8872–8885 *Nucleic Acids Research, 2013, Vol. 41, No. 19* *Published online 8 August 2013*
doi:10.1093/nar/gkt685

Upstream mononucleotide A-repeats play a *cis*-regulatory role in mammals through the DICER1 and Ago proteins

Chatchawit Aporntewan[1], Piyapat Pin-on[2], Nachol Chaiyaratana[3,4], Monnat Pongpanich[1], Viroj Boonyaratanakornkit[5] and Apiwat Mutirangura[6,*]

[1]Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok 10330, Thailand, [2]Inter-Department Program of Biomedical Sciences, Faculty of Graduate School, Chulalongkorn University, Bangkok 10330, Thailand, [3]Department of Electrical and Computer Engineering, Faculty of Engineering, King Mongkut's University of Technology North Bangkok, Bangkok 10800, Thailand, [4]Division of Molecular Genetics, Department of Research and Development, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok 10700, Thailand, [5]Department of Clinical Chemistry, Faculty of Allied Health Sciences, Chulalongkorn University, Bangkok 10330, Thailand and [6]Center of Excellence in Molecular Genetics of Cancer and Human Diseases, Department of Anatomy, Faculty of Medicine, Chulalongkorn University, Bangkok 10330, Thailand

Received February 27, 2013; Revised July 11, 2013; Accepted July 13, 2013

ABSTRACT

A-repeats are the simplest form of tandem repeats and are found ubiquitously throughout genomes. These mononucleotide repeats have been widely believed to be non-functional 'junk' DNA. However, studies in yeasts suggest that A-repeats play crucial biological functions, and their role in humans remains largely unknown. Here, we showed a non-random pattern of distribution of sense A- and T-repeats within 20 kb around transcription start sites (TSSs) in the human genome. Different distributions of these repeats are observed upstream and downstream of TSSs. Sense A-repeats are enriched upstream, whereas sense T-repeats are enriched downstream of TSSs. This enrichment directly correlates with repeat size. Genes with different functions contain different lengths of repeats. In humans, tissue-specific genes are enriched for short repeats of $<$10 bp, whereas housekeeping genes are enriched for long repeats of $\geq$10 bp. We demonstrated that DICER1 and Argonaute proteins are required for the *cis*-regulatory role of A-repeats. Moreover, in the presence of a synthetic polymer that mimics an A-repeat, protein binding to A-repeats was blocked, resulting in a dramatic change in the expression of genes containing upstream A-repeats. Our findings suggest a length-dependent *cis*-regulatory function of A-repeats and that Argonaute proteins serve as *trans*-acting factors, binding to A-repeats.

INTRODUCTION

A microsatellite or a tandem repeat (TR) is a concatenation of the same nucleotide sequence, called a unit. In other words, a TR is a repeat of the same unit of nucleotides from the beginning to the end of the repeat (1,2). For example, 'AAAAA' represents five repeats of 'A', whereas 'CATCATCATCAT' represents four repeats of 'CAT'. Traditionally, these repeats were believed to be generated by DNA replication slippage and to have no function, and they were called 'junk' or 'selfish' DNA (3). TRs have a propensity for evolvability because there is a high degree of variation within the TRs among related species (4). Repeat variation can be measured in terms of repeat size and sequence similarity (5). TRs are found ubiquitously in both coding and non-coding regions. In coding regions, TRs enable functional variability among genes. In non-coding regions, specifically within gene promoters, repeat variability correlates with variations in gene expression (4,6). This diversity of expression can produce phenotypic variants. Several lines of evidence showing phenotypic variations due to TRs have been reviewed (2). The evolvability of gene modulation is vital for coping with environmental changes and for the emergence of new species. TRs in promoters mediate transcription in

*To whom correspondence should be addressed. Tel: +66 2256 4281 (Ext 1713); Fax: +66 2256 4281 (Ext 1713); Email: Apiwat.Mutirangura@gmail.com

several ways (2). First, repeat units may serve as binding sites for transcription factors. The number of binding sites determines the rate of transcription. Second, the expansion and shrinkage of TRs can change the distance between two functional elements. Third, TRs can affect chromatin structure and consequently mediate transcription (2). A correlation between TR enrichment and nucleosome-depleted regions suggests that TRs mediate transcription by inhibiting nucleosome formation (7). Finally, frequent deletions of mononucleotide repeats in 3′ or 5′ untranslated regions (UTRs) were observed in tumors with microsatellite instability. This observation also suggested that mononucleotide repeats in 3′ or 5′ UTRs may perform specific functions (8). Currently, it is accepted that the number of units in trinucleotide repeats in both coding and non-coding regions is a crucial factor in the development of neurodegenerative diseases and certain phenotypic traits (2).

Mononucleotide repeats are the simplest class of TRs. In eukaryotes, poly(dA:dT) tracts are ubiquitously distributed throughout the entire genome (9) (Supplementary Figure S1). Extensive studies in yeasts suggest that these non-coding repeats may perform crucial biological functions (10). Poly(dA:dT) tracts are correlated with nucleosome-depleted regions in yeasts (7,11) and in humans (12,13). Moreover, these nucleosome-depleted tracts are evolutionarily conserved among four species of yeast (14). It is hypothesized that one intrinsic property of poly(dA:dT) is to resist sharp DNA bending (15). Thus, poly(dA:dT) tracts within gene promoters can block nucleosome formation and increase transcription factor accessibility. A recent study showed that gene transcription can be fine-tuned by varying poly(dA:dT) tract length and continuity (16). However, transcriptional regulation is a dynamic and competitive process involving nucleosomes, chromatin structure and transcription factors (17).

Although the functional role of non-coding poly(dA:dT) is well established, the mechanism underlying this function remains largely unknown. In addition to the theories about the intrinsic properties of poly(dA:dT), it is believed that these poly(dA:dT) tracts may serve as *cis*-regulatory elements or binding sites for *trans*-acting factors. Protein complexes that form with a certain repeat sequence may regulate specific biological functions. However, no *trans*-acting poly(dA:dT) binding proteins have been reported to date. Currently, it is well accepted that small RNAs are key players in target recognition. In addition, small RNAs can play a regulatory role in controlling gene expression (18). The discovery of RNA interference (RNAi), for which the 2006 Nobel Prize in Physiology was awarded, suggests that small RNAs play important roles in epigenetics (19,20). RNAi is characterized by the binding of a small interfering RNA to a messenger RNA (mRNA), which targets that mRNA for degradation. As a result, the corresponding gene is downregulated. During the first step of the RNAi pathway, double-stranded RNAs or pre-microRNAs are cleaved by the Dicer protein into small double-stranded RNA fragments (20–25 bp). Second, a single-stranded RNA is selected by Argonaute proteins and then loaded onto an RNA-induced silencing complex (RISC). Third, the RISC complex binds to the target mRNA by recognizing its complementary sequence. In contrast to RNAi, which functions at the post-transcriptional level, promoter targeting by small RNAs may either silence or activate gene transcription (21–24). Argonaute is a family of proteins (25,26). In humans, members of the Argonaute family are evolutionarily conserved and can be subdivided into the Ago and the Piwi subfamilies. Only Ago proteins are expressed ubiquitously, and they cooperate with small RNAs for target recognition. Piwi proteins are expressed exclusively in the germline. The Ago protein family consists of four members: AGO1, AGO2, AGO3 and AGO4. Broadly, small RNAs serve as components of a cellular surveillance system. Cells produce small RNAs to help maintain the overall epigenetic state of the genome.

The research question herein arose from our observation of sense A-repeats upstream of transcription start sites (TSSs). We observed that sense A-repeats are often more enriched upstream than downstream in humans and mice, but not in yeasts. Although studies on poly(dA:dT) tracts are largely conducted in yeasts, our observations suggest that A-repeats may possess regulatory functions distinct from those found in yeasts. Therefore, we set out to investigate three specific aims. First, we aimed to demonstrate that A-repeats are *cis*-regulatory elements and correlate with gene expression. Second, we aimed to identify the corresponding *trans*-acting factors, with Dicer and members of the Ago family as our candidate proteins. Third, we incorporated information from several public databases in our experiments, allowing us to perform an integrated genome-wide analysis of human sequence, expression and gene regulation data.

To explore the role of mononucleotide repeats in humans, we performed a computational analysis by integrating data from a number of relevant databases, including whole-genome sequences (27), Gene Expression Omnibus (GEO) data sets (28) and Ago-binding sites (29). Six model organisms were used in our analysis: *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Rattus norvegicus*, *Mus musculus* and *Homo sapiens* (Supplementary Table S1). The distribution frequency of mononucleotide repeats has previously been investigated by counting poly(dA:dT) tracts in double-stranded DNA (9). However, this previously used counting method might not be well suited to uncovering biological functions because it does not reflect the imbalance of A- and T-repeats between the two DNA strands. In this study, poly(dA:dT) tracts were counted separately as sense A- and sense T-repeats relative to the TSS.

MATERIALS AND METHODS

UCSC genome browser database

The organisms included in our analysis are listed in Supplementary Table S1. Their whole genomes were downloaded from the UCSC Genome Browser Database (27). We used human genome build 36 (hg18) because it is

compatible with the CLIPZ database (29). Sequences 10 000 bp upstream and 10 000 bp downstream of a TSS were extracted from the whole genomes for further statistical analyses.

CLIPZ database

The CLIPZ database lists all the known binding sites of Ago proteins in the whole genome of human embryonic kidney (HEK)-293 cells (29). The database contains two important files:

mapped_sequences	RNA sequences bound by Argonaute (AGO1-4).
genome_mappings	The locations of these RNA sequences, mapped to the whole genome. The mapping begins at chromosome 1, and the mapping is stopped for RNA sequences that can be mapped to >30 locations (mostly repeat sequences).

