristics, respectively, were significant, respectively. Among these characteristics, 2 different chromosomes were significant in multiple myeloma and prostate cancer. In lung adenocarcinoma, characteristics that promoted gene repression were chromosomes 15 (OR (95% CI) = 3.27 (1.98-5.42)) and 18 (OR (95% CI) = 2.12 (1.21-3.72)) and ORF1 frame shifts (Test statistic = 2.57905, $p = 0.009967$), and uncommon characteristics were intragenic LINE-1s on chromosomes 8 (OR (95% CI) = 0.46 (0.29-0.72)) and 14 (OR (95% CI) = 0.45 (0.24-0.82)). For promoter hypermethylation, a conserved ORF Start-Stop was a preferred characteristic (OR (95% CI) = 7.19 (2.89-17.92)), whereas a lower number of ORF2 was conserved (OR (95% CI) = 0.29 (0.11-0.75)). In multiple myeloma, intragenic LINE-1s that promoted gene repression and promoter hypermethylation were located on different chromosomes. A striking example is on chromosome X. While many gene-repressing intragenic LINE-1s were presented on chromosome X (OR (95% CI) = 2.39 (1.65-3.47)), they were less likely to promote promoter methylation (OR (95% CI) = 0.19 (0.11-0.32)). Finally, in prostate cancer, the characteristics that promoted gene repression were chromo-

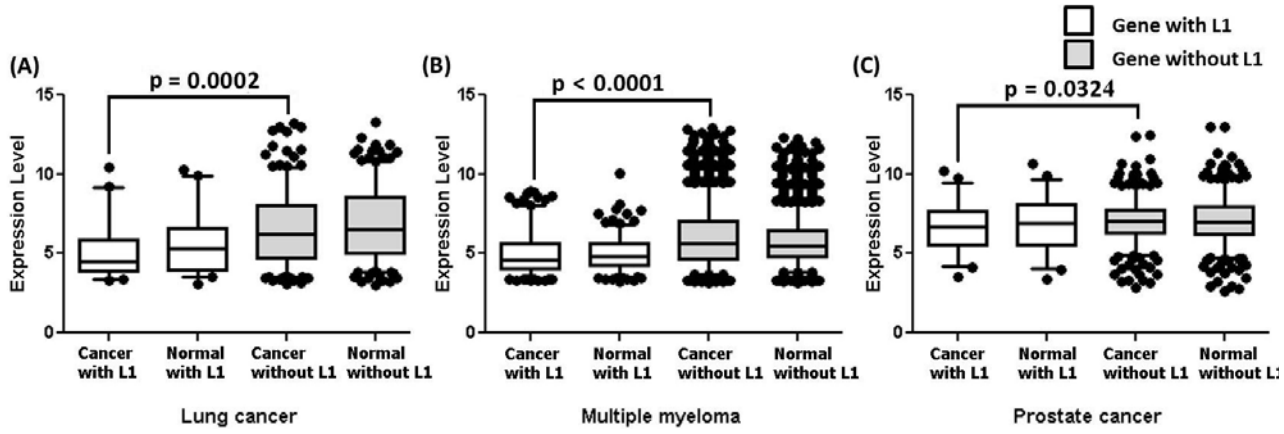


Figure 2. The expression profiles of promoter-hypermethylated genes with and without LINE-1s in lung adenocarcinoma, multiple myeloma and prostate cancer. A comparison of expression levels of hypermethylated genes was performed between the genes with and without LINE-1s from three case/control cancer experiments: (A) lung adenocarcinoma and non-tumor lung, (B) multiple myeloma and plasma cells of healthy individuals and (C) prostate cancer and normal prostate. The grey and white boxes represent hypermethylated genes with and without LINE-1s, respectively. The Student's t-test p-values between cancer with and without LINE-1s are shown.

somes 2 (OR (95% CI) = 2.04 (1.29-3.22)) and 3 (OR (95% CI) = 2.35 (1.53-3.62)) and GAGG of Ta subfamilies (OR (95% CI) = 6.74 (1.95-23.25)) possessing shared sequence variants, and uncommon characteristics were chromosome 10 (OR = 0.00). For promoter hypermethylation, there was a lower number of intragenic LINE-1s on chromosomes 13 (OR (95% CI) = 0.40 (0.21-0.77)) and 14 (OR (95% CI) = 0.36 (0.19- 0.69)) and smaller size of short target site duplications (Test statistic = -2.59198, p = 0.009627) than the other genes.

Intragenic *Alu* and methylated genes in lung adenocarcinoma, multiple myeloma and prostate cancer

To evaluate if promoter hypermethylation also associated with other intragenic repeat elements, intragenic *Alu* was selected to evaluate the three methylation arrays. The prevalence of promoter hypermethylation in lung adenocarcinoma was not significant in genes containing *Alu* (OR (95% CI) = 0.84 (0.60-1.17), slightly higher in multiple myeloma 1.11 (1.03-1.19) but lower in prostate cancer 0.86 (0.80-0.93). Unlike LINE-1s, there is no evidence of ubiquitous association between promoter hypermethylation and intragenic *Alu*.

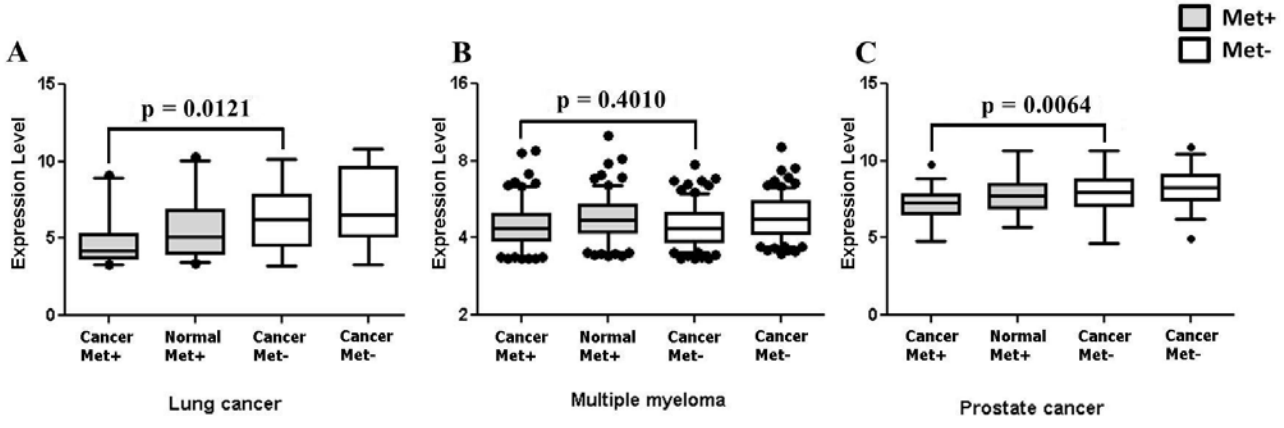


Figure 3. The expression profiles of down-regulated genes containing LINE-1s with and without promoter hypermethylation in lung adenocarcinoma, multiple myeloma and prostate cancer. A comparison of expression levels of genes containing LINE-1s was performed between hypermethylated (Met+) and non-hypermethylated (Met-) genes from three case/control cancer experiments: (A) lung adenocarcinoma and non-tumor lung, (B) multiple myeloma and plasma cells of healthy individuals and (C) prostate cancer and normal prostate. The grey and white boxes represent Met+ and Met- genes containing LINE-1s, respectively. The Student's t-test p-values between met+ and met- of each cancer are shown.

Table 5. The list of LINE-1 characteristics associated with gene promoter methylation and repressed gene regulation in lung adenocarcinoma, multiple myeloma and prostate cancer. (A) Chi-square test and (B) Student's *t*-test was used as statistical analyses both considering significant value to be p<0.01.

	Lung adenocarcinoma		Multiple myeloma		Prostate cancer	
	Down vs Not down	Met+ vs Met–	Down vs Not down	Met+ vs Met–	Down vs Not down	Met+ vs Met–
(A) Chi-square test [OR (95% CI)]						
Chromosome location						
Chromosome 2	N	N	0.51 (0.33 - 0.78)	N	2.04 (1.29 - 3.22)	N
Chromosome 3	N	N	N	N	2.35 (1.53 - 3.62)	N
Chromosome 4	N	N	0.49 (0.32 - 0.75)	N	N	N
Chromosome 8	0.46 (0.29 - 0.72)	N	N	N	N	N
Chromosome 9	N	N	N	2.71 (1.61 - 4.58)	N	N
Chromosome 10	N	N	4.24 (2.63 - 6.85)	N	0(-)	N
Chromosome 12	N	N	0.48 (0.28 - 0.83)	N	N	N
Chromosome 13	N	N	N	N	N	0.40 (0.21 - 0.77)
Chromosome 14	0.45 (0.24 - 0.82)	N	N	N	N	0.36 (0.19 - 0.69)
Chromosome 15	3.27 (1.98 - 5.42)	N	N	N	N	N
Chromosome 18	2.12 (1.21 - 3.72)	N	N	N	N	N
Chromosome 20	N	N	7.34 (2.08 - 25.87)	N	N	N
Chromosome X	N	N	2.39 (1.65 - 3.47)	0.19 (0.11 - 0.32)	N	N
ORF StartStop						
ORF2 cons	N	0.29 (0.11 - 0.75)	N	N	N	N
StartStop cons	N	7.19 (2.89 - 17.92)	N	N	N	N
Ta SSVs						
GAGG	N	N	N	N	6.74 (1.95 - 23.25)	N
Ta1-nd/d						
Ta1-nd	N	N	0.71 (0.58 - 0.88)	N	N	N
(B) Student's *t*-test (p-value)						
ORF1 frameshift	2.57905 (0.01)	N	N	N	N	N
find TSDs	N	N	N	N	N	-2.59198 (0.01)

N represents not significant value at p-value threshold 0.01.

Discussion

The functions of CU-DREAM and CU-DREAMX are to classify genes under two specific conditions and statistically test the gene distribution. If a significant number of genes share these two conditions with more than random distribution, the possibility that the two conditions are related is high. These two programs perform an analysis for the whole genome, so the number of genes for each test is usually high and the results have strong statistical significance. Note that CU-DREAM and CU-DREAMX possess two limitations. First, they cannot detect an association when the association is limited to a subset

Table 6. Promoter methylation statuses of the genes containing *Alu* in cancers.

Cancer types	*Alu*	Met + (%)	Met – (%)	Odds ratio	95% CI	p-value
Lung adenocarcinoma	*Alu*	217 (28.26)	360 (46.88)	0.84	Lower : 0.60	2.93E-01
	No *Alu*	80 (10.42)	111 (14.45)		Upper : 1.17	
Multiple myeloma	*Alu*	3,812 (26.33)	6,471 (44.70)	1.11	Lower : 1.03	7.78E-03
	No *Alu*	1,456 (10.06)	2,737 (18.91)		Upper : 1.19	
Prostate cancer	*Alu*	3,332 (23.02)	6,951 (48.02)	0.86	Lower : 0.80	1.34E-04
	No *Alu*	1,497 (10.34)	2,696 (18.62)		Upper : 0.93	

Met+ represents "hypermethylated", and Met- represents "unchanged" or hypomethylated".

of small size for the certain condition. For example, the association between up-regulated gene expression and promoter hypermethylation was significantly present in prostate cancer but not in lung adenocarcinoma and multiple myeloma. This missing may be due to lacking or limiting the mechanism causing the association in the latter two groups. Second, these programs usually yield less significant statistical values, higher p-values and lower OR (when OR>1), than the actual values of significance because of inter-assay variations between the two conditions. Here, we used intragenic LINE-1s from L1base (Penzkofer et al., 2005). These LINE-1s are found in most populations. A number of LINE-1 insertion dimorphisms (LIDs) (Badge et al., 2003; Pornthanakasem and Mutirangura, 2004) were not taken into account for this analysis. The genes containing LIDs were counted as genes without LINE-1s and resulted in increased p-values of CU-DREAM for the LINE-1 test. When comparing two GSEs from different patient groups by CU-DREAM, heterogeneity can cause some genes to be counted as not associated. In summary, CU-DREAM and CU-DREAMX can perform statistical significant tests but their results must be interpreted with caution when no statistical significance is reported.