The Ago protein family members are AGO1, AGO2, AGO3 and AGO4. We downloaded the following files from http://test.mirz.unibas.ch/smirnaWeb/geneBio/smiRNA/temp/10544043421949953483/samples in the following subfolders (October, 2011):

AGO1:	/230/mapped_sequences, /230/genome_mappings
AGO2:	/238/mapped_sequences, /238/genome_mappings
AGO3:	/239/mapped_sequences, /239/genome_mappings
AGO4:	/240/mapped_sequences, /240/genome_mappings

The numbers of mononucleotide repeats bound by Ago proteins are listed in Supplementary Table S2.

The Ago-bound sequences from CLIPZ are highly redundant. For instance, three reads, AAAC, AAACG, AAACGT, are sequenced from the same source. In this case, AAA is counted too many times. To remove any possible bias, we excluded sequence reads that are contained in other reads. The exclusion method is illustrated in Supplementary Figure S2. This exclusion strategy will eliminate the argument that Ago-bound A-repeats are actually poly-A tails, which are numerous. If the poly-A tails argument is valid, only the single longest A-repeat will be counted because the other repeats will be excluded.

Housekeeping and tissue-specific genes

A total of 575 housekeeping genes in the human genome were identified by Eisenberg and Levanon (30), and 7261 tissue-specific genes in the human genome were identified in the Tissue-specific Gene Expression and Regulation database (31). The list of housekeeping genes was downloaded from http://www.compugen.co.il/supp_info/Housekeeping_genes.html (October, 2011). The list of tissue-specific genes was downloaded from http://bioinfo.wilmer.jhu.edu/tiger/download/ref2tissue-Table.txt (October, 2011). The two gene sets contain 122 overlapping genes.

GEO

We searched for microarray experiments that involved DICER1 knockdown (KD) and AGO1-4 KD in the GEO database (28) and found that microarray experiment GSE4246 used the same HEK-293 cell line (32). Selected experiments and samples are listed in Supplementary Table S3. Both up- and downregulated genes were identified using our software, called CU-DREAM (33). We classified samples into experimental and control groups and performed Student's *t*-test on each probe. The significance threshold was set at $P < 0.01$. Transcripts with significantly higher or lower means of expression in the experimental group compared with those in the control group were considered up- or downregulated, respectively. Transcripts without significant differences between experimental and control groups were considered neutral (neither up- nor downregulated).

Statistical methods

The imbalance between repeat enrichment upstream and downstream of TSSs was determined using Student's *t*-test. The first 10 bins (bin 1–10) represent the sequence 2001–10 000 bp upstream, whereas the last 10 bins (bin 16–25) represent the sequence 2001–10 000 bp downstream of the TSS. The five middle bins (bin 11–15) were not analyzed because the numbers of A- and T-repeats drop sharply in the immediate vicinity of the TSS. The number of repeats was calculated as described in Supplementary Figure S3. Finally, an unpaired *t*-test was conducted between the numbers of repeats in the first and the last 10 bins.

Based on the results of microarray experiments, we divided genes into three groups, downregulated (Dn), upregulated (Up) and non-regulated (Nu - neutral). Within a specific region around the TSS, we expected to see the mean difference between the amount of A-repeats in the first (Dn or Up) and second (Nu) sets. Next, a permutation test was used to determine the statistical significance of the original mean difference (34). Every gene was labeled with '1st' or '2nd', indicating the first or the second group. In each replicate, the labels were randomly shuffled, and then the mean difference was recalculated. A total of 1000 replicates were performed. The permutation *P*-value is defined as the number of times that the replicated mean difference was greater than or equal to the original mean difference divided by the total number of replicates (1000). If the dividend equals zero, then the permutation *P*-value was considered <0.001.

There are a total of 44 hypotheses per microarray experiment, $|\{\text{Length} = 1, \text{Length} = 15 \text{ to } 30\} \times \{\text{Dn versus Nu, Up versus Nu}\} \times \{\text{bin 1 to 10, bin 1, bin 2, bin 3, bin 4, bin 5, bin 6, bin 7, bin 8, bin 9, bin 10}\}| = 44$. Multiple hypothesis correction was performed using false-discovery rate (FDR) analysis (35). The QVALUE package for R statistical software was used. All default settings were maintained except that the range of λ was set at [0, 0.5], stepped by 0.05, and the bootstrap option was selected instead of the smoother option for the estimation of π_0 (36). The obtained $\hat{\pi}_0 \ll 1$, suggesting that the number of significant *P*-values is high. When the *q*-value was restricted to ≤ 0.05, the number of significant *P*-values was 82.05% (361/440). A more stringent restriction of *q*-value to ≤ 0.01 yields 7.95% (35/440).

We presented fold change instead of mean difference in the figures because fold changes can be compared without additional normalization. The fold change is defined as the ratio between the number of A-repeats (number of bp per gene) in the first and the second groups, respectively. The numerical fold change, *P*-value and *q*-value data are provided in Supplementary Table S4.

A(15) inhibitor transfection and microarray

HEK-293 cells were grown and maintained in Dulbecco's modified Eagle's medium (Gibco-BRL) supplemented with 2 mM L-glutamine, 10% heat-inactivated fetal calf serum and 10 mg/ml antibiotic/antimycotic (Invitrogen) before and after transfection. A peptide nucleic acid (PNA) oligo containing a long A-repeat sequence [A(15)] (37) was used to inhibit AGO binding to A-repeats. Here, PNA oligos were modified by adding an 8-amino-3, 6-dioxaoctanoic acid to their 5′ ends. Duplicate sets of HEK-293 cells were transfected with either PNA-A(15) or scramble (control) PNA oligo (PNA-ACgTTCg CgCAACgA) at 50 nM using the TransIT-siQUEST transfection reagent (Mirus). At 48 h after transfection, RNAs were extracted and purified using Trizol (Invitrogen) according to the manufacturer's protocol. cDNAs were prepared according to the manufacturer-recommended protocols (Affymetrix and NuGEN). Labeled cDNAs were hybridized to Affymetrix Human Gene 1.0 ST arrays for 18 h at 45°C with rotation for 18 h. The arrays were then washed and stained using FS450_0007 fluidics protocol and scanned with Affymetrix Gene ChIP Scanner 3000 7G. Scanned images were inspected for hybridization efficiency, and the data were converted to expression values from hybridization efficiency intensity to expression values. CEL files from GeneChip Operating Software were imported into expression console (EC) 1.2 software for array quality control (QC). The matrix was created by Affymetrix Power Tools—Release 1.14.3 on Windows 7. The command used was 'aptprobeset-summarize -a rma -d HuGene-1_0-st-v1.r3.cdf -o out –cel-files cel_list.txt'. The probe group file was HuGene-1_0-st-v1.r4.pgf, and the meta-probe-set file was HuGene-1_0-st-v1.r4.mps. Data obtained from two independent array hybridization experiments were uploaded into Analyst from GeneChip® Operating Software (Genedata AG; Basel, Switzerland) and normalized simultaneously. Expression values were estimated using the GC-RMA algorithm provided by Genedata. Statistical analysis was performed using Analyst. Genes were required to pass an N-way ANOVA with a $P < 0.05$ and/or have a median fold change of ≥ 1.5 between one or more pairs of conditions. All original microarray data were deposited in the NCBI GEO database (series record GSE43185) (Supplementary Table S3). The scanned images were inspected for hybridization efficiency, and CEL files generated from GeneChip Operating Software were imported into EC 1.2 software for array QC. RMA normalization was performed to generate the QC metrics that we routinely use to determine data quality. These include perfect match mean (PM_Mean), background mean (Bgd_Mean), positive and negative probes (POS versus NEG AUC), bacterial spike controls and polyA controls.

Chromatin immunoprecipitation and PCR

For chromatin immunoprecipitation (ChIP) assays, human HEK-293 cells were treated with PNA-A(15) and/or scramble sequence and grown in a 75 cm^2 flask to 80% confluence. The cells were harvested, and ChIP assays were carried out as previously described (38). Chromatin fragments were immunoprecipitated with anti-AGO2 monoclonal antibody (SC-32659, Santa Cruz Biotechnology) or control non-immunized goat antibody (SC-2028, Santa Cruz). Immunoprecipitated DNA fragments were analyzed by PCR amplification and DNA electrophoresis.

Oligos

Two sets of oligos for AGO2-bound long A-repeat locations:

(1) AGO2+, A-repeat+
NM_006068 AAggTTgTggATTCAAAgggA and TTTTA AAgCAATAATTTCTCCCATCT
(2) AGO2+, A-repeat+
NM_005216 TCTAAgCTCAgTggCAAgACCTA and A AAAACAACCACCACCACCCATg.

Two sets of oligos for AGO2-bound non-A-repeat locations:

(1) AGO2+, A-repeat-
NM_007225 ACgCTggCATgggAAAACCAAg and ACT TCTACCgAgTgCTCCTTAgA
(2) AGO2+, A-repeat-
NM_005481 TgTTgTATATgTgTgCgCgCgT and ATAA AACCggCTCTTAggACCgT.

One set of oligos for AGO2-unbound sequences:

AGO2-, A-repeat-
NM_001143943 gCCTAATCAgCAAATTAggCA and T TTTTATATACCCACACTACCTAg

RESULTS

A-repeats are not randomly distributed around TSSs

The distribution of sense A- and T-repeats within the 10 000 bp upstream and downstream of TSSs was examined. A total sequence of 20 000 bp was divided into 25 bins of 800 bp each. The TSS was centered in the 13th bin; lower number bins contain upstream sequences, and higher number bins contain downstream sequences (Figure 1A). The distribution of sense A- and T-repeats is non-random (Figure 1, Supplementary Figures S4–S9). The counting method is illustrated in Supplementary Figure S3. In invertebrates and yeast (Figure 1B–D), sense A-repeats are clearly enriched at the TSS, except in *C. elegans*, in which T-repeats are enriched at TSS but depleted in the 14th bin, immediately downstream of the TSS. In mammals (Figure 1E–G), the distribution of repeats drops sharply at the TSS. Most strikingly, the distribution of A- and T-repeats upstream and downstream of the TSS is not

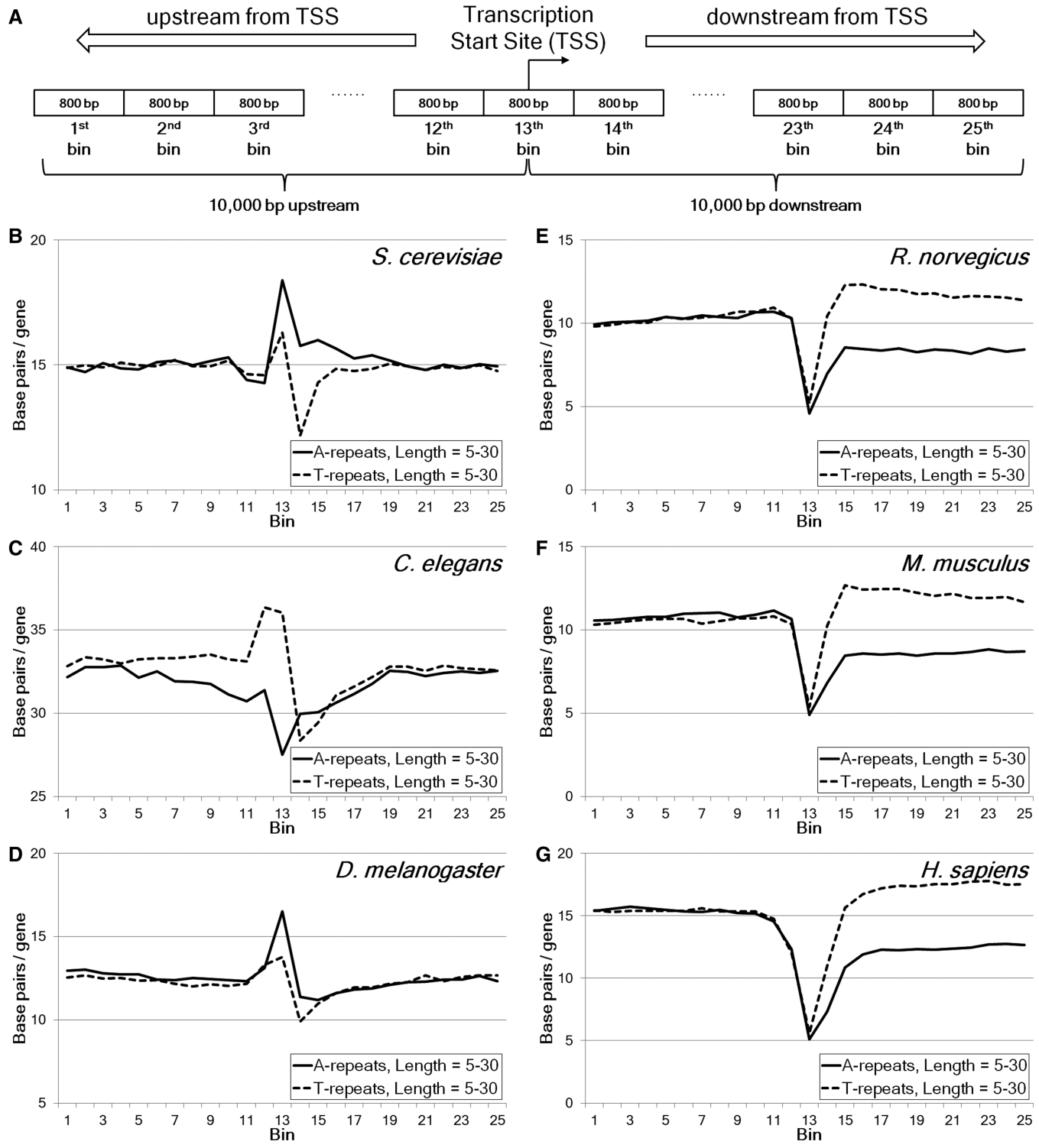

Figure 1. Distributions of sense A- and T-repeats around TSSs. (**A**) Bin structure around TSS. There are 25 bins. Each bin covers 800 bp, and 10 000 bp upstream and 10 000 bp downstream of the TSS were analyzed. The TSS is centered in the 13th bin. (**B–G**) Distributions of sense A- and T-repeats. Repeats with a length of 5–30 bp around TSSs in the whole genome are shown. The horizontal axis consists of 25 bins. The vertical axis represents the number of base pairs normalized by the total number of genes. (B) *S. cerevisiae*. (C) *C. elegans*. (D) *D. melanogaster*. (E) *R. norvegicus*. (F) *M. musculus*. (G) *H. sapiens*.

symmetrical. We defined an A-singleton as a single nucleotide 'A' next to any other nucleotide base (C, G or T). A-singletons (non-repeats) were used as a control group, whereas A-repeats (length ≥ 2) served as the experimental group. Because the A-singletons are not repeats, differences in the occurrence of A-singletons and A-repeats should be attributable to the repetitive nature of the sequence. Figure 2A and B show a comparison between A-singletons

and A-repeats (length = 5–30) in humans, indicating that long sense A-repeats are enriched upstream of TSSs compared with downstream sequence, whereas long sense T-repeats are enriched in the opposite direction. In addition, the degree of asymmetry increases with repeat length (Supplementary Figures S7–S9). Figure 3 shows the result of an unpaired *t*-test between bins 1 to 10 and bins 16 to 25. It is clear that the numbers of A- and T-repeats in upstream and downstream repeats are not equal. A- and T-repeats yield *P*-values of 2.97E-15 and 6.44E-10, respectively. The conservation of the imbalance between the A- and T-repeat distribution upstream and downstream of the TSS across several mammalian species suggests that these mononucleotide repeats may have functional roles in mammalian genomes.

The enrichment of A-repeats correlates with gene functions

In yeasts, mononucleotide repeats are characteristic of certain gene families. Poly(dA:dT) tracts are enriched in the promoters of growth-related genes, whereas stress-related genes tend to contain TATA boxes (10,17). In the human genome, 575 housekeeping genes and 7261 tissue-specific genes were identified (30,31). The frequencies of A- and T-repeats in these two categories are dependent on repeat size. Short A- and T-repeats (2–9 bp) are more abundant in tissue-specific genes (Figure 4A and B, Supplementary Figures S10 and S11), whereas long A- and T-repeats (10–30 bp) are more abundant in housekeeping genes (Figure 4C and D, Supplementary Figures S10 and S11). Our findings suggest that non-random distributions of A- and T-repeats around the TSS correlate with gene function.

A-repeats are preferential targets of Ago binding

In humans, the Ago proteins form a subfamily of the Argonaute proteins (25,26). Ago is a ribonucleoprotein that is required by the RISC (19,20). Ago proteins have been shown to bind mononucleotide repeats (29). This complex contains a small RNA and requires Dicer protein for ribonucleoprotein assembly. The small RNA guides the Ago–Dicer complex to specific gene targets. In humans, Dicer is called DICER1, and the Ago subfamily consists of AGO1, AGO2, AGO3 and AGO4. Recent data from cross-linking and immunoprecipitation coupled with deep sequencing provided the locations of all Ago-binding sites across the whole genome of HEK-293 cells (29). We counted the number of repeats in all sequence reads from

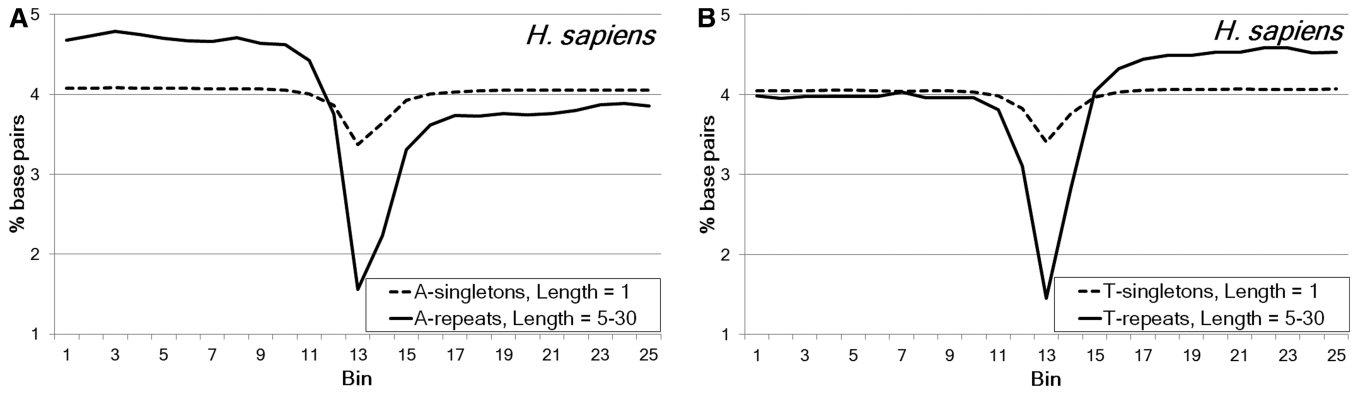

Figure 2. Comparisons between A/T-singletons (length = 1) and A/T-repeats (length = 5–30 bp) around TSS in the whole genome. The horizontal axis consists of 25 bins. Each bin covers 800 bp, with 10 000 bp upstream and 10 000 bp downstream of TSS in total. The TSS is centered in the 13th bin. The vertical axis represents the number of base pairs normalized to the percentage of all nucleotides. (**A**) A-repeats. (**B**) T-repeats.