We found that genes containing LINE-1s are frequently down-regulated and possess promoter hypermethylation in LINE-1-hypomethylated cancers. Intragenic LINE-1s and promoter hypermethylation synergistically repressed gene expression. Finally, few characteristics of LINE-1s influenced LINE-1 regulations. Among these significant characteristics, the "location" on different chromosomes was detected most often. Moreover, gene repression and promoter hypermethylation were influenced by different chromosomes. Therefore, cancer-associated intragenic LINE-1 epigenetic changes not only repress gene expression but also promote promoter methylation. The mechanisms of gene expression and promoter hypermethylation are influenced independently by different genome locations. Nevertheless, these two events synergistically repress genes.

Our data suggest that the association between LINE-1 hypomethylation and gene promoter hypermethylation may be connected by the cancer-associated changes in intragenic LINE-1s. Therefore, the association between promoter hypermethylation and LINE-1 hypomethylation that has been reported should be due to cancer development processes influenced by the epigenetic changes of intragenic LINE-1s. Nevertheless, more than one mechanism may potentially cause promoter hypermethylation, and some may be independent of the global hypomethylation process. As a result, some studies have revealed no correlation between promoter methylation and genome-wide hypomethylation.

Hypomethylated LINE-1s consequently increase LINE-1 transcription (Aporntewan et al., 2011). Antisense intronic RNA promoting promoter methylation has been reported in several biological events (Wutz et al., 1997; Tufarelli et al., 2003; Navarro et al., 2005). Therefore, it is reasonable to hypothesise that intragenic LINE-1s in cancer may promote promoter hypermethylation via LINE-1 RNA. Notably, LINE-1 RNA has been proven to mediate the cis regulatory function of LINE-1 hypomethylation. In cancer, hypomethylated intragenic LINE-1s produce LINE-1 RNA to form a complex with pre mRNA and AGO2 (Aporntewan et al., 2011). In X-inactivation, LINE-1s form a complex with *Xist* RNA to condense chromatin and inactivate the chromosome (Chow et al., 2010). LINE-1s on the inactive X have recently been shown to possess a lower methylation level (Singer et al., 2012). Nevertheless, the mechanism for how intragenic antisense RNA initiates promoter methylation requires further investigation. To the best of our knowledge, there are two interesting hypotheses. First, antisense RNA in plants can form a RISC complex to mediate DNA methylation. This process is known as RNA-directed DNA methylation (RdDM) (Zhang and Zhu, 2011). However, the evidence for RdDM in human cancers is lacking (Ting et al., 2005). Second, an example exists for a gene promoter to be hypermethylated when RNA processing is halted by an expanded triplet repeat sequence. For example, in Fragile X syndrome, the CGG repeat sequence at the 5'-untranslated region of the *FMR1* gene is expanded and subsequently halts RNA polymerisation. Similar to antisense RNA, the failure in RNA elongation by triplet repeat expansion is believed to promote *FMR1* promoter methylation (Jin and Warren, 2000). Further exploration of whether any of these hypotheses represent the mechanism for how intragenic LINE-1s induce promoter methylation will be interesting.

Although the cancer-associated LINE-1 hypomethylation mechanism is a generalised process that reduces LINE-1 methylation genome-wide, the most common LINE-1 character-

istics associated with whether LINE-1s mediated promoter methylation or gene repression was chromosome location. Some gene regulatory events and promoter methylations occur in a specific chromosomal region, such as in genomic imprinting (Edwards and Ferguson-Smith, 2007) and X-inactivation (Sharp et al., 2011). It will be interesting to further explore whether some similarities exist in terms of the mechanisms of these three epigenetic events.

A better understanding of these processes may lead to better applications for investigation and treatment of cancers and other human diseases or debilitating conditions. Several broader aspects should be considered in future studies. First, cancer-associated global hypomethylation reduces the methylation of other sequences in addition to LINE-1s (Kitkumthorn and Mutirangura, 2011). Other intragenic sequences, such as small genes in introns, may produce antisense RNA and cause promoter hypermethylation. Additionally, global hypomethylation can be found in other diseases, such as Systemic Lupus Erythematosus (Nakkuntod et al., 2011), and some physiologic conditions, such as cell differentiation (Phokaew et al., 2008) and aging (Jintaridth and Mutirangura, 2010; Zhu et al., 2011). Some cases of gene promoter methylation in these conditions may be mediated by a similar mechanism.

Second, some LINE-1s are active retrotransposable elements, and retrotransposition events have occurred frequently throughout human evolution (Gogvadze and Buzdin, 2009). The retrotransposition events create thousands of LIDs that are ancestor-specific (Badge et al., 2003; Pornthanakasem and Mutirangura, 2004; Ewing and Kazazian, 2011). It is tempting to hypothesise that some LIDs are non-synonymous polymorphisms and can increase the risk of certain diseases (Kitkumthorn and Mutirangura, 2011). These insertions could cause diseases by structural variants in the genomes (Ewing and Kazazian, 2011). Nevertheless, some of these if locate in intron and are hypomethylated, can cause gene repression (Aporntewan et al., 2011) or promoter hypermethylation. As a result, some disease- or cancer-related gene repression and/or promoter methylation may be specific to certain races.

Demethylating agents, such as 5-aza-deoxycytidine and 5-azacytidine, have been commonly used to revert promoter hypermethylation and up-regulate certain tumor suppressor genes for identification and cancer treatment (Das and Singal, 2004). However, our findings imply that a limitation exists for the use of demethylating agents on some promoters. DNA demethylating agents reduce methylation not only at promoters, but also elsewhere, including within the body of genes. The gene body demethylation will up-regulate intragenic antisense RNA and consequently undesirably repress gene expression.

In conclusion, by comparing expression and methylation using a genome-wide array, we found that cancer-associated intragenic LINE-1 changes do not only repress genes but also promote promoter hypermethylation. The mechanisms causing repression and methylation may be independently influenced by chromosome locations. Nevertheless, these events synergistically repress gene expression.

Acknowledgments This study was supported in part by the Research Chair Grant 2011, National Science and Technology Development Agency (NSTDA), Thailand, Four Seasons Hotel Bangkok's 4th Cancer Care Charity Fund Run in coordination with the Thai Red Cross Society, Rachadapiseksompoch Endowment Fund "Emerging Health Risks Cluster" and Chulalongkorn University. Suphakit Khowutthitham is supported by a Royal Golden Jubilee Ph.D. grant (PHD/0151/2546), the Thailand Research Fund and the 90th Anniversary of Chulalongkorn University Fund (Ratchadaphiseksomphot Endowment Fund), Chulalongkorn University. Chatchawit Aporntewan is supported by the Thailand Research Fund (Grant number RSA5580042).

References

Aporntewan C and Mutirangura A (2011) Connection up- and down-regulation expression analysis of microarrays (CU-DREAM): a physiogenomic discovery tool. Asian Biomed. 5: 257-262.

Aporntewan C, Phokaew C, Piriyapongsa J, Ngamphiw C, Ittiwut C, Tongsima S and Mutirangura A (2011) Hypomethylation of intragenic LINE-1 represses transcription in cancer cells through AGO2. PLoS One 6: e17934.

Baba Y, Huttenhower C, Nosho K, Tanaka N, Shima K, Hazra A, Schernhammer ES, Hunter DJ, Giovannucci EL, Fuchs CS, et al. (2010) Epigenomic diversity of colorectal cancer indicated by LINE-1 methylation in a database of 869 tumors. Mol. Cancer 9: 125.

Badge RM, Alisch RS and Moran JV (2003) ATLAS: A system to selectively identify human-specific L1 insertions. Am. J. Hum. Genet. 72: 823-838.

Barrett T, Suzek TO, Troup DB, Wilhite SE, Ngau WC, Ledoux P, Rudnev D, Lash AE, Fujibuchi W and Edgar R (2005) NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res. 33: D562-D566.

Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Marshall KA, et al. (2009) NCBI GEO: archive for high-throughput functional genomic data. Nucleic Acids Res. 37: D885-D890.

Belancio VP, Hedges DJ and Deininger P (2006) LINE-1 RNA splicing and influences on mammalian gene expression. Nucleic Acids Res. 34: 1512-1521.

Bollati V, Fabris S, Pegoraro V, Ronchetti D, Mosca L, Deliliers GL, Motta V, Bertazzi PA, Baccarelli A and Neri A (2009) Differential repetitive DNA methylation in multiple myeloma molecular subgroups. Carcinogenesis 30: 1330-1335.

Brait M and Sidransky D (2011) Cancer epigenetics: above and beyond. Toxicol. Mech. Methods 21: 275-288.

Chalitchagorn K, Shuangshoti S, Hourpai N, Kongruttanachok N, Tangkijvanich P, Thong-ngam D, Voravud N, Sriuranpong V and Mutirangura A (2004) Distinctive pattern of LINE-1 methylation

level in normal tissues and the association with carcinogenesis. Oncogene 23: 8841-8846.

Chen RZ, Pettersson U, Beard C, Jackson-Grusby L and Jaenisch R (1998) DNA hypomethylation leads to elevated mutation rates. Nature 395: 89-93.

Cho NY, Kim BH, Choi M, Yoo EJ, Moon KC, Cho YM, Kim D and Kang GH (2007) Hypermethylation of CpG island loci and hypomethylation of LINE-I and Alu repeats in prostate adenocarcinoma and their relationship to clinicopathological features. J. Pathol. 211: 269-277.

Choi IS, Estecio MR, Nagano Y, Kim DH, White JA, Yao JC, Issa JP and Rashid A (2007) Hypomethylation of LINE-1 and Alu in well-differentiated neuroendocrine tumors (pancreatic endocrine tumors and carcinoid tumors). Mod. Pathol. 20: 802-810.

Chow JC, Ciaudo C, Fazzari MJ, Mise N, Servant N, Glass JL, Attreed M, Avner P, Wutz A, Barillot E, et al. (2010) LINE-1 activity in facultative heterochromatin formation during X chromosome inactivation. Cell 141: 956-969.

Christensen BC, Marsit CJ, Houseman EA, Godleski JJ, Longacker JL, Zheng S, Yeh RF, Wrensch MR, Wiemels JL, Karagas MR, et al. (2009) Differentiation of lung adenocarcinoma, pleural mesothelioma, and nonmalignant pulmonary tissues using DNA methylation profiles. Cancer Res. 69: 6315-6321.

Cordaux R and Batzer MA (2009) The impact of retrotransposons on human genome evolution. Nat. Rev. Genet. 10: 691-703.

Das PM and Singal R (2004) DNA methylation and cancer. J. Clin. Oncol. 22: 4632-4642.

Daskalos A, Logotheti S, Markopoulou S, Xinarianos G, Gosney JR, Kastania AN, Zoumpourlis V, Field JK and Liloglou T (2011) Global DNA hypomethylation-induced Delta Np73 transcriptional activation in non-small cell lung cancer. Cancer Lett. 300: 79-86.

Edwards CA and Ferguson-Smith AC (2007) Mechanisms regulating imprinted genes in clusters. Curr. Opin. Cell Biol. 19: 281-289.

Ehrlich M (2002) DNA methylation in cancer: too much, but also too little. Oncogene 21: 5400-5413.

Ehrlich M, Jiang G, Fiala E, Dome JS, Yu MC, Long TI, Youn B, Sohn OS, Widschwendter M, Tomlinson GE, et al. (2002) Hypomethylation and hypermethylation of DNA in Wilms tumors. Oncogene 21: 6694-6702.

Ehrlich M, Woods CB, Yu MC, Dubeau L, Yang F, Campan M, Weisenberger DJ, Long TI, Youn B, Fiala ES, et al. (2006) Quantitative analysis of associations between DNA hypermethylation, hypomethylation, and DNMT RNA levels in ovarian tumors. Oncogene 25: 2636-2645.

Ewing AD and Kazazian HH (2011) Whole-genome resequencing allows detection of many rare LINE-1 insertion alleles in humans. Genome Res. 21: 985-990.

Feinberg AP and Tycko B (2004) The history of cancer epigenetics. Nat. Rev. Cancer 4: 143-153.

Florl AR, Steinhoff C, Muller M, Seifert HH, Hader C, Engers R, Ackermann R and Schulz WA (2004) Coordinate hypermethylation at specific genes in prostate carcinoma precedes LINE-1 hypomethylation. Br. J. Cancer 91: 985-994.

Gogvadze E and Buzdin A (2009) Retroelements and their impact on genome evolution and functioning. Cell. Mol. Life Sci. 66: 3727-3742.

Gutierrez NC, Ocio EM, Rivas JDL, Maiso P, Delgado M, Ferminan E, Arcos MJ, Sanchez ML, Hernandez JM and Miguel JFS (2007) Gene expression profiling of B lymphocytes and plasma cells from Waldenstrom's macroglobulinemia: comparison with expression patterns of the same cell counterparts from chronic lymphocytic leukemia, multiple myeloma and normal individuals. Leukemia 21: 541-549.

Herman JG (2005) Epigenetic changes in cancer and preneoplasia. Cold Spring Harb.Symp. Quant. Biol. 70: 329-333.

Iacopetta B, Heyworth J, Girschik J, Grieu F, Clayforth C and Fritschi L (2009) The MTHFR C677T and DeltaDNMT3B C-149T polymorphisms confer different risks for right- and left-sided colorectal cancer. Int. J. Cancer 125: 84-90.

Igarashi S, Suzuki H, Niinuma T, Shimizu H, Nojima M, Iwaki H, Nobuoka T, Nishida T, Miyazaki Y, Takamaru H, et al. (2010) A novel correlation between LINE-1 hypomethylation and the malignancy of gastrointestinal stromal tumors. Clin. Cancer Res. 16: 5114-5123.

Iramaneerat K, Rattanatunyong P, Khemapech N, Triratanachat S and Mutirangura A (2011) HERV-K hypomethylation in ovarian clear cell carcinoma is associated with a poor prognosis and platinum resistance. Int. J. Gynecol. Cancer 21: 51-57.

Jhavar S, Brewer D, Edwards S, Kote-Jarai Z, Attard G, Clark J, Flohr P, Christmas T, Thompson A, Parker M, et al. (2009) Integration of ERG gene mapping and gene-expression profiling identifies distinct categories of human prostate cancer. BJU Int. 103: 1256-1269.

Jin P and Warren ST (2000) Understanding the molecular basis of fragile X syndrome. Hum. Mol. Genet. 9: 901-908.

Jintaridth P and Mutirangura A (2010) Distinctive patterns of age-dependent hypomethylation in interspersed repetitive sequences. Physiol. Genomics 41: 194-200.

Kim BH, Cho NY, Shin SH, Kwon HJ, Jang JJ and Kang GH (2009) CpG island hypermethylation and repetitive DNA hypomethylation in premalignant lesion of extrahepatic cholangiocarcinoma. Virchows Arch. 455: 343-351.

Kitkumthorn N and Mutirangura A (2011) Long interspersed nuclear element-1 hypomethylation in cancer: biology and clinical applications. Clin. Epigenetics 2: 315-330.

Kobayashi Y, Absher DM, Gulzar ZG, Young SR, McKenney JK, Peehl DM, Brooks JD, Myers RM and Sherlock G (2011) DNA methylation profiling reveals novel biomarkers and important roles for DNA methyltransferases in prostate cancer. Genome Res. 21: 1017-1027.

Kongruttanachok N, Phuangphairoj C, Thongnak A, Ponyeam W, Rattanatanyong P, Pornthanakasem W and Mutirangura A (2010) Research Replication independent DNA double-strand break retention may prevent genomic instability. Mol. Cancer 9: 70.

Levy A, Sela N and Ast G (2008) TranspoGene and microTranspoGene: transposed elements influence on the transcriptome of seven vertebrates and invertebrates. Nucleic Acids Res. 36: D47-D52.

Nakkuntod J, Avihingsanon Y, Mutirangura A and Hirankarn N (2011) Hypomethylation of LINE-1 but not Alu in lymphocyte subsets of systemic lupus erythematosus patients. Clin. Chim. Acta. 412: 1457-1461.

Navarro P, Pichard S, Ciaudo C, Avner P and Rougeulle C (2005) Tsix transcription across the Xist gene alters chromatin conformation without affecting Xist transcription: implications for X-chromosome inactivation. Genes Dev. 19: 1474-1484.

Ogino S, Kawasaki T, Nosho K, Ohnishi M, Suemoto Y, Kirkner GJ and Fuchs CS (2008) LINE-1 hypomethylation is inversely associated with microsatellite instability and CpG island methylator phenotype in colorectal cancer. Int. J. Cancer 122: 2767-

2773.

Park SY, Yoo EJ, Cho NY, Kim N and Kang GH (2009) Comparison of CpG island hypermethylation and repetitive DNA hypomethylation in premalignant stages of gastric cancer, stratified for Helicobacter pylori infection. J. Pathol. 219: 410-416.

Penzkofer T, Dandekar T and Zemojtel T (2005) L1Base: from functional annotation to prediction of active LINE-1 elements. Nucleic Acids Res. 33: D498-D500.

Phokaew C, Kowudtitham S, Subbalekha K, Shuangshoti S and Mutirangura A (2008) LINE-1 methylation patterns of different loci in normal and cancerous cells. Nucleic Acids Res. 36: 5704-5712.

Poage GM, Houseman EA, Christensen BC, Butler RA, Avissar-Whiting M, McClean MD, Waterboer T, Pawlita M, Marsit CJ and Kelsey KT (2011) Global hypomethylation identifies Loci targeted for hypermethylation in head and neck cancer. Clin. Cancer Res. 17: 3579-3589.

Pornthanakasem W, Kongruttanachok N, Phuangphairoj C, Suyarnsestakorn C, Sanghangthum T, Oonsiri S, Ponyeam W, Thanasupawat T, Matangkasombut O and Mutirangura A (2008) LINE-1 methylation status of endogenous DNA double-strand breaks. Nucleic Acids Res. 36: 3667-3675.

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

Sanchez-Palencia A, Gomez-Morales M, Gomez-Capilla JA, Pedraza V, Boyero L, Rosell R and Farez-Vidal ME (2011) Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. Int. J. Cancer 129: 355-364.

Sharp AJ, Stathaki E, Migliavacca E, Brahmachary M, Montgomery SB, Dupre Y and Antonarakis SE (2011) DNA methylation profiles of human active and inactive X chromosomes. Genome Res. 21: 1592-1600.

Singer H, Walier M, Nusgen N, Meesters C, Schreiner F, Woelfle J, Fimmers R, Wienker T, Kalscheuer VM, Becker T, et al. (2012) Methylation of L1Hs promoters is lower on the inactive X, has a tendency of being higher on autosomes in smaller genomes and shows inter-individual variability at some loci. Hum. Mol. Genet. 21: 219-235.

Ting AH, Schuebel KE, Herman JG and Baylin SB (2005) Short double-stranded RNA induces transcriptional gene silencing in human cancer cells in the absence of DNA methylation. Nat. Genet. 37: 906-910.

Trankenschuh W, Puls F, Christgen M, Albat C, Heim A, Poczkaj J, Fleming P, Kreipe H and Lehmann U (2010) Frequent and distinct aberrations of DNA methylation patterns in fibrolamellar carcinoma of the liver. PLoS One 5.

Tufarelli C, Stanley JAS, Garrick D, Sharpe JA, Ayyub H, Wood WG and Higgs DR (2003) Transcription of antisense RNA leading to gene silencing and methylation as a novel cause of human genetic disease. Nat. Genet. 34: 157-165.

Walker BA, Wardell CP, Chiecchio L, Smith EM, Boyd KD, Neri A, Davies FE, Ross FM and Morgan GJ (2011) Aberrant global methylation patterns affect the molecular pathogenesis and prognosis of multiple myeloma. Blood 117: 553-562.

Woloszynska-Read A, Zhang W, Yu J, Link PA, Mhawech-Fauceglia P, Collamat G, Akers SN, Ostler KR, Godley LA, Odunsi K, et al. (2011) Coordinated cancer germline antigen promoter and global DNA hypomethylation in ovarian cancer: association with the BORIS/CTCF expression ratio and advanced stage. Clin. Cancer Res. 17: 2170-2180.

Wutz A, Smrzka OW, Schweifer N, Schellander K, Wagner EF and Barlow DP (1997) Imprinted expression of the Igf2r gene depends on an intronic CpG island. Nature 389: 745-749.

Yamamoto E, Toyota M, Suzuki H, Kondo Y, Sanomura T, Murayama Y, Ohe-Toyota M, Maruyama R, Nojima M, Ashida M, et al. (2008) LINE-1 hypomethylation is associated with increased CpG island methylation in Helicobacter pylori-related enlarged-fold gastritis. Cancer Epidemiol. Biomarkers Prev. 17: 2555-2564.

Zhang HM and Zhu JK (2011) RNA-directed DNA methylation. Curr. Opin. Plant Biol. 14: 142-147.

Zhu ZZ, Sparrow D, Hou LF, Tarantini L, Bollati V, Litonjua AA, Zanobetti A, Vokonas P, Wright RO, Baccarelli A, et al. (2011) Repetitive element hypomethylation in blood leukocyte DNA and cancer incidence, prevalence, and mortality in elderly individuals: the Normative Aging Study. Cancer Causes Control 22: 437-447.

Gene Ontology-Based Analysis Reveals a Physiological Role of Upstream Mononucleotide A-repeats in Mammals

Chatchawit Aporntewan[12*] and Apiwat Mutirangura[3]

[1] Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok, 10330, Thailand

[2] Program in Bioinformatics and Computational Biology, Graduate School, Chulalongkorn University, Bangkok, Thailand.

[3] Center of Excellence in Molecular Genetics of Cancer and Human Diseases, Department of Anatomy, Faculty of Medicine, Chulalongkorn University, Bangkok, 10330, Thailand

* Corresponding author
E-mail: Chatchawit.A@chula.ac.th

Abstract

Herein, we performed a thorough investigation in the genomes of 8 mammals and 12 non-mammals. The imbalance of A-repeats between upstream and downstream of transcription start sites is conserved in all mammals and is also found in some non-mammalian vertebrates. Moreover, short repeats (2 to 9 bp) are enriched in tissue-specific genes, whereas long repeats (10 to 30 bp) are enriched in housekeeping genes. A joint analysis of *H. sapiens*, *M. musculus*, and *R. norvegicus* reveals 10 genes and 25 Gene Ontology terms that are significantly enriched with upstream A-repeats. These results indicate that A-repeats play a regulatory role in DNA/RNA metabolism, including viral infection, as validated by two microarray experiments involving infection with influenza A or Epstein-Barr virus. Our findings suggest that upstream A-repeats may act as *cis*-regulatory elements in a variety of mammalian gene groups.

Introduction

A microsatellite is a repeat of the same unit [1]. The repeated unit is a sequence of nucleotides (A, T, C, or G). For instance, 'CAGCAGCAGCAG' is a microsatellite that repeats a unit of 'CAG' four times. A more general term of microsatellites is tandem repeats (TRs). A form of mutation called DNA replication slippage contracts and expands the length of TRs by decreasing and increasing the number of repeat units. TRs are found ubiquitously in both coding and non-coding regions of eukaryotic genomes. The mutation of coding TRs obviously alters subsequent mRNAs, protein structures, and phenotypes. Huntington's disease is a well-known phenotype associated with tri-nucleotide repeats in the coding region. The number of CAG units ≥ 36 causes a malfunction of Huntingtin protein. Traditionally, non-coding TRs were believed to have no functions, and they were called 'junk' or 'selfish' DNA [2]. In fact, non-coding TRs play a crucial role in cell physiology. Several lines of evidence show that there is a great deal of high variability in TRs [3,4]. This repeat variability, specifically within gene promoters, correlates with variations in gene expression, which in turn connected with phenotypes [5]. The adaptation via gene modulation is a vital key for survival in ecological niches and coping with environmental changes. A high degree of variation within the TRs among related species suggests that TRs have been evolved through natural selection and drive the emergence of new species [6].

Mononucleotide repeats (unit size = 1) is the simplest class, but constitute the largest part of TRs. Extensive studies in yeasts show that non-coding poly(dA:dT) tracts correlate with nucleosome-depleted regions [7,8]. These poly(dA:dT) tracts are in proximity with gene promoters and evolutionarily conserved. Hypothetically, an intrinsic property of poly(dA:dT) tracts is to resist sharp DNA bending [9]. Thus, poly(dA:dT) tracts inhibit the nucleosome formation and enable more accessibility to transcription factors. Although the functional role of poly(dA:dT) is well established in yeasts, the underlying mechanisms are still largely unknown.