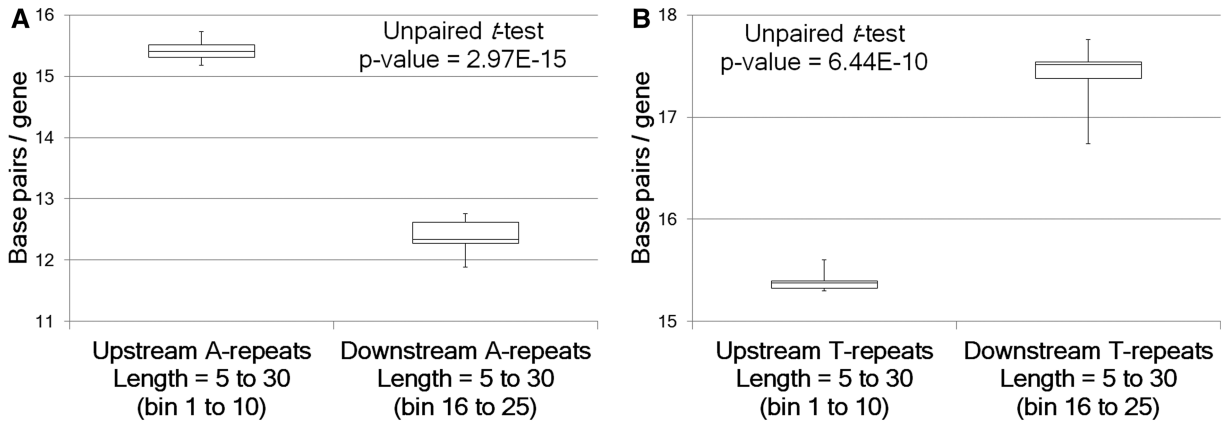

Figure 3. Box plots of numbers of A- and T-repeats in bins 1–10 and bins 16–25. The unpaired *t*-test was used to compare the mean difference between upstream repeats (bins 1–10) and downstream repeats (bins 16–25).

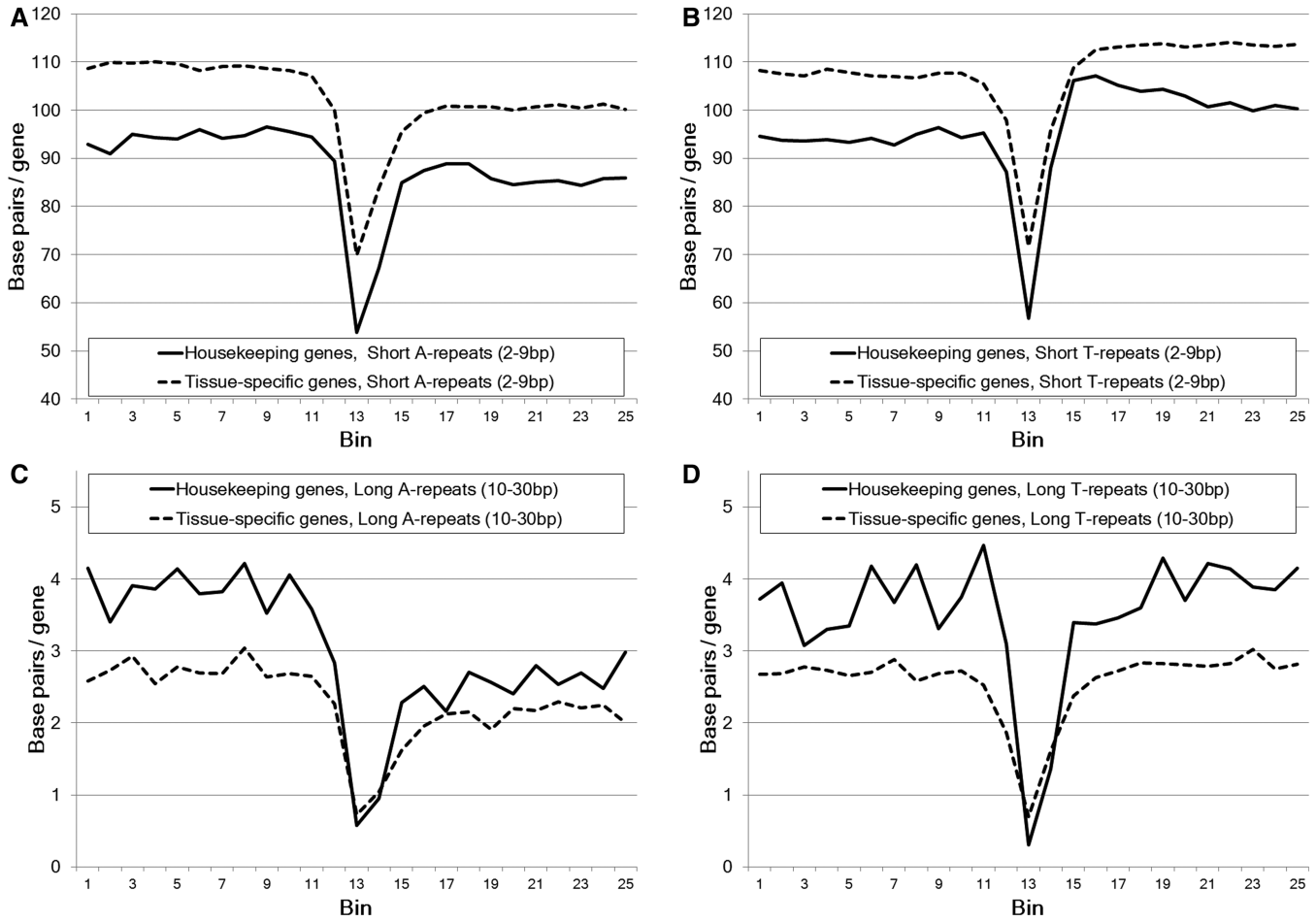

Figure 4. A comparison of human housekeeping and tissue-specific genes. (**A** and **B**) Short A- and T-repeats. (**C** and **D**) Long A- and T-repeats.

this data set. Sequence reads that are part of other reads were excluded. The exclusion method is illustrated in Supplementary Figure S2. We found that all members of the Ago protein family preferentially bind A-repeats. Moreover, Ago-binding ability increases with repeat length (Figure 5).

A-repeats are *cis*-regulatory elements

Ago proteins bind sequences around the TSS and control transcription in human cells (21–24). Therefore, Ago-bound repeats may serve as *cis*-regulatory elements in mammals. DICER1 is an essential protein in Ago complex assembly. DICER1 KD should inhibit all Ago complexes, independent of Ago member or binding site (Figure 6A and B). The genes in HEK-293 cells with DICER1 KD (Supplementary Table S3) were grouped into three categories: downregulated (Dn), upregulated (Up) and non-regulated (Nu - neutral). The methods used to calculate the fold change, *P*-values and *q*-values are described in the 'Materials and Methods' section.

First, we analyzed HEK-293 cells that had been subjected to DICER1 KD for 6 days (Figure 7A). At A-singleton, repeat length = 1, the fold change, i.e. the ratio between the number of A-singletons in two groups of genes, were almost constant at 1.0, indicating that single A-nucleotides do not correlate with gene expression. In contrast, A-repeats of 15–30 bp in length show distinct fold changes, and each bin shows a similar pattern of deviation. In DICER1 KD HEK293 cells, A-repeats tend to be enriched upstream of the TSS in downregulated genes (fold change > 1) and tend to be depleted upstream of the TSS in upregulated genes (fold change < 1). As shown in the leftmost column of Figure 7A, integrating bin 1–10 together yields highly significant $P < 0.001$ and $q = 6.67\text{E-}04$ (length = 15–30). Surprisingly, the first, third and forth bins, which are 6801–10 000 bp upstream far from TSS show striking fold changes. Two-day DICER1 KD experiments yielded results similar to those of the six-day DICER1 KD (Figure 7B). DICER1 KD was also explored in other cell lines in a similar manner. The results obtained from DICER1 KD in mouse embryo, mouse liver and HeLa cell lines confirmed the regulatory role of A-repeats. The presence of A-repeats upstream of the TSS suppresses gene expression in DICER1 KD (Figure 8A–C). However, the pattern of A-repeat distribution was not the same as that in the HEK-293 cell line. For example, in both mouse tissues, the seventh bin shows the largest fold change.