Recently, the imbalance of A-repeats between the upstream and the downstream of transcription start sites (TSSs) has been discovered [10]. Although the number of poly(dA:dT) tracts is balance, counting only sense A-repeats and excluding antisense A-repeats reveals the unexpected disproportion. Sense A-repeats are enriched upstream, but depleted downstream of TSSs. The imbalance was observed in three mammals, *Rattus norvegicus*, *Mus musculus* and *Homo sapiens*, but not found in non-mammalian genomes, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, and *Drosophila melanogaster*. The imbalance of A-repeats does not occur at random, and thus it implies a physiological function. Moreover, the imbalance of A-repeats found in mammals but not in yeasts suggests a novel role and underlying mechanism. Consequently, it has been proved that upstream A-repeats serve as *cis*-regulatory elements, and Argonaute (Ago) proteins as a *trans*-acting factor bind to upstream A-repeats and regulate the expression of the host genes [10]. The enrichment of long A-repeats (10 to 30 bp) in 'human' housekeeping genes suggests an approximated function [10]. Nevertheless, the specific role and functionality of A-repeats in mammals is still unknown.

In this paper, we extended the previous work, which conducted the study only in a total of six species, three mammals and three non-mammals. Therefore, we set out three specific aims. First, we aimed to analyze the imbalance of A-repeats in other mammals. We also aimed to analyze the enriched A-repeats in housekeeping genes in species other than humans. Additionally, the analysis may allow us to trace back the origin of the imbalance in the genomes of 20 species to see how the evolution has shaped up the distribution of A-repeats around TSSs. Second, we aimed to identify a specific function of upstream A-repeats. Basically, we searched for pathways or gene sets that were significantly enriched with upstream A-repeats. Third, we aimed to validate the function of A-repeats using a computational approach. We searched in public databases for microarray datasets in which experimental and control groups were conditioned by the function of A-repeats, i.e. virus infection. We expected to see the correlation between the number of A-repeats and the expression level of their host genes.

To accomplish our aims, we investigated the distribution of A-repeats in the whole genomes of 20 living organisms in the National Center for Biotechnology Information (NCBI) database [11]. Next, we performed a computational analysis of 2,403 Gene Ontology (GO) terms [12]. The significant GO terms involved DNA/RNA metabolisms and virus infection, which led us to perform functional validation by microarray experiments. Finally, two public datasets were selected from the Gene Expression Omnibus (GEO) [11,13]. The experimental group was infected with influenza A or Epstein-Barr virus (EBV). The functional role of upstream long A-repeats was confirmed by the differential number of repeats between regulated and unregulated genes.

Materials and Methods

Distribution of Mononucleotide Repeats

Our study was limited to only 20,000 bp around TSSs (Fig. 1). A total of 20,000 bp were divided into 25 bins of 800 bp each. Bins 1 to 10 were referred to as the upstream region, and bins 16 to 25 were referred to as the downstream region. The TSSs were located at the middle of the 13th bin. The number of repeats including the repeats that spanned across two bins were counted in the unit of base pairs, and were accumulated in each bin (S1 Fig.). Next, the amount of repeats in each bin was normalized by dividing by the total number of genes so that the number of repeats between two unequal sets of genes could be compared. Thus, the measurement unit of each bin was base pairs per gene.

Fig. 1. Bin structure around the transcription start sites (TSSs). There are 25 bins of 800 bp each. The first 10 bins (the 1st to 10th bins) are referred to as upstream of TSSs, and the last 10 bins (the 16th to the 25th bins) are referred to as downstream of TSSs. The TSS is centered at the 13th bin.

The binning technique was required for the Student's *t*-test because a *t*-test needed at least two samples per group. Thus, at least two bins were required of a *t*-test. Too large bins reduced the specificity of genomic locations, while too small bins entailed insufficient number of repeats for further analysis. The number of bins and bin size used in this paper were compatible with those in the previous work [10].

Genomes of Living Organisms

Initially, the genomes of all available organisms at NCBI [11] were considered on December 21, 2013. Next, we found that only the living organisms with at least about 1,000 known protein-coding transcripts were suitable for our analysis. In addition, the transcript status must be 'reviewed', 'validated' or 'provisional'. The 20 living organisms that passed the conditions were *Ashbya gossypii, Encephalitozoon cuniculi, Schizosaccharomyces pombe, Saccharomyces cerevisiae, Arabidopsis thaliana, Glycine max, Solanum lycopersicum, Oryza sativa, Drosophila melanogaster, Caenorhabditis elegans, Danio rerio, Gallus gallus, Canis lupus familiaris, Sus scrofa, Bos taurus, Mus musculus, Rattus norvegicus, Pan troglodytes, Pongo abelii, and Homo sapiens* (S1 Table). This number of living organisms was comparable to those of Tandem Repeats Database [14].

Housekeeping and Tissue-Specific Genes

A total of 575 housekeeping genes and 7,261 tissue-specific genes in humans were identified by Eisenberg and Levanon [15] and the Tissue-specific Gene Expression and Regulation (TiGER) database [16], respectively. The list of housekeeping genes was downloaded from http://www. compugen.co.il/supp_info/Housekeeping_genes.html. The list of tissue-specific genes was downloaded from http:// bioinfo.wilmer.jhu.edu/tiger/download/ref2tissue-Table.txt.

HomoloGene

Although the housekeeping and tissue-specific genes in humans were identified, very few genes had been reported in other living organisms. To compare across species, the HomoloGene Build 67 was used to map the genes of other organisms to human genes [11]. Subsequently, the human genes could be classified as housekeeping or tissue-specific genes [15,16]. Only the homolog genes of *C. l. familiaris, S. scrofa, B. taurus, M. musculus, R. norvegicus, P. troglodytes, P. abelii* were available. Note that a homolog of a human gene existed only if an organism was relatively close to humans. Some species were not included in HomoloGene because of their limited or incomplete genomic information (< 10,000 UniGene entries).

Gene Ontology (GO) Terms

A total of 14,271 GO terms [12] were downloaded from http://www.geneontology.org/gene-associations/gene_association.goa_human.gz on March 29, 2014. Each GO term, which is associated with a set of proteins, belonged to one of three domains: cellular component, molecular function, or biological process. A mapping file between the proteins and genes was downloaded from ftp://ftp.ebi.ac.uk/pub/databases/uniprot/current_release/knowledgebase/idmapping/idmapping.dat.gz on March 30, 2014. The GO terms associated with less than 20 genes were discarded. Finally, 2,403 GO terms remained (S2 Table).

Statistical Selection of Genes and GO terms that were enriched with upstream A-repeats

The number of upstream A-repeats of a single gene was the sum of the amount from bins 1 to 10. For each species, the top 500 highly enriched genes were identified. Next, the correlation between species was determined by a 2×2 contingency table. The row and the column counted the number of homolog genes of the first and second species, respectively. The first row and the first column counted the number of highly enriched genes (the top 500 genes), while the second row and the second column counted the rest. The odds ratio (OR) and Pearson's chi-squared *p*-value were calculated according to the 2×2 table.

The number of upstream A-repeats of a single GO term (a set of gene) was obtained using the sum of each gene. Additionally, the significance of the enrichment was determined by a paired *t*-test between a gene set of interest (a GO term) and a set of all of the genes in the entire genome. Multiple hypothesis testing was implemented due to the 2,403 GO terms. The multiple hypothesis correction was performed using false-discovery rate (FDR) analysis [17]. The QVALUE package for R statistical software was equipped with default parameters and the bootstrap option instead of the smoother option for the estimation of π_0 [18]. The estimation of π_0 was denoted by $\hat{\pi}_0$, which was an estimate of the overall proportion of true null hypotheses. The proportion of significant tests is 1 – $\hat{\pi}_0$.

The obtained $\hat{\pi}_0$ values were 0.47, 0.72, and 0.68 in *H. sapiens*, *M. musculus* and *R. norvegicus*, respectively. When the *q*-value was restricted to < 0.05, the number of significant GO terms was 6.41% (154 ÷ 2,403), 2.37% (57 ÷ 2,403), and 2.04% (49 ÷ 2,403) in *H. sapiens*, *M. musculus* and *R. norvegicus*, respectively.

Note that Student's *t*-test was frequently used throughout this study to compare the number of repeats in two cases. First, an unpaired *t*-test was used to compare different regions (i.e., upstream vs. downstream) of a gene set. Second, a paired *t*-test was used to compare the same region between two sets of genes (i.e., housekeeping vs. tissue-specific genes). The default parameters of a *t*-test are two-tailed distribution and two-sample unequal variance, if not paired.

Microarray Experiment Datasets

Two microarray experiments were selected from the Gene Expression Omnibus (GEO) [11,13]. The public datasets described the alteration in gene expression that was subject to influenza A (GSE24533) and Epstein-Barr virus (GSE45829) [19,20]. The regulated genes were identified by the software CU-DREAM [21]. The CU-DREAM parameters are shown in S3 Table.

Results

The imbalance between the upstream and downstream repeats is conserved in mammals.

The bin structure around transcription start sites (TSSs) is shown in Fig. 1. A sequence of 20,000 bp is divided into 25 bins of 800 bp each. The TSS is centered in the 13th bin. Bins 1 to 10 are referred to as the upstream region, and bins 16 to 25 are referred to as the downstream region. The counting and normalization methods are illustrated in S1 Fig. Bins 11 to 15 are the region of CpG islands [22,23]. The CG-rich sequences automatically prevent the occurrence of A- and T-repeats. Thus, the number of repeats drops very sharply and forms a V-shape (Figs. 2A and 2B). We excluded bins 11 to 15 in further analysis because the V-shape is symmetric between downstream and upstream and therefore irrelevant to the imbalance.

Fig. 2. Genome-wide distribution of A- and T-repeats in 8 mammals. The repeats with lengths of 10 to 30 bp located around the TSSs throughout the entire genome are shown. **(A,B)** The horizontal axis consists of 25 bins. The vertical axis represents the normalized number of repeats in base pairs per gene. **(C)** An unpaired *t*-test compares the number of repeats between the upstream (bins 1 to 10) and downstream (bins 16 to 25) regions, yielding the *p*-values as summarized in Table 1A.

All the eight mammals exhibited a similar pattern. Sense long A-repeats (10 to 30 bp) are more abundant upstream (Fig. 2A), while sense long T-repeats (10 to 30 bp) are more abundant downstream (Fig. 2B). The number of A- and T-repeats between the upstream (bins 1 to 10) and

downstream (bins 16 to 25) regions were compared by unpaired *t*-tests (Fig. 2C). The *t*-test *p*-values are summarized in Table 1A.

Table 1. The tabulation of *t*-test *p*-values. (A) The *p*-values of unpaired *t*-test in Figure 2C. **(B)** The *p*-values of paired *t*-test in Figure 3A & 3B.