Next, HEK-293 cell lines subjected to AGO1 KD, AGO2 KD, AGO3 KD and AGO4 KD were analyzed

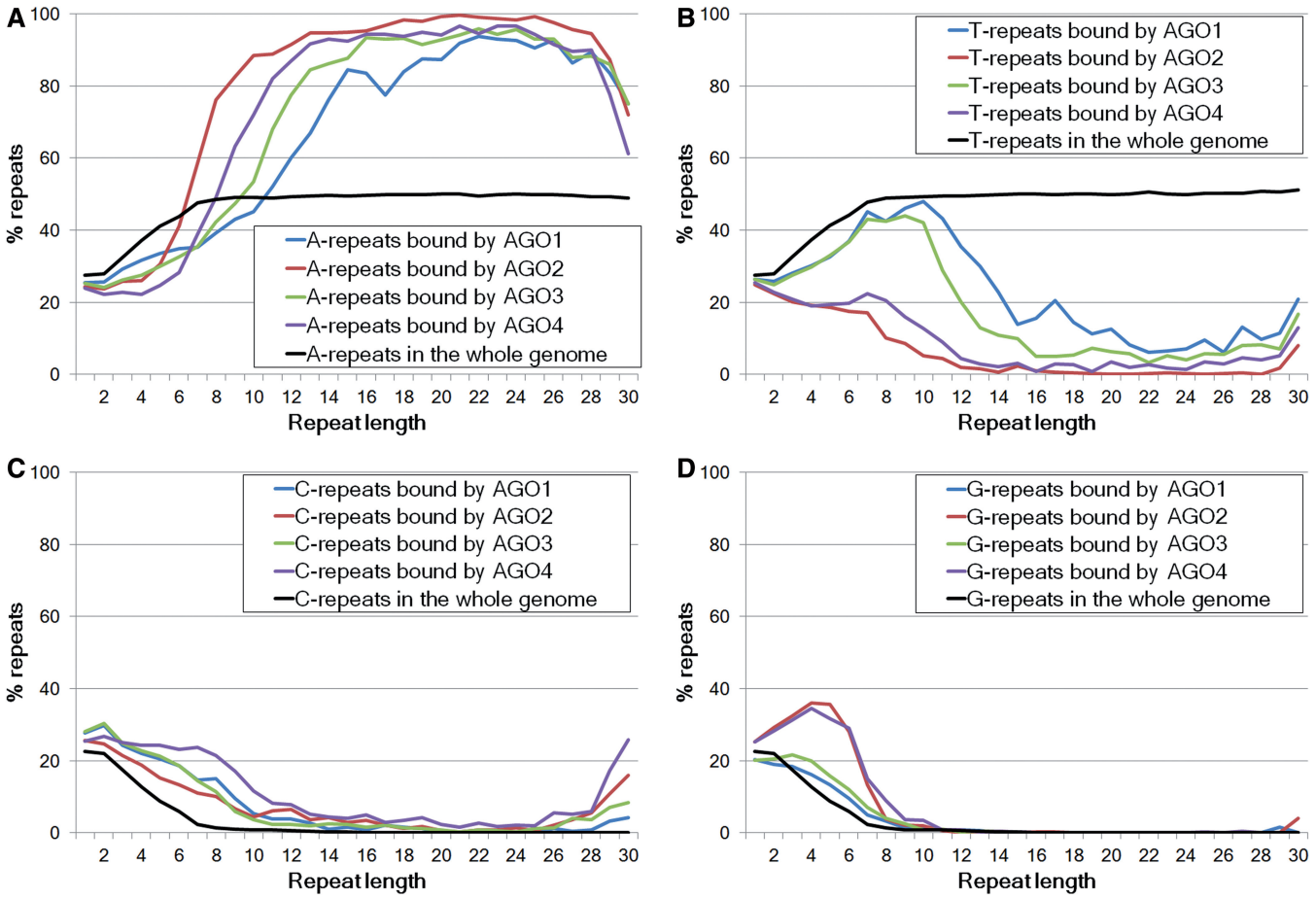

Figure 5. Ago protein affinity for A-, T-, C- and G-repeats. (**A–D**) A whole-genome comparison of mononucleotide repeats binding to Argonaute proteins in the HEK-293 cell line. The vertical axis represents numbers of repeats of the same length, normalized to the overall base composition (%A + %T + %C + %G = 100%). The horizontal axis is repeat length.

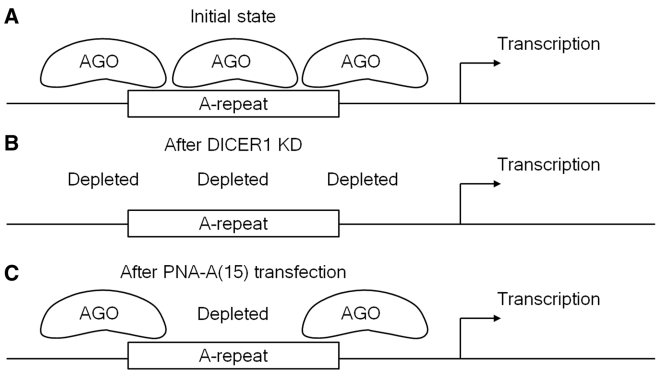

Figure 6. Expected results of DICER1 KD and PNA-A(15) transfection. (**A**) In the initial state, three Ago complexes bind upstream of the TSS. (**B**) DICER1 KD depletes all Ago binding, independent of target sequences. (**C**) PNA-A(15) transfection depletes only Ago complexes that bind to intact A(15).

(Figure 9A–D). The fold changes in the number of A-repeats (A-singleton, length = 1) remain constant at 1.0, suggesting no regulatory role for the A-singletons. Long A-repeats (length = 15–30) show fold changes of greater or less than 1. In AGO1 KD, the fold change pattern is consistent in each bin and is the opposite of the pattern found in DICER1 KD. The genes that are upregulated due to AGO1 KD are more enriched in A-repeats (fold change > 1), whereas A-repeats in the downregulated genes are more depleted (fold change < 1). Although the overall *P*-values do not reach the statistical significance, a striking enrichment of A-repeats appears in the eighth bin, 3601–4400 bp upstream of the TSS (Up versus Nu, $P < 0.001$, q = 1.00E-03). In the AGO2 KD and AGO3 KD experiments, the fold change pattern is not consistent and varies in each bin. In addition, the corresponding *P*-values do not reach robust statistical significance (all $q < 0.05$ but > 0.01). For the last Ago protein analyzed, AGO4, no significant change was observed in any inspected bin (all $q > 0.05$).

Ago proteins are *trans*-acting factors

To confirm the regulatory roles of Ago-bound A-repeat sequences, we transfected HEK-293 cells with a synthetic polymer mimicking the A-repeat, i.e. the PNA-AAAAAA AAAAAAAAA [PNA-A(15)] oligo (Supplementary Table S3). The injection of this polymer should inhibit protein binding to A-repeats (Figure 6A and C). Using ChIP, we showed that the PNA-A(15) interfered with

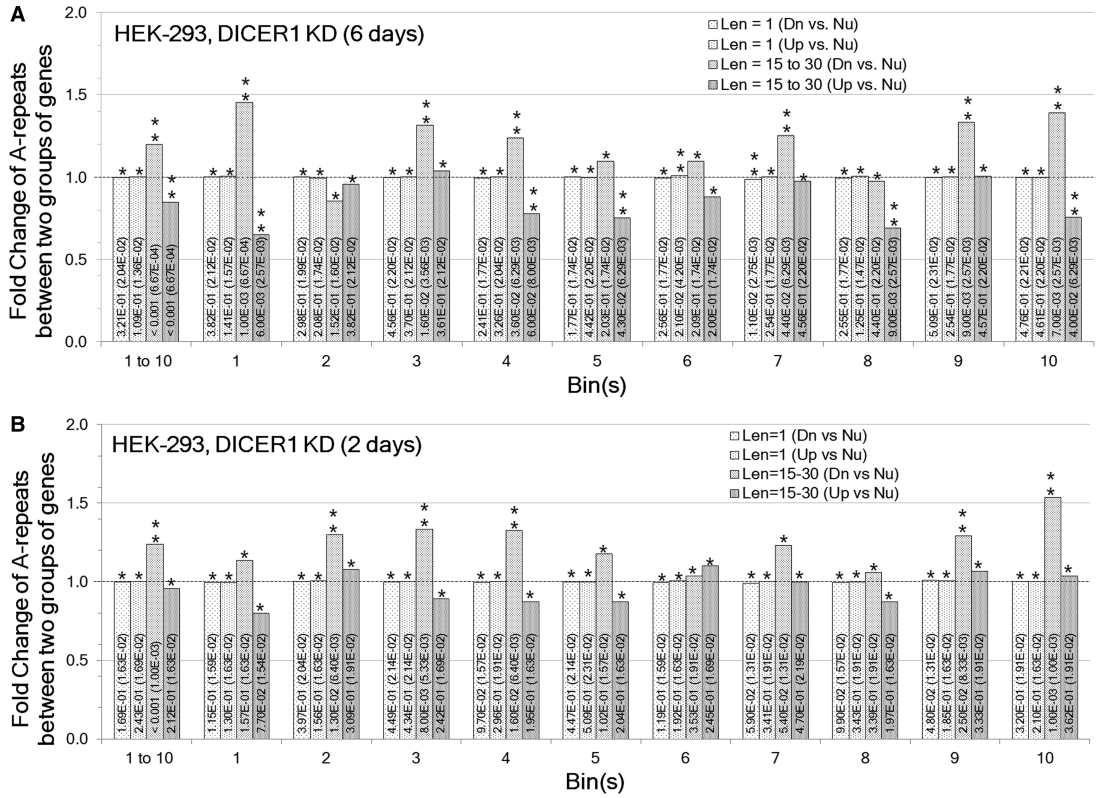

Figure 7. The fold changes in the number of A-repeats in Dn/Up and Nu, which denote downregulated, upregulated and non-regulated (neutral) genes, respectively. The horizontal axis represents the bin location. The first column shows the sums of bins 1–10. The two numbers '*p* (*q*)' in each bar are the *P*-value and *q*-value, respectively. Single stars indicate low confidence (FDR or $q \leq 0.05$), and double stars indicate high confidence (FDR or $q \leq 0.01$). (**A**) HEK-293 with DICER1 KD for 6 days. (**B**) HEK-293 with DICER1 KD for 2 days.

Ago binding to long A-repeat sequences (Figure 10A). We tested five distinct locations. The first two locations were AGO2-bound A(15) repeats (AGO2+, A-repeat+). The second two locations were AGO2-bound unique sequences (AGO2+, A-repeat-). Finally, the third location was a sequence that CLIPZ database listed as having no AGO2 binding. The binding of AGO2 to known AGO2-bound sequences (AGO2+) was confirmed. Moreover, PNA-A(15) transfection specifically reduced AGO2 binding to the two genomic locations containing AGO2-bound A(15) repeats (Figure 10A).

We also performed a microarray experiment to compare the PNA-A(15)-transfected group and the scrambled PNA-transfected control group. As shown in Figure 10B, we counted only the repeats bound by Ago proteins (AGO1-4). Ago proteins are thought to bind an A-repeat if the repeat overlaps with at least 1 bp of an Ago-bound sequence in the CLIPZ database (29). Both sense and antisense overlaps were permitted. The Ago-bound length indicates the length of the repeat that was actually bound by Ago proteins, not the whole repeat length. The difference in abundance of A-repeats among the regulated genes due to PNA-A(15) transfection is indicated by significant fold changes in several bins. The most striking change is a dramatic fold change (8.55) in the ninth bin, 2801–3600 bp upstream of the TSSs ($P = 1.40\text{E-}02$, $q = 1.28\text{E-}02$). However, this fold change does not imply that most upregulated genes contain AGO-bound A-repeats (length ≥ 15) in the ninth bin. Only 5 of 46 upregulated genes contain an A-repeat, but this ratio is ~9 times greater than that in the non-regulated genes (143 of 11 878) (odds ratio = 10.01, unadjusted Fisher's exact test *P*-value = 2.51E-04). The list of all genes with A-repeat sequences in the ninth bin and detailed calculations are shown in Supplementary Tables S5, S6 and S7.

The CLIPZ database provides information about Ago-binding sites in HEK-293 cells. However, the binding sites in the CLIPZ database may not be reliable because a read sequence could be mapped to multiple genomic locations. To find Ago-binding sites, each read sequence was aligned with the whole-human genome starting from chromosome 1. None of the read sequences could be uniquely aligned to a single binding site. The CLIPZ database displays multiple binding sites. However, the alignment was stopped if the number of binding sites exceeded a threshold of 30, which typically occurs for common sequences. Thus, the alignment halted immediately at chromosome 1 due to the detection of a number of binding sites, exceeding the threshold. To improve the accuracy of our search for binding site, we adjusted the threshold to <30 and recalculated the fold change in the ninth bin of PNA-A(15) transfection experiment (Figure 10C). We observed that the fold change increased with the use of a more stringent threshold. Using a threshold of 15, the fold change reaches almost 20. In other words, A-repeats were enriched in the upregulated genes 20 times. At thresholds

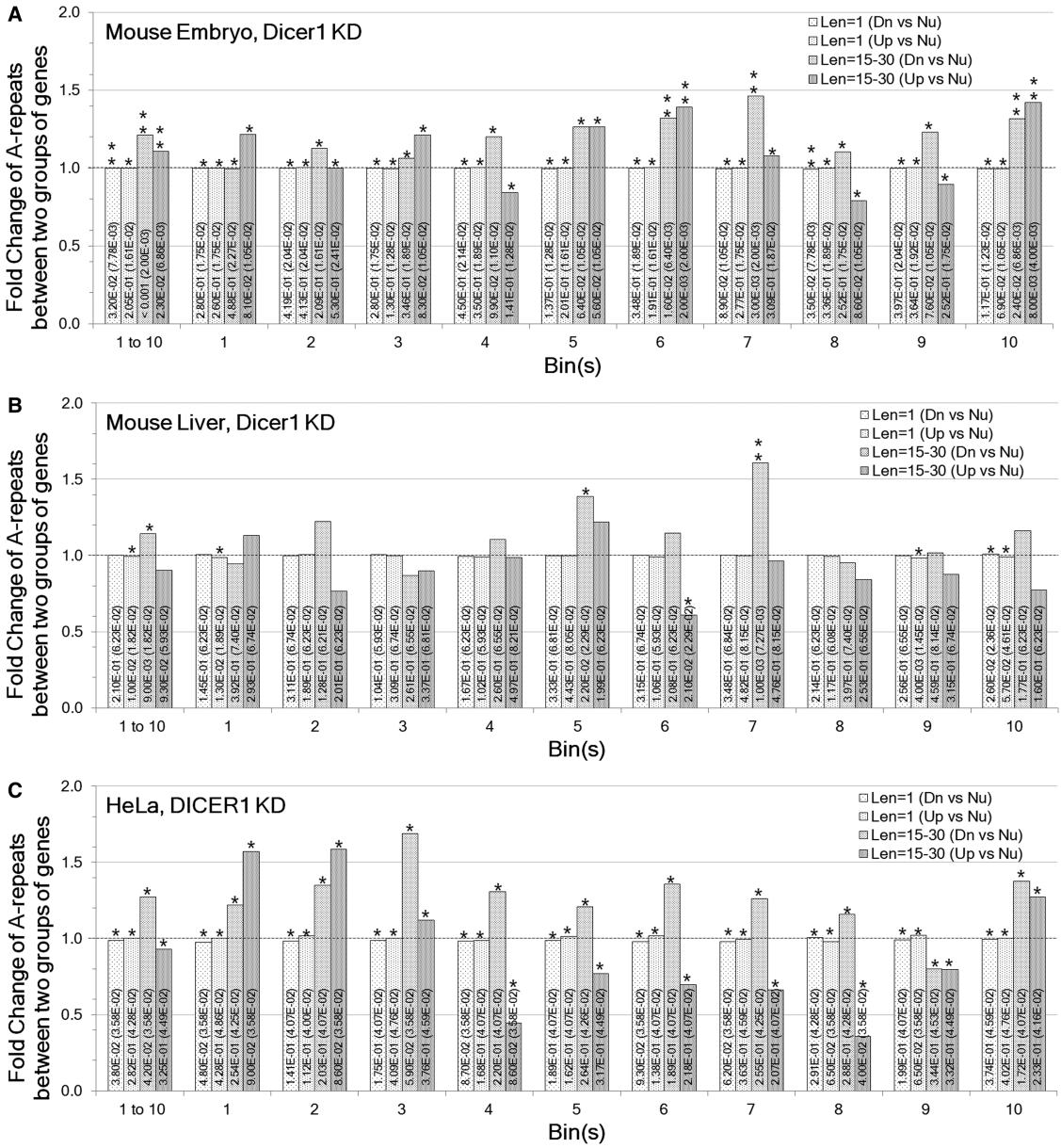

Figure 8. The fold changes in the number of A-repeats between Dn/Up and Nu, which denote downregulated, upregulated and non-regulated (neutral) genes, respectively. The horizontal axis represents the bin location. The first column shows the sums of bins 1–10. The two numbers '*p* (*q*)' in each bar are the *P*-value and *q*-value, respectively. Single stars indicate low confidence (FDR or $q \leq 0.05$), and double stars indicate high confidence (FDR or $q \leq 0.01$). (**A**) Mouse Embryo Dicer1 KD. (**B**) Mouse Liver Dicer1 KD. (**C**) HeLa DICER1 KD.

of 10, 5 and 1, no read sequence passed the threshold limit, and no fold change in A-repeat enrichment was observed.

DISCUSSION

Mononucleotide repeats are traditionally thought of as junk DNA that serves no function. However, our findings suggest a length-dependent *cis*-regulatory function of A-repeats, with Ago proteins as *trans*-acting factors. Nevertheless, other mechanisms, such as chromatin organization or physical property of repeat sequences, in addition to AGO-associated regulation may also direct A-repeats regulate transcription. Further evaluation into the precise role of repeats in mammalian promoter regions is desirable.