Species	(A) Upstream vs. Downstream Unpaired *t*-test *p*-values		(B) Housekeeping vs. Tissue-specific genes Paired *t*-test *p*-values			
	Long A-repeat	Long T-repeat	Short A-repeat	Short T-repeat	Long A-repeat	Long T-repeat
Mammals						
C. l. familiaris	2.58E-05	3.31E-04	2.75E-07	9.86E-01	1.53E-02	4.46E-04
S. scrofa	8.48E-10	2.77E-02	NA	NA	NA	NA
B. taurus	4.29E-10	8.22E-09	4.58E-13	4.12E-11	9.99E-04	9.95E-01
M. musculus	4.60E-11	4.55E-05	5.11E-15	1.76E-09	5.71E-06	4.64E-08
R. norvegicus	7.59E-10	1.02E-03	2.29E-13	1.84E-09	1.76E-07	5.30E-10
P. troglodytes	5.43E-08	1.05E-03	1.23E-07	2.74E-04	6.26E-02	6.82E-07
P. abelii	2.32E-10	3.92E-10	NA	NA	NA	NA
H. sapiens	4.09E-13	1.15E-06	6.55E-19	1.95E-13	7.83E-07	1.03E-09
Non-mammals						
A. gossypii	5.17E-01	5.96E-01	3.84E-07	5.70E-06	9.51E-01	4.11E-01
E. cuniculi	1.32E-01	2.47E-01	NA	NA	NA	NA
S. pombe	2.62E-02	3.60E-01	6.90E-01	2.45E-01	1.42E-01	9.00E-01
S. cerevisiae	8.64E-02	5.84E-03	5.41E-01	3.76E-01	8.09E-01	5.55E-01
A. thaliana	2.53E-02	1.06E-03	7.20E-01	1.97E-04	1.03E-01	3.71E-01
G. max	1.32E-01	3.32E-01	NA	NA	NA	NA
S. lycopersicum	1.76E-02	5.69E-02	NA	NA	NA	NA
O. sativa	4.94E-01	1.34E-05	4.31E-01	6.48E-01	8.69E-01	4.12E-01
C. elegans	1.23E-01	4.84E-02	NA	NA	NA	NA
D. melanogaster	3.76E-01	3.33E-03	NA	NA	NA	NA
D. rerio	7.59E-01	4.98E-08	4.97E-03	4.59E-01	3.92E-03	9.36E-02
G. gallus	1.84E-05	1.73E-09	1.78E-11	2.42E-09	2.99E-03	3.55E-08

The imbalance between the upstream and downstream regions varies among species. The most significant imbalance of A-repeats in mammals are found in *H. sapiens*, *M. musculus*, *P. abelii*, *B. taurus*, *R. norvegicus*, *S. scrofa*, *P. troglodytes*, and *C. l. familiaris*, respectively (Table 1A). There is a significant imbalance of A-repeats in a non-mammalian vertebrate (*G. gallus*). However, this imbalance is not significant (*p*-value > 0.01) in another non-mammalian vertebrate (*D. rerio*) and in the other non-mammalian organisms (Table 1A, S2 Fig.).

Although the normalization was conducted, the number of A- and T-repeats varies among the species. The highest number is observed in *S. scrofa* and primates (*H. sapiens, P. abelii* and *P. troglodytes*), while a lower number is in *C. l. familiaris*, *R. norvegicus*, *M. musculus*, and *B. taurus* (Figs. 2A and 2B). The number of A-repeats was correlated with body temperature, body mass, and circadian period (Table 2).

Table 2. Correlations of number of A-repeats with body temperature, body mass, and circadian period. The total number of A-repeats is the sum of bins 1 to 25. The circadian period, body temperature, and body mass were taken from the website of Department of Psychology, Boise State University, USA at URL: http://www.circadian.org/animal.html.

Mammals	Total number of A-repeats (bp/gene)	Body temperature (°C)	Body mass (kg)	Circadian period (hours)
C. l. familiaris	48.70	39	30	24.4
B. taurus	22.03	38	800	unknown
M. musculus	33.94	36.9	0.03	23.6
R. norvegicus	36.66	37.3	0.4	24.2
H. sapiens	61.78	37	70	24.6
Correlation with A-repeats		**-0.0558**	**-0.6239**	**0.8532**
***P*-value (two-tailed probability)**		**0.9290**	**0.2607**	**0.1468**

The enrichment of long A- and T-repeats in housekeeping genes is conserved in mammals

Despite the imbalance of repeats between the upstream and downstream regions, there is a differential number of A-repeats between housekeeping and tissue-specific genes as shown in the previous work [10]. Herein, we extended the scope of study from only humans to include other mammals. The housekeeping and tissue-specific genes in non-human organisms were identified using HomoloGene [11]. The differential number of repeats between housekeeping and tissue-specific genes was determined using bins 1 to 10, bins 16 to 25, and a paired *t*-test (Figs. 3A and 3B). The *t*-test *p*-values are summarized in Table 1B.

Fig. 3. Comparison of the number of repeats between housekeeping (HK) and tissue-specific (TS) genes in six mammals. The repeats in the upstream (bins 1 to 10) and downstream (bins 16 to 25) regions between HK and TS genes are compared. The paired *t*-test *p*-values are summarized in Table 1B. **(A)** Short repeats (2 to 9 bp). **(B)** Long repeats (10 to 30 bp).

The previous findings in humans can be generalized to other mammals. Short A- and T-repeats (2 to 9 bp) are enriched in tissue-specific genes, whereas long A- and T-repeats (10 to 30 bp) are enriched in housekeeping genes. In contrast, non-mammals do not show the differential number of repeats between the two sets of genes because most of their *p*-values are not significant (Table 1B, S3 Fig.). The enrichment of long A-repeats in housekeeping genes plus their frequency upstream in a wide range of mammals suggests that long A-repeats may correlate with gene functions. Moreover, these functions may be conserved only in mammals.

Genes enriched with upstream long A-repeats are common among related organisms

We found that the genes highly enriched with upstream long A-repeats (10 to 30 bp) are shared among related species. First, the top 500 genes enriched with upstream long A-repeats were

identified from 12,504 homologous genes of *H. sapiens*, *M. musculus*, and *R. norvegicus*. Note that only the set of organisms that share a large number of homolog genes are suitable for further analysis. Second, two species were correlated by a 2×2 contingency table. Third, the odds ratio (OR) and Pearson's chi-squared *p*-values were obtained (Fig. 4, S4 Table). The correlation of genes across species could be achieved through HomoloGene [11].

Fig. 4. Correlations of genes enriched with upstream long A-repeats (10 to 30 bp). Each 2×2 table associates the top 500 homologous genes between two species. **(A)** *M. musculus* vs. *R. norvegicus*. **(B)** *H. sapiens* vs. *M. musculus*. **(C)** *H. sapiens* vs. *R. norvegicus*.

The significant and high values of the odds ratio (OR) in Fig. 4 suggest that the genes that are highly enriched with upstream long A-repeats are shared among mammals. Moreover, two related species, *M. musculus* and *R. norvegicus*, show a stronger correlation compared to *H. sapiens* with *M. musculus* and *R. norvegicus*. A Venn diagram illustrates the exact number of enriched genes that are shared by *H. sapiens*, *M. musculus*, and *R. norvegicus* (Fig. 5, S5 Table). The intersection of these three species resulted in 10 genes: *CAPZB*, *EIF2B1*, *EIF4H*, *GALR2*, *PSMG2, RNF114*, *SLC25A20*, *TSNAXIP1*, *UBN1*, and *UBTF* (written in human gene symbols and sorted in alphabetical order). The non-identical but correlated gene sets that were identified in the three mammals suggest that A-repeats may possess multiple and overlapping functions. Some of these functions might be shared across the entire class of Mammalia, and some functions might be specifically developed and conserved in related species.

Fig. 5. The intersection of the top 500 homologous genes. These genes were enriched with upstream long A-repeats (10 to 30 bp) in *H. sapiens*, *M. musculus*, and *R. norvegicus*. A total of 10 genes were highly enriched in all three of the species.

Physiological functions of the upstream long A-repeats as identified by a GO-based analysis

A-repeats are *cis*-regulatory elements and are ubiquitously found throughout the entire genomes [10]. Thus, A-repeats can regulate multiple genes simultaneously. To identify the physiological functions of genes containing A-repeats, a statistical test was performed for each of 2,403 Gene Ontology (GO) terms (see Material and Methods).

A Venn diagram illustrates the intersection of *H. sapiens*, *M. musculus*, and *R. norvegicus* (Fig. 6, S6 Table). Only 25 GO terms are significantly enriched with upstream long A-repeats. The 25 GO terms were classified into three domains that consisted of 12 biological processes, 6 cellular components, and 7 molecular functions. Most of the significant GO terms constitute the fundamental basis of cells, for instance, gene expression, nucleus, and DNA/RNA binding. These terms are difficult to assign a biological condition for an experimental group. In contrast, the viral terms are more convenient for

functional validation. The experimental group was set to be virus infection, and the control group was chosen to be mock infection.

Fig. 6. The intersection of the significant Gene Ontology (GO) terms. These GO terms were enriched with upstream long A-repeats (10 to 30 bp) in *H. sapiens*, *M. musculus*, and *R. norvegicus*. A total of 25 GO terms were significantly enriched in all three of the species (q-value ≤ 0.05).

The function of upstream long A-repeats was validated by microarray experiments

To validate the function of upstream long A-repeats, we selected two viral microarray experiments in Gene Expression Omnibus [11,13]. The first experiment was influenza A infection (GSE24533), and the second experiment was Epstein-Barr virus infection (GSE45829). A microarray experiment compared the gene expression between an experimental and a control group. As a result, the genes were statistically classified into three sets: down-regulation (Dn), up-regulation (Up), and unchanged expression (Nc). If A-repeats did not play any roles in virus infection, differential numbers of A-repeats among Dn, Up, and Nc would not be observed. A differential number is a significant result.

The normalized number of upstream A-repeats in the three different gene sets (Dn, Up, and Nc) was compared (Fig. 7, S7 Table). The significance of the differential number of A-repeats (bins 1 to 10) was determined by a paired *t*-test. GSE24533 (influenza A) yielded the following *p*-values: 4.76e-04 (Dn vs. Nc) and 1.00e-01 (Up vs. Nc). GSE45829 (Epstein-Barr virus) yielded the following *p*-values: 6.60e-04 (Dn vs. Nc) and 2.51e-06 (Up vs. Nc). The results suggest that upstream long A-repeats are statistically associated with gene regulation due to virus infection.

Fig. 7. Comparison of the number of upstream long repeats (10 to 30 bp, bins 1 to 10) between the genes with regulated and unchanged expression levels. Dn, Up, Nc denote down-regulated, up-regulated, and unchanged expression genes, respectively. **(A)** Influenza A infection dataset (GSE24533) yielded the following paired *t*-test *p*-values: 4.76e-04 (Dn vs. Nc) and 1.00e-01 (Up vs. Nc). **(B)** EBV infection dataset (GSE45829) yielded the following paired *t*-test *p*-values: 6.60e-04 (Dn vs. Nc) and 2.51e-06 (Up vs. Nc).

Discussion

In the previous work [11], although only six model organisms (*S. cerevisiae*, *D. melanogaster*, *C. elegans*, *R. norvegicus*, *M. musculus*, and *H. sapiens*) were investigated, the *cis*-regulatory roles of upstream A-repeats and AGO proteins as *trans*-acting factors of mammals were identified. In this paper, the investigation was extended to other mammals (*C. l. familiaris*, *S. scrofa*, *B. taurus*, *P. troglodytes*, and *P. abelii*) and non-mammals (*A. gossypii*, *E. cuniculi*, *S. pombe*, *A. thaliana*, *G. max*, *S. lycopersicum*, *O. sativa*, *D. rerio*, and *G. gallus*).

The similar findings across the eight mammals suggest that the imbalance of A-repeats around TSSs might be a common characteristic of mammalian genomes (Fig. 2, Table 1A). A lower degree of imbalance was also observed in a non-mammalian vertebrate (*G. gallus*) but was not observed in another non-mammalian vertebrate (*D. rerio*) (S2 Fig.). This suggests that the imbalance of A-repeats may develop in vertebrates or its predecessors.

The enrichment of long A-repeats upstream of mammalian housekeeping genes (Fig. 3, Table 1B) indicates that these repeats may play regulatory roles and mediate a large number of genes in concert. Interestingly, the number of A-repeats varies greatly among species (Fig. 2). The numbers of A-repeats does not correlate with body temperature or body mass as suggested in the previous studies [24,25], but the number of A-repeats correlates with circadian period (r = 0.8532, p = 0.1468) (Table 2). The p-value does not meet the significant threshold because of too small sample size. However, the exact circadian periods of mammals are still largely unknown.

The strong correlation of highly enriched genes between related species (*M. musculus* vs. *R. norvegicus*) provides additional evidence of repeat conservation (Fig. 4). The weaker correlation between unrelated species (*H. sapiens* vs. *M. musculus* and *H. sapiens* vs. *R. norvegicus*) suggests that the variability of A-repeats may correlate with the divergence of species.

The 10 genes that were commonly enriched with upstream long A-repeats in *H. sapiens, M. musculus, and R. norvegicus* were identified (Fig. 5). The individual function of each gene was known; however, their collective function that was subject to A-repeats was obscured. Therefore, we analyzed a set of genes or a GO term instead of a single gene (Fig. 6). The 25 GO terms constitutes the fundamental basis of cells, for instance, transcription and translation processes, primary cellular components (nucleus, nucleolus, cytoplasm, and mitochondrion), and molecular functions, such as DNA/RNA/protein binding. A brief summary of the 25 GO terms would be “DNA/RNA metabolisms”. At least three GO terms are directly associated with virus infection. In fact, all viral strains are made of either DNA or RNA, and thus, the viral terms overlap with DNA/RNA metabolisms.

A clue to the functional role of A-repeats may lie in the viral genomes. The observed to the expected (O/E) number of mononucleotide repeats (MNRs) ratio in viruses positively correlates with the range of hosts that the viruses can infect [26]. For instance, the O/E ratios for viruses in mammals have a tendency to be higher than those of bacterial viruses. The MNRs in viral genomes, similarly to A-repeats in mammals, may function as *cis*-regulatory elements, but viruses cannot produce transcription factors by themselves. The large number of MNRs may allow viruses to exploit the host’s transcription factors because A-repeats are common *cis*-regulatory elements in housekeeping genes of mammals. Therefore, the MNRs in viral genomes may enhance the capability of viruses to operate in a wide range of mammalian hosts. The absence of MNRs may increase the complexity of *cis* elements, and require more specific transcription factors.

The evidence for the functional role of upstream long A-repeats in viral infection is strengthened by the results of two microarray experiments. Public datasets demonstrate that regulated genes contain more repeats than non-regulated genes (Fig. 7). These results suggest that upstream A-repeats may increase the susceptibility to alteration of gene expression due to influenza A and EBV infection. We searched the highly-enriched genes (Fig. 5) in the literature on virology and found additional supporting evidence.

CAPZB denotes capping protein (actin filament) muscle Z-line, beta. This gene encodes a member of the F-actin capping protein family. CAPZB inhibits the growth of actin filaments by capping at the barbed ends. The actin filament system plays a role in measles virus maturation [27].

EIF2B1 denotes eukaryotic translation initiation factor 2B, subunit 1 alpha. This gene might be dysregulated by several strains of virus including Hepatitis C [28], Chikungunya [29], Dengue [30], and simian immunodeficiency viruses [31]. This gene is also required for the translational suppression of vesicular stomatitis virus [32].

EIF4H denotes eukaryotic translation initiation factor 4H. This gene crucially interplays with the herpes simplex virus virion host shutoff protein [33-37].

GALR2 denotes galanin receptor 2. Galanin is an important neuromodulator in the brain, gastrointestinal system, and hypothalamopituitary axis. The rotavirus infection of the murine small intestine causes colonic secretion via galanin receptor 1 expression [38]. The mean diarrheal scores were significantly reduced in Galr1 knockout mice [38]. However, the role of galanin receptor 2 in rotavirus infection is still unknown.

RNF114 denotes ring finger protein 114. This gene plays a role in the immune responses to double-stranded RNA. RNF114 enhances the dsRNA-induced production of type I interferon [39].

UBN1 denotes ubinuclein 1. This gene encodes a member of the ubinuclein proteins that is a partner of viral *trans*-activator EB1. *UBN1* overexpression represses the Epstein-Barr Virus (EBV) productive cycle [40].

UBTF denotes upstream binding transcription factor, RNA polymerase I. This gene is up-regulated (4.5-fold change) after the medicinal treatment of chronic hepatitis C infection [41]. This gene also exhibits differential expression between susceptible and resistant strains of mice after infection with the H5N1 virus [42].

How does a virus alter gene expression via A-repeats? As a possible explanation, we suggest the following hypothesis. Influenza virus infection causes the blockade of mRNA transportation [43]. Only virus RNAs are exported from the nucleus to the cytoplasm. Selective nuclear export is a general strategy by which the virus seizes the host cells, although different viral strains use different mechanisms [44]. Subsequently, an abundance of host mRNAs with poly(A) tails is acquired in the nuclei of the infected cells [45]. According to a previous report [10], upstream A-repeats and

Argonaute proteins are identified as *cis*-regulatory elements and *trans*-acting factors, respectively. The binding activity between A-repeats and Argonaute can be reduced by transfecting peptide nucleic acid (PNA) oligonucleotides. An oligo [A(15)] is made of 15 bases of adenines in a row [10]. The synthetic PNA-A(15) competes with genomic A-repeats for the binding proteins. Thus, PNA-A(15) behaves like an inhibitor of A-repeats. The accumulated poly(A) tails in nuclei due to the virus infection provide a substantial supply of A-repeat inhibitor. As a result, the protein binding at upstream A-repeats is reduced, and the expression of their host genes is altered.

In summary, we generalized the imbalance of A-repeats and their enrichment in housekeeping genes to a wider range of mammals. A GO-based analysis identified 10 genes and 25 GO terms that are enriched with upstream long A-repeats and shared among *H. sapiens, M. musculus,* and *R. norvegicus*. Our findings demonstrate that upstream long A-repeats play crucial roles in the fundamental basis of cells; these roles can be summarized as DNA/RNA metabolism. Our findings also suggest a *cis*-regulatory role of A-repeats in response to virus infection. However, our conjectures in the discussion are based on a computational approach. Further study is required to elucidate the underlying mechanisms, e.g., involved proteins and the precise genomic locations of the functional repeats.

Funding

Research Chair Grant from the National Science and Technology Development Agency (NSTDA) of Thailand [R13/2554]; Four Seasons Hotel Bangkok's 4th Cancer Care charity fun run in coordination with the Thai Red Cross Society; and Thailand Research Fund and Office of the Higher Education Commission and Chulalongkorn University [RSA5580042]. Funding for the open access charge: [NSTDA R13/2554] and Chulalongkorn University.

References

1. Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet. 2004 Jun;5(6):435-45.
2. Birney E. Journey to the genetic interior. Interview by Stephen S. Hall. Sci Am. 2012 Oct;307(4):80-2, 84.
3. Legendre M, Pochet N, Pak T, Verstrepen KJ. Sequence-based estimation of minisatellite and microsatellite repeat variability. Genome Res. 2007 Dec;17(12):1787-96.
4. Duitama J, Zablotskaya A, Gemayel R, Jansen A, Belet S, Vermeesch JR, et al. Large-scale analysis of tandem repeat variability in the human genome. Nucleic Acids Res. 2014 May;42(9):5728-41.
5. Vinces MD, Legendre M, Caldara M, Hagihara M, Verstrepen KJ. Unstable tandem repeats in promoters confer transcriptional evolvability. Science. 2009 May 29;324(5931):1213-6.
6. Gemayel R, Vinces MD, Legendre M, Verstrepen KJ. Variable tandem repeats accelerate evolution of coding and regulatory sequences. Annu Rev Genet. 2010;44:445-77

7. Rando OJ, Winston F. Chromatin and transcription in yeast. Genetics 2012 Feb; 190(2)351-387.
8. Segal E, Widom J. Poly(dA:dT) tracts: major determinants of nucleosome organization. Curr Opin Struct Biol. 2009 Feb;19(1):65-71.
9. Nelson HCM, Finch JT, Luisi BF, Klug A. The structure of an oligo(dA)•oligo(dT) tract and its biological implications. Nature. 1987 Nov 19-25;330(6145):221-6.
10. Aporntewan C, Pin-On P, Chaiyaratana N, Pongpanich M, Boonyaratanakornkit V, Mutirangura A. Upstream mononucleotide A-repeats play a cis-regulatory role in mammals through the DICER1 and Ago proteins. Nucleic Acids Res. 2013 Oct;41(19):8872-85.
11. NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2015 Jan;43(Database issue):D6-17.
12. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25-9.
13. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, et al. NCBI GEO: archive for functional genomics data sets - 10 years on. Nucleic Acids Res. 2011 Jan;39(Database issue):D1005-10.
14. Gelfand Y, Rodriguez A, Benson G. TRDB – the Tandem Repeats Database. Nucleic Acids Res. 2007 Jan;35(Database issue):D80-7.
15. Eisenberg E, Levanon EY. Human housekeeping genes are compact. Trends Genet. 2003 Jul;19(7):362-5.
16. Liu X, Yu X, Zack D, Zhu H, Qian J. TiGER: A database for tissue-specific gene expression and regulation. BMC Bioinformatics. 2008 Jun 9;9:271.
17. Storey JD, Tibshirani R. Statistical significance for genomewide studies. PNAS. 2003 May:100(16):9440–9445.
18. Storey JD, Taylor JE, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. J R Statist Soc B. 2004 Feb;66(1):187–205.
19. Lee SM, Chan RW, Gardy JL, Lo CK, Sihoe AD, Kang SS, et al. Systems-level comparison of host responses induced by pandemic and seasonal influenza A H1N1 viruses in primary human type I-like alveolar epithelial cells in vitro. Respir Res. 2010 Oct 28;11:147.
20. Smith N, Tierney R, Wei W, Vockerodt M, Murray PG, Woodman CB, et al. Induction of interferon-stimulated genes on the IL-4 response axis by Epstein-Barr virus infected human b cells; relevance to cellular transformation. PLoS One. 2013 May 27;8(5):e64868.
21. Aporntewan C, Mutirangura A. Connection up- and down-regulation expression analysis of microarrays (CU-DREAM): a physiogenomic discovery tool. Asian Biomed. 2011 April;5(2):257–262.
22. Gardiner-Garden M, Frommer M. CpG islands in vertebrate genomes. J Mol Biol. 1987 Jul 20;196(2):261-82.
23. Akan P, Deloukas P. DNA sequence and structural properties as predictors of human and mouse promoters. Gene. 2008 Feb 29;410(1):165-76.

24. Tuntiwechapikul W, Salazar M. Mechanism of in vitro expansion of long DNA repeats: effect of temperature, repeat length, repeat sequence, and DNA polymerases. Biochemistry. 2002 Jan 22;41(3):854-60.
25. Amos W, Filipe LN. Microsatellite frequencies vary with body mass and body temperature in mammals, suggesting correlated variation in mutation rate. PeerJ. 2014 Nov 6;2:e663.
26. Qin L, Ma Y, Liang P, Tan Z, Li S. Differential distributions of mononucleotide repeat sequences in 256 viral genomes and its potential implications. Gene. 2014 Jul 10;544(2):159-64.
27. Dietzel E, Kolesnikova L, Maisner A. Actin filaments disruption and stabilization affect measles virus maturation by different mechanisms. Virol J. 2013 Aug 2;10:249.
28. Caillot F, Hiron M, Goria O, Gueudin M, Francois A, Scotte M, et al. (2009) Novel serum markers of fibrosis progression for the follow-Up of hepatitis C virus-infected patients. Am J Pathol. 2009 Jul;175(1):46-53.
29. Thio CL, Yusof R, Abdul-Rahman PS, Karsani SA. Differential proteome analysis of Chikungunya virus infection on host cells. PLoS One. 2013 Apr 10;8(4):e61444.
30. Villas-Bôas CS, Conceição TM, Ramírez J, Santoro AB, Da Poian AT, Montero-Lomelí M. Dengue virus-induced regulation of the host cell translational machinery. Braz J Med Biol Res. 2009 Nov;42(11):1020-6.
31. Ndolo T, George M, Nguyen H, Dandekar S. Expression of simian immunodeficiency virus Nef protein in CD4+ T cells leads to a molecular profile of viral persistence and immune evasion. Virology. 2006 Sep 30;353(2):374-87.
32. Elsby R, Heiber JF, Reid P, Kimball SR, Pavitt GD, Barber GN. The alpha subunit of eukaryotic initiation factor 2B (eIF2B) is required for eIF2-mediated translational suppression of vesicular stomatitis virus. J Virol. 2011 Oct;85(19):9716-25.
33. Everly DN Jr, Feng P, Mian IS, Read GS. mRNA degradation by the virion host shutoff (Vhs) protein of herpes simplex virus: genetic and biochemical evidence that Vhs is a nuclease. J Virol. 2002 Sep;76(17):8560-71.
34. Doepker RC, Hsu WL, Saffran HA, Smiley JR. Herpes simplex virus virion host shutoff protein is stimulated by translation initiation factors eIF4B and eIF4H. J Virol. 2004 May;78(9):4684-99.
35. Feng P, Everly DN Jr, Read GS. mRNA decay during herpes simplex virus (HSV) infections: protein-protein interactions involving the HSV virion host shutoff protein and translation factors eIF4H and eIF4A. J Virol. 2005 Aug;79(15):9651-64.
36. Sarma N, Agarwal D, Shiflett LA, Read GS. Small interfering RNAs that deplete the cellular translation factor eIF4H impede mRNA degradation by the virion host shutoff protein of herpes simplex virus. J Virol. 2008 Jul;82(13):6600-9.
37. Page HG, Read GS. The virion host shutoff endonuclease (UL41) of herpes simplex virus interacts with the cellular cap-binding complex eIF4. J Virol. 2010 Jul;84(13):6886-90.
38. Hempson SJ, Matkowskyj K, Bansal A, Tsao E, Habib I, Benya R, et al. Rotavirus infection of murine small intestine causes colonic secretion via age restricted galanin-1 receptor expression. Gastroenterology. 2010 Jun;138(7):2410-7.