Several lines of evidence, including the findings of this study, suggest that sense A-repeats function as *cis*-regulatory elements and could play an important role in transcriptional regulation. First, the distribution of A-repeats within the genome is non-random. The enrichment of A-repeats upstream of TSSs correlates with the biological functions of the corresponding genes. An increase in the number of upstream sense A-repeats in several species, including rat, mouse and human, suggests that A-repeats are evolutionarily conserved and may perform essential functions in mammals. A sharp drop in the numbers of

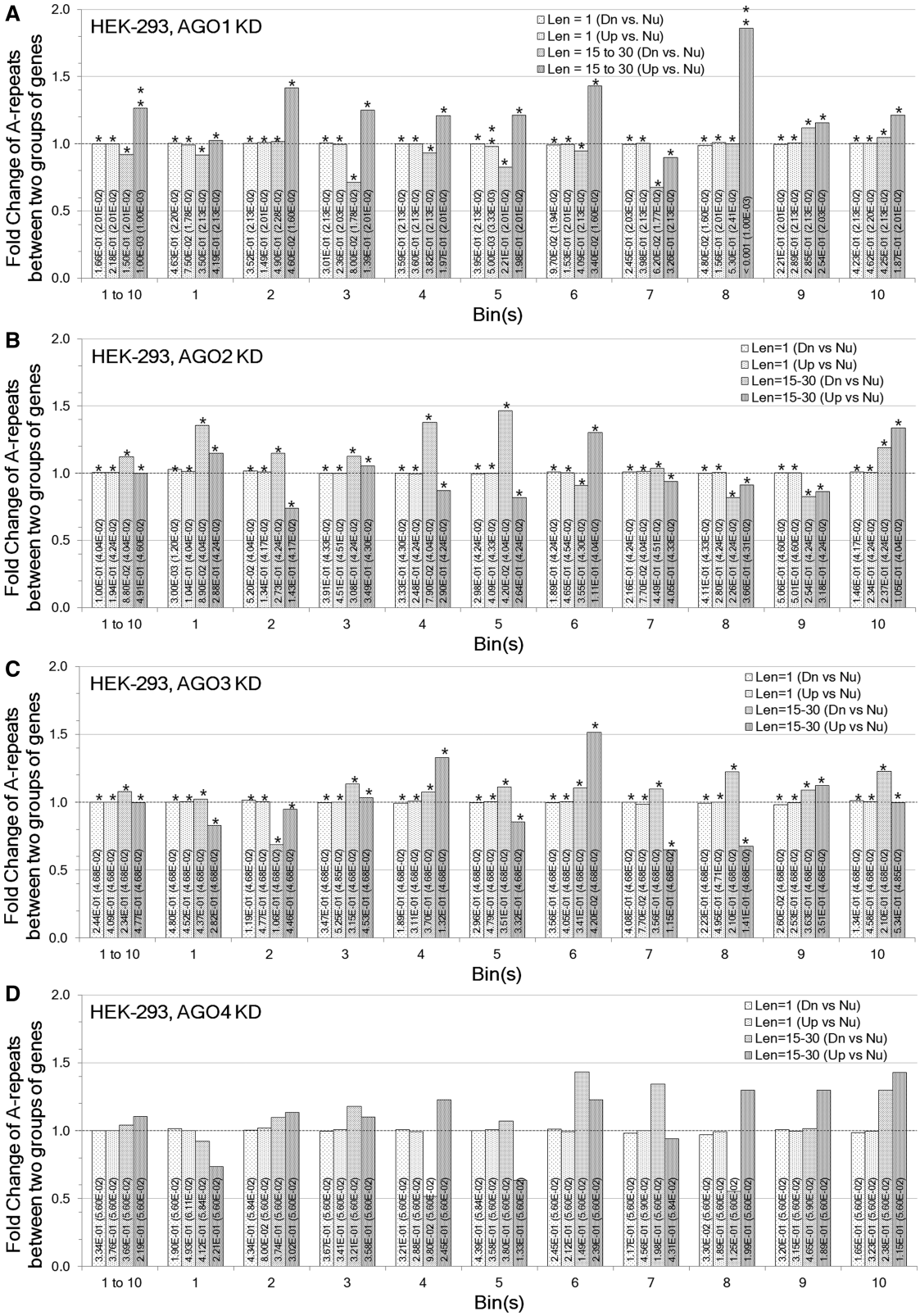

Figure 9. The fold changes in the number of A-repeats between Dn/Up and Nu, which denote downregulated, upregulated and non-regulated (neutral) genes, respectively. The horizontal axis represents the bin location. The first column shows the sums of bins 1–10. The two numbers 'p (q)' in each bar are the P-value and q-value, respectively. Single stars indicate low confidence (FDR or $q \leq 0.05$), and double stars indicate high confidence (FDR or $q \leq 0.01$). (**A**) HEK-293 AGO1 KD. (**B**) HEK-293 AGO2 KD. (**C**) HEK-293 AGO3 KD. (**D**) HEK-293 AGO4 KD.

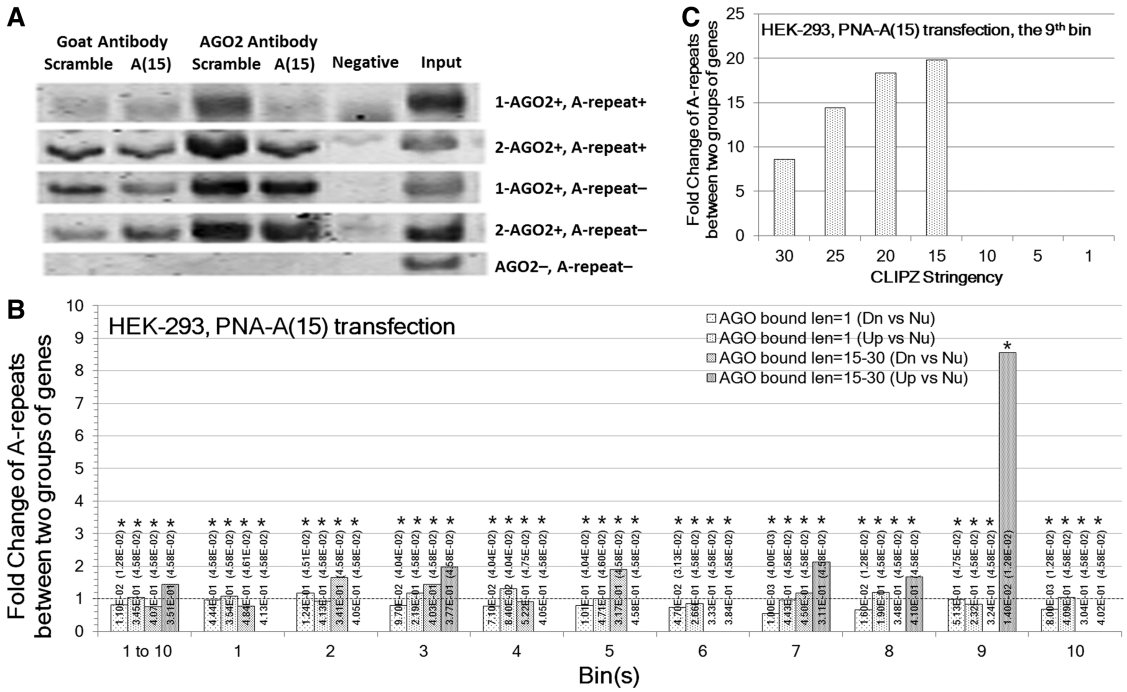

Figure 10. Inhibition of AGO binding to A-repeats. (**A**) PNA-A(15) reduced AGO2 binding to long A repeats. ChIP assays for AGO2-bound sequences were performed in HEK-293 cells transfected with control PNA (scramble) or test PNA-A(15) [A(15)] oligos. A control goat antibody was used as the negative control for the AGO2-bound sequences. Distilled water and sonicated DNA input were used as negative and positive controls for PCR, respectively. Reductions of AGO2 binding to A repeats in the presence of PNA-A(15) were demonstrated at AGO2 bound A-repeats using two AGO2+ and A-repeat+ PCR primer sets (AGO2+, A-repeat+). No reduction of AGO2 binding to unique DNA sequences in the presence of PNA-A(15) was observed using two AGO2+ and A-repeat-PCR primer sets (AGO2+, A-repeat–). No AGO2 binding to AGO2-negative locations was detected (AGO2–, A-repeat–). (**B**) HEK-293 PNA-A(15) transfection. The fold changes in the number of A-repeats between Dn/Up and Nu, which denote down/upregulated and non-regulated (neutral) genes, respectively. The horizontal axis indicates the bin location. The Ago-bound length is the length of the repeat that is actually bound by Ago proteins, not the entire repeat length. The first column shows the sums of bins 1–10. The two numbers 'p (q)' in each bar are the P-value and q-value, respectively. Single stars indicate low confidence (FDR or $q \leq 0.05$), and double stars indicate high confidence (FDR or $q \leq 0.01$). (**C**) The fold changes in the number of A-repeats between upregulated (Up) and non-regulated (Nu) genes in the ninth bin of HEK-293 PNA-A(15) transfected cells. The horizontal axis is the stringency threshold; lower numbers correspond to greater stringency. AGO-bound sequences that can be aligned to a number of genomic loci greater than the stringency threshold were excluded from the fold-change calculation.

mononucleotide repeats at the TSS occurs due to the presence of CpG islands (39) around the TSSs of most mammalian genes (40). It is also possible that repeats are inherently incompatible with a defined TSS. Additionally, a drop in the number of long A-repeats can be observed downstream of the TSS. Because Ago proteins preferentially bind A-repeats, these A-repeats may function as targets recruiting the RNAi RISC complex to transcribed mRNAs (in addition to their functions as *cis*-regulatory elements); thus, the presence of A-repeats within genes may be disadvantageous.

Second, A-repeats regulate gene expression through DICER1 and AGO1-4 binding. DICER1 silencing produced a consistent pattern and significant fold change in almost every bin, whereas AGO1-4 KD silencing produced different patterns of results. It is possible that proteins within Ago complexes have both distinct and shared functions and that some AGO subfamily members may substitute for each other. For example, both AGO1 and AGO2 are required for mammalian transcriptional silencing (41). Moreover, Ago proteins may cooperate with other factors, such as tissue-specific factors, to control gene expression. We hypothesized that the silencing of a single Ago protein at a time might produce variable results. Here, we observed different results when AGO1-4 was silenced. The non-random distribution of A-repeats between regulated and neutral genes was more significant in AGO1 KD cells than in AGO2-4 KD cells. Although AGO4 binds to A-repeats, AGO4 KD failed to show any correlation with the non-random distribution of A-repeats. Our findings suggest that AGO1 may perform a non-redundant regulatory role related to A-repeats that cannot be compensated by any other member of the Ago subfamily. In contrast, AGO4 may have only a minor role related to A-repeats or may have a redundant function that can be performed by other Ago proteins.