39. Bijlmakers MJ, Kanneganti SK, Barker JN, Trembath RC, Capon F. Functional analysis of the RNF114 psoriasis susceptibility gene implicates innate immune responses to double-stranded RNA in disease pathogenesis. Hum Mol Genet. 2011 Aug 15;20(16):3129-37.
40. Gruffat H, Lupo J, Morand P, Boyer V, Manet E. The nuclear and adherent junction complex component protein ubinuclein negatively regulates the productive cycle of Epstein-Barr virus in epithelial cells. J Virol. 2011 Jan;85(2):784-94.
41. Grinde B, Hetland G, Johnson E. Effects on gene expression and viral load of a medicinal extract from Agaricus blazei in patients with chronic hepatitis C infection. Int Immunopharmacol. 2006 Aug;6(8):1311-4.
42. Boon AC, deBeauchamp J, Hollmann A, Luke J, Kotb M, Rowe S, et al. Host genetic variation affects resistance to infection with a highly pathogenic H5N1 Influenza A virus in mice. J. Virol. 2009 Oct;83(20):10417-26.
43. Chen Z, Krug RM. Selective nuclear export of viral mRNAs in influenza-virus-infected cells. Trends Microbiol. 2000 Aug;8(8):376-83.
44. Rozanne MS. Viral regulation of mRNA export. J Virol. 2004 May;78(9):4389-96.
45. Rubio RM, Mora SI, Romero P, Arias CF, López S. Rotavirus prevents the expression of host responses by blocking the nucleocytoplasmic transport of polyadenylated mRNAs. J Virol. 2013 Jun;87(11):6336-45.

Supporting Information

S1 Fig. Counting and normalization methods. A bin is defined as a genomic region of 800 bp. Given a set of genes, the number of repeats in each bin is the sum of base pairs that fall into the bin. The number of repeats is normalized by dividing by the total number of genes. Finally, the number of repeats in each bin is measured in the unit of base pairs per gene.

S2 Fig. Genome-wide distribution of A- and T-repeats in 12 non-mammals. The repeats with lengths of 10 to 30 bp located around the TSSs throughout the entire genome are shown. **(A,B)** The horizontal axis consists of 25 bins. The vertical axis represents the normalized number of repeats in base pairs per gene. **(C)** An unpaired *t*-test compares the number of repeats between the upstream (bins 1 to 10) and downstream (bins 16 to 25) regions, yielding the *p*-values as summarized in Table 1A.

S3 Fig. Comparison of the number of repeats between housekeeping (HK) and tissue-specific genes (TS) in 7 non-mammals. The repeats in the upstream (bins 1 to 10) and downstream (bins 16 to 25) regions between HK and TS genes are compared. The paired *t*-test *p*-values are summarized in Table 1B. **(A)** Short repeats (2 to 9 bp). **(B)** Long repeats (10 to 30 bp).

S1 Table. The selected 20 living organisms from NCBI database. Only the organisms with more than 1,000 known protein-coding transcripts were suitable for our analysis. In addition, the transcript status must be ‘reviewed’, ‘validated’ or ‘provisional’.

S2 Table. The selected 2,403 Gene Ontology (GO) terms. Only GO terms with at least 20 genes were selected.

S3 Table. CU-DREAM parameters. The software CU-DREAM requires these parameters to analyze the microarray datasets GSE24533 and GSE45829. CU-DREAM classifies genes into 3 classes: down-regulation, up-regulation, and unchanged expression.

S4 Table. The top 500 homologous genes. These genes were enriched with upstream long A-repeats (10 to 30 bp) in *H. sapiens*, *M. musculus*, and *R. norvegicus*. The genes in the contingency tables of Fig. 4 are completely listed.

S5 Table. The intersection of the top 500 homologous genes. These genes were enriched with upstream long A-repeats (10 to 30 bp) in *H. sapiens*, *M. musculus*, and *R. norvegicus*. The intersected genes are completely listed. H, M, and R denote the sets of the top 500 genes in *H. sapiens, M. musculus*, and *R. norvegicus*, respectively.

S6 Table. The intersection of the significant Gene Ontology (GO) terms. These GO terms were enriched with upstream long A-repeats (10 to 30 bp) in *H. sapiens*, *M. musculus*, and *R. norvegicus*. The intersected GO terms are completely listed. *H*, *M,* and *R* denote the sets of the significant GO terms in *H. sapiens*, *M. musculus*, and *R. norvegicus*, respectively. A total of 2,403 GO terms included in the study are completely listed with *P*-value and *q*-values.

S7 Table. The complete list of genes. The transcripts of these genes were down-regulated, up-regulated, and unchanged expression in the microarray datasets GSE24533 and GSE45829.

Fig 1.

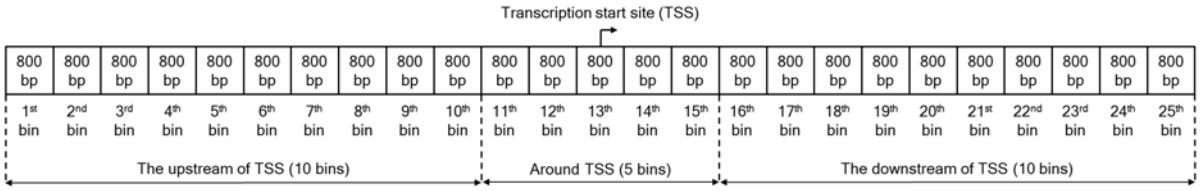

Transcription start site (TSS)
800 bp
1st bin
2nd bin
3rd bin
4th bin
5th bin
6th bin
7th bin
8th bin
9th bin
10th bin
11th bin
12th bin
13th bin
14th bin
15th bin
16th bin
17th bin
18th bin
19th bin
20th bin
21st bin
22nd bin
23rd bin
24th bin
25th bin
The upstream of TSS (10 bins)
Around TSS (5 bins)
The downstream of TSS (10 bins)

Fig 2.

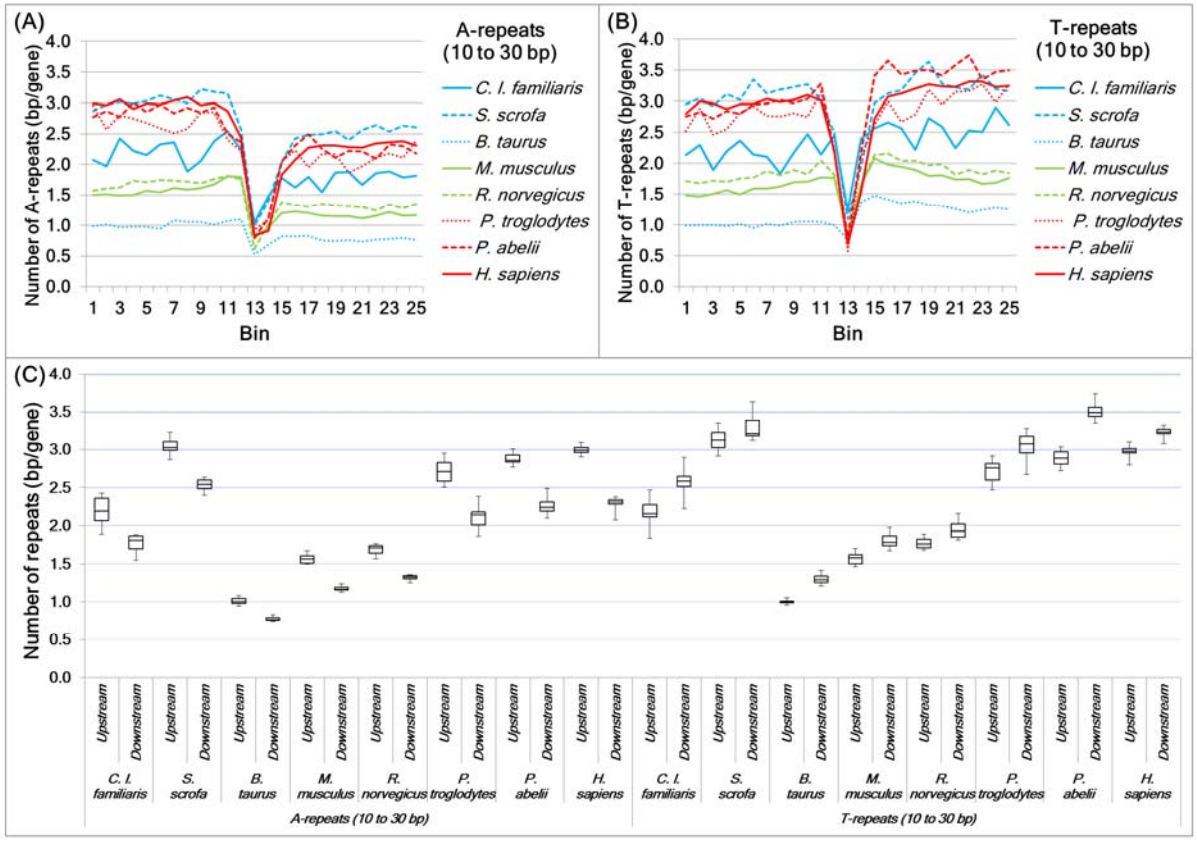

(A)
Number of A-repeats (bp/gene)
A-repeats (10 to 30 bp)
C. l. familiaris
S. scrofa
B. taurus
M. musculus
R. norvegicus
P. troglodytes
P. abelii
H. sapiens
Bin
(B)
Number of T-repeats (bp/gene)
T-repeats (10 to 30 bp)
(C)
Number of repeats (bp/gene)
Upstream
Downstream
A-repeats (10 to 30 bp)
T-repeats (10 to 30 bp)

Fig 3.

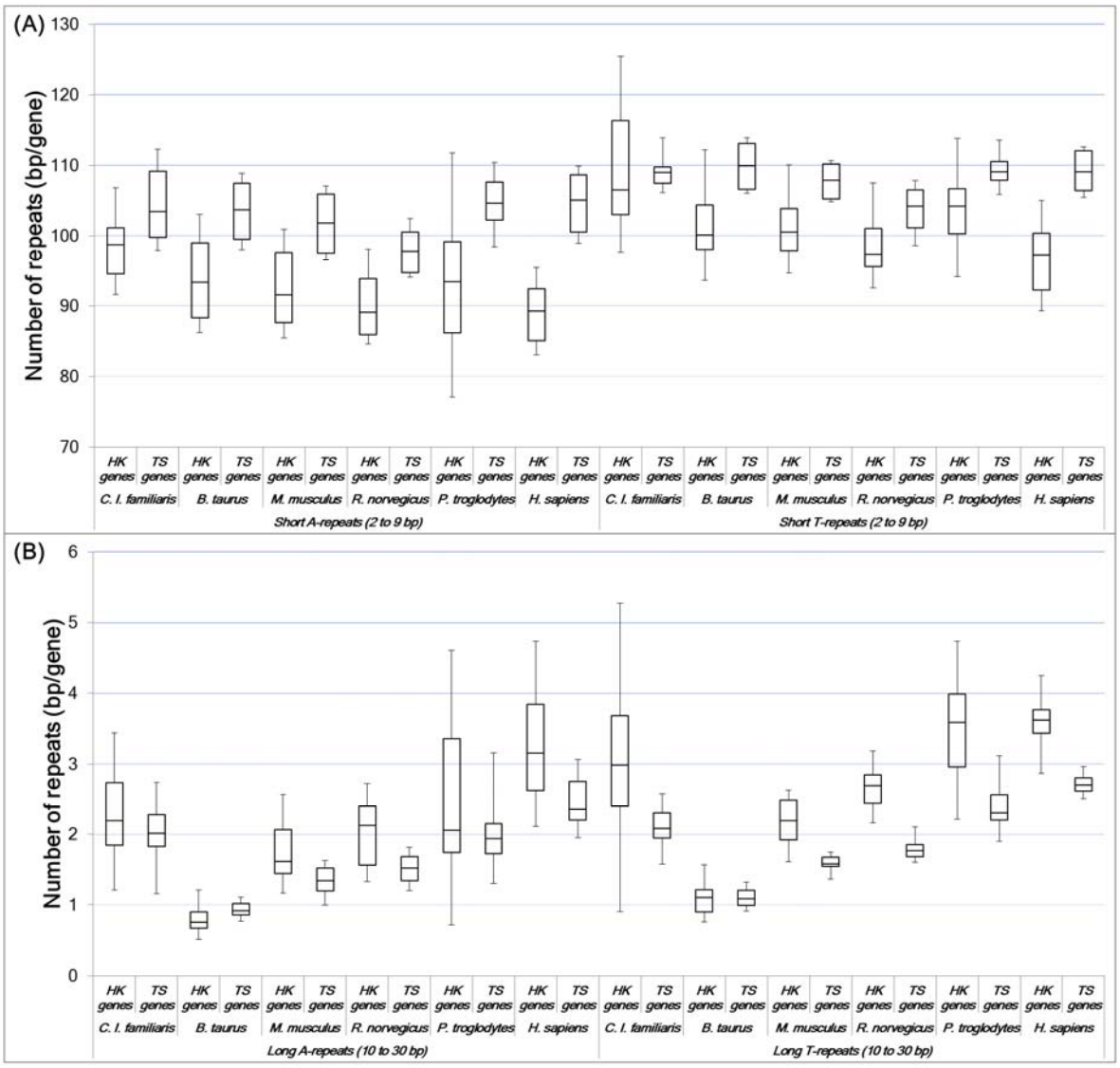

(A)
Number of repeats (bp/gene)
130
120
110
100
90
80
70
HK genes
TS genes
C. l. familiaris
B. taurus
M. musculus
R. norvegicus
P. troglodytes
H. sapiens
Short A-repeats (2 to 9 bp)
Short T-repeats (2 to 9 bp)
(B)
Number of repeats (bp/gene)
6
5
4
3
2
1
0
Long A-repeats (10 to 30 bp)
Long T-repeats (10 to 30 bp)

Fig 4.

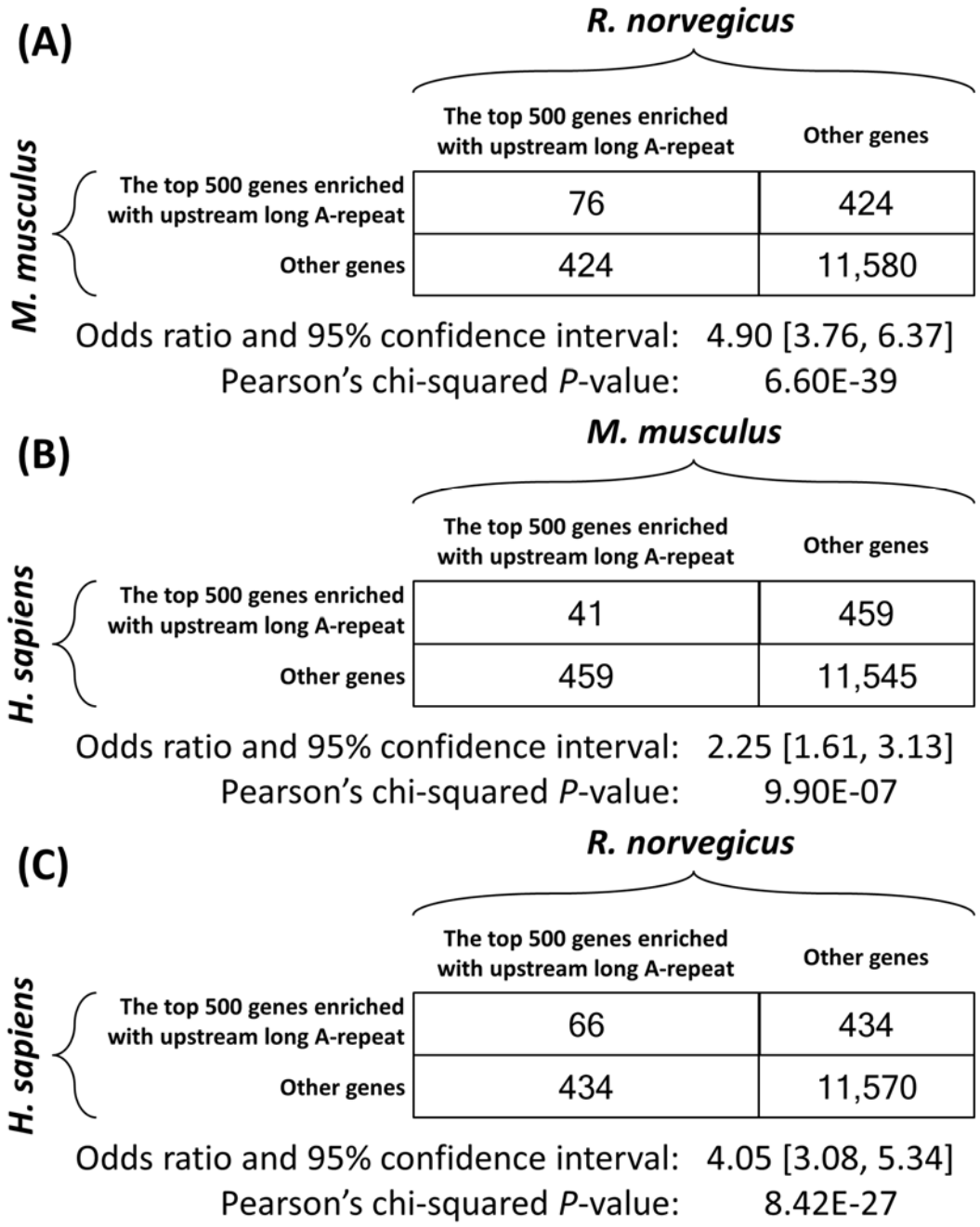
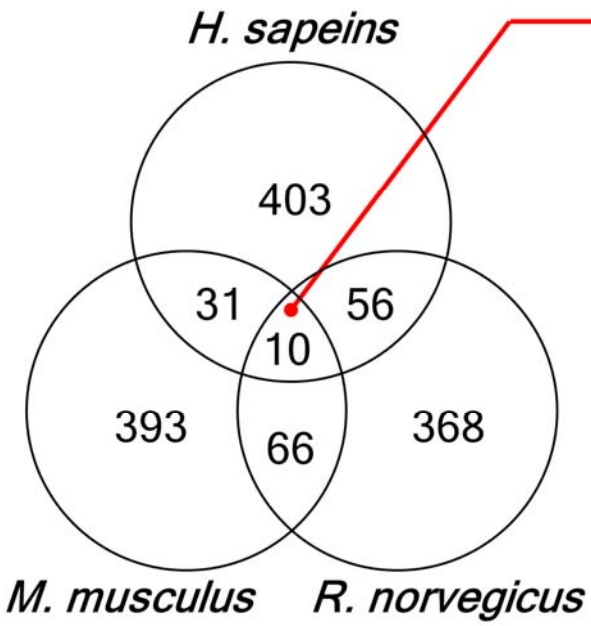
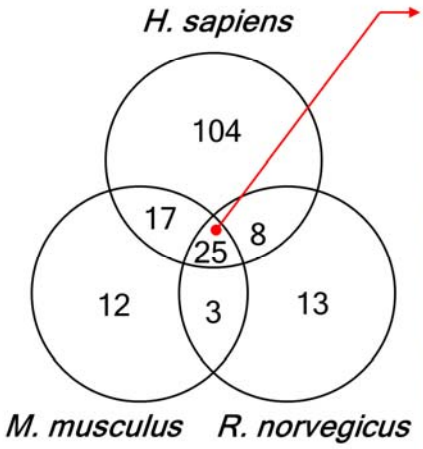

(A)
R. norvegicus
The top 500 genes enriched with upstream long A-repeat
Other genes
M. musculus
The top 500 genes enriched with upstream long A-repeat
76
424
Other genes
424
11,580
Odds ratio and 95% confidence interval: 4.90 [3.76, 6.37]
Pearson's chi-squared P-value: 6.60E-39
(B)
M. musculus
The top 500 genes enriched with upstream long A-repeat
Other genes
H. sapiens
The top 500 genes enriched with upstream long A-repeat
41
459
Other genes
459
11,545
Odds ratio and 95% confidence interval: 2.25 [1.61, 3.13]
Pearson's chi-squared P-value: 9.90E-07
(C)
R. norvegicus
The top 500 genes enriched with upstream long A-repeat
Other genes
H. sapiens
The top 500 genes enriched with upstream long A-repeat
66
434
Other genes
434
11,570
Odds ratio and 95% confidence interval: 4.05 [3.08, 5.34]
Pearson's chi-squared P-value: 8.42E-27

Fig 5.



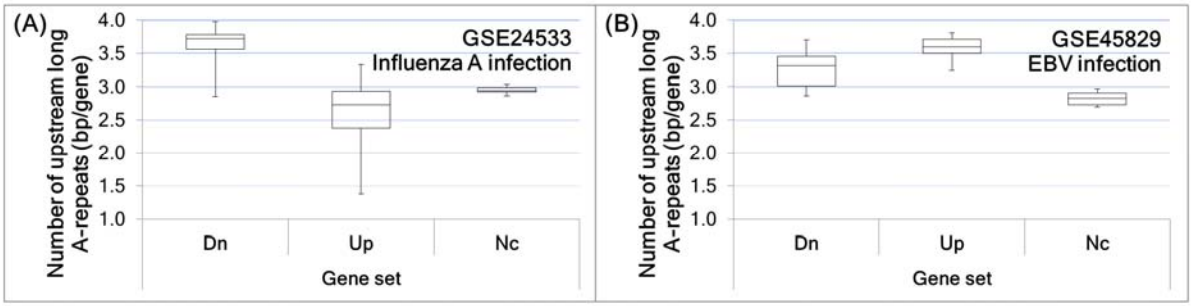
Gene ID	Gene Symbol
832	*CAPZB*
1967	*EIF2B1*
7458	*EIF4H*
8811	*GALR2*
56984	*PSMG2*
55905	*RNF114*
788	*SLC25A20*
55815	*TSNAXIP1*
29855	*UBN1*
7343	*UBTF*

Fig 6.



GO ID	Domain	Title
GO:0000122	Biological process	Negative regulation of transcription from RNA polymerase II promoter
GO:0000278	Biological process	Mitotic cell cycle
GO:0006351	Biological process	Transcription, DNA-templated
GO:0006355	Biological process	Regulation of transcription, DNA-templated
GO:0006614	Biological process	SRP-dependent cotranslational protein targeting to membrane
GO:0006974	Biological process	Cellular response to DNA damage stimulus
GO:0008380	Biological process	RNA splicing
GO:0010467	Biological process	Gene expression
GO:0016032	Biological process	Viral process
GO:0016070	Biological process	RNA metabolic process
GO:0019058	Biological process	Viral life cycle
GO:0019083	Biological process	Viral transcription
GO:0005634	Cellular component	Nucleus
GO:0005654	Cellular component	Nucleoplasm
GO:0005730	Cellular component	Nucleolus
GO:0005737	Cellular component	Cytoplasm
GO:0005739	Cellular component	Mitochondrion
GO:0005829	Cellular component	Cytosol
GO:0003677	Molecular function	DNA binding
GO:0003723	Molecular function	RNA binding
GO:0003735	Molecular function	Structural constituent of ribosome
GO:0004842	Molecular function	Ubiquitin-protein ligase activity
GO:0005515	Molecular function	Protein binding
GO:0008270	Molecular function	Zinc ion binding
GO:0044822	Molecular function	Poly(A) RNA binding

Fig 7.


(A)
Number of upstream long A-repeats (bp/gene)
4.0
3.5
3.0
2.5
2.0
1.5
1.0
GSE24533
Influenza A infection
Dn
Up
Nc
Gene set
(B)
Number of upstream long A-repeats (bp/gene)
4.0
3.5
3.0
2.5
2.0
1.5
1.0
GSE45829
EBV infection
Dn
Up
Nc
Gene set