Third, the transfection of PNA-A(15) into HEK-293 cells altered the expression of genes enriched with A-repeats. An increase in the expression of A-repeat-enriched genes implies that *trans*-acting factor binding to A-repeats normally inhibits gene transcription in HEK-293 cells. The transfected PNA-A(15) competes with genomic A-repeats for binding to *trans*-acting factors, resulting in lower levels of *trans*-acting factor binding to the genomic A-repeats. A ChIP assay was conducted to demonstrate

that AGO proteins bound A-repeats and that the presence of PNA-A(15) decreased AGO-binding activity. However, the effect of PNA-A(15) transfection is not identical to those of DICER1 KD or AGO1-4 KD. This discrepancy may be because PNA-A(15) cannot compete with AGOs under all conditions. In particular, PNA-A(15) prevents AGO binding to A-repeats for repeats ≥15 bp. PNA-A(15) may fail to compete with AGO if a target A-repeat is too short (<15 bp) and AGOs can partially bind to other flanking sequences (Figure 6A and C).

Although there have been few studies investigating this issue to date, we believed that the length variation of A-repeats at certain loci may determine disease susceptibility. The enrichment of upstream sense A-repeats increases with repeat size. This size dependence may provide a selective advantage for long repeats compared with short repeats to support regulatory functions. A-repeats and AGOs may be under *cis-trans* co-evolution (42,43). Repeat length is a key factor for evolutionary advantage. We found that AGO1-4 prefers to bind A-repeats, and the binding preference increases with repeat size (Figure 5A). A loss of the essential regulatory functions of A-repeats may be disadvantageous. Therefore, A-repeat mutations that disrupt these repeats may be negatively selected. Moreover, genes with different functions may contain repeats of different sizes and locations. Long A-repeats are often found in constitutively expressed housekeeping genes. Therefore, in humans, housekeeping genes may exploit similar nucleotide repeat patterns to allow simultaneous gene expression. From an evolutionary perspective, poly(dA:dT) tracts and Argonaute proteins are found mostly in eukaryotes. Therefore, it would be interesting to identify the point in time during evolution when these *cis-trans* elements emerged and acquired a function in transcription regulation.

In conclusion, we report that the distribution of sense A- and T-repeats around the TSS is non-random. The distribution patterns in mice, rats and humans are similar and are distinct from those of invertebrates and yeast. In humans, different distributions of A- and T-repeats are observed for housekeeping and tissue-specific genes. Argonaute proteins bind to A-repeats and regulate gene expression. Nevertheless, further research is required to directly demonstrate and further elucidate the role of poly(A) repeats in (mammalian) promoter sequences.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors thank Dr Siwanon Jirawatnotai for critical review of this manuscript.

FUNDING

Research Chair Grant from the National Science and Technology Development Agency (NSTDA) of Thailand [R13/2554]; a Chulalongkorn University Dusadeepipat scholarship (to P.P.); 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.

Conflict of interest statement. None declared.

REFERENCES

1. Ellegren,H. (2004) Microsatellites: simple sequences with complex evolution. *Nat. Rev. Genet.*, **5**, 435–445.
2. Gemayel,R., Vinces,M.D., Legendre,M. and Verstrepen,K.J. (2010) Variable tandem repeats accelerate evolution of coding and regulatory sequences. *Annu. Rev. Genet.*, **44**, 445–477.
3. Birney,E. (2012) Journey to the genetic interior. Interview by Stephen S. Hall. *Sci. Am.*, **307**, 80–84.
4. Vinces,M.D., Legendre,M., Caldara,M., Hagihara,M. and Verstrepen,K.J. (2009) Unstable tandem repeats in promoters confer transcriptional evolvability. *Science*, **324**, 1213–1216.
5. Legendre,M., Pochet,N., Pak,T. and Verstrepen,K.J. (2007) Sequence-based estimation of minisatellite and microsatellite repeat variability. *Genome Res.*, **17**, 1787–1796.
6. Tirosh,I., Barkai,N. and Verstrepen,K.J. (2009) Promoter architecture and the evolvability of gene expression. *J. Biol.*, **8**, 95.
7. Jansen,A., van der Zande,E., Meert,W., Fink,G.R. and Verstrepen,K.J. (2012) Distal chromatin structure influences local nucleosome positions and gene expression. *Nucleic Acids Res.*, **40**, 3870–3885.
8. Suraweera,N., Iacopetta,B., Duval,A., Compoint,A., Tubacher,E. and Hamelin,R. (2001) Conservation of mononucleotide repeats within 3′ and 5′ untranslated regions and their instability in MSI-H colorectal cancer. *Oncogene*, **20**, 7472–7477.
9. Tóth,G., Gáspári,Z. and Jurka,J. (2000) Microsatellites in different eukaryotic genomes: survey and analysis. *Genome Res.*, **10**, 967–981.
10. Rando,O.J. and Winston,F. (2012) Chromatin and transcription in yeast. *Genetics*, **190**, 351–387.
11. Segal,E. and Widom,J. (2009) Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.*, **19**, 65–71.
12. Tillo,D., Kaplan,N., Moore,I.K., Fondufe-Mittendorf,Y., Gossett,A.J., Field,Y., Lieb,J.D., Widom,J., Segal,E. and Hughes,T.R. (2010) High nucleosome occupancy is encoded at human regulatory sequences. *PLoS One*, **5**, e9129.
13. Valouev,A., Johnson,S.M., Boyd,S.D., Smith,C.L., Fire,A.Z. and Sidow,A. (2011) Determinants of nucleosome organization in primary human cells. *Nature*, **474**, 516–520.
14. Yuan,G., Liu,Y., Dion,M.F., Slack,M.D., Wu,L.F., Altschuler,S.J. and Rando,O.J. (2005) Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science*, **309**, 626–630.
15. Nelson,H.C.M., Finch,J.T., Luisi,B.F. and Klug,A. (1987) The structure of an oligo(dA)•oligo(dT) tract and its biological implications. *Nature*, **330**, 221–226.
16. Raveh-Sadka,T., Levo,M., Shabi,U., Shany,B., Keren,L., Lotan-Pompan,M., Zeevi,D., Sharon,E., Weinberger,A. and Segal,E. (2012) Manipulating nucleosome disfavoring sequences allows fine-tune regulation of gene expression in yeast. *Nat. Genet.*, **44**, 743–750.
17. Cairns,B.R. (2009) The logic of chromatin architecture and remodelling at promoters. *Nature*, **461**, 193–198.
18. He,L. and Hannon,G.J. (2004) MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.*, **5**, 522–531.
19. Daneholt,B. (2006), The 2006 Nobel Prize in Physiology or Medicine - Advanced Information. *RNA interference*. Nobel Media AB 2013. http://www.nobelprize.org (23 July 2013, date last accessed).
20. Fire,A.Z., Xu,S., Montgomery,M.K., Kostas,S.A., Driver,S.E. and Mello,C.C. (1998) Potent and specific genetic interference by

double-stranded RNA in *Caenorhabditis elegans*. *Nature*, **391**, 806–811.
21. Verdel,A., Jia,S., Gerber,S., Sugiyama,T., Gygi,S., Grewal,S.I.S. and Moazed,D. (2004) RNAi-mediated targeting of heterochromatin by the RITS Complex. *Science*, **303**, 672–676.
22. Janowski,B.A., Younger,S.T., Hardy,D.B., Ram,R., Huffman,K.E. and Corey,D.R. (2007) Activating gene expression in mammalian cells with promoter-targeted duplex RNAs. *Nat. Chem. Biol.*, **3**, 166–173.
23. Schwartz,J.C., Younger,S.T., Nguyen,N., Hardy,D.B., Monia,B.P., Corey,D.R. and Janowski,B.A. (2008) Antisense transcripts are targets for activating small RNAs. *Nat. Struct. Mol. Biol.*, **15**, 842–848.
24. Morris,K.V., Santoso,S., Turner,A.M., Pastori,C. and Hawkins,P.G. (2008) Bidirectional transcription directs both transcriptional gene activation and suppression in human cells. *PLoS Genet.*, **4**, e1000258.
25. Höck,J. and Meister,G. (2008) The Argonaute protein family. *Genome Biol.*, **9**, 210.
26. Ender,C. and Meister,G. (2010) Argonaute proteins at a glance. *J. Cell. Sci.*, **123**, 1819–1823.
27. Dreszer,T.R., Karolchik,D., Zweig,A.S., Hinrichs,A.S., Raney,B.J., Kuhn,R.M., Meyer,L.R., Wong,M., Sloan,C.A., Rosenbloom,K.R. *et al.* (2012) The UCSC Genome Browser database: extensions and updates 2011. *Nucleic Acids Res.*, **40**, D918–D923.
28. Barrett,T., Troup,D.B., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M. *et al.* (2011) NCBI GEO: archive for functional genomics data sets - 10 years on. *Nucleic Acids Res.*, **39**, D1005–D1010.
29. Khorshid,M., Rodak,C. and Zavolan,M. (2011) CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic Acids Res.*, **39**, D245–D252.
30. Eisenberg,E. and Levanon,E.Y. (2003) Human housekeeping genes are compact. *Trends Genet.*, **19**, 362–365.
31. Liu,X., Yu,X., Zack,D., Zhu,H. and Qian,J. (2008) TiGER: A database for tissue-specific gene expression and regulation. *BMC Bioinformatics*, **9**, 271.
32. Schmitter,D., Filkowski,J., Sewer,A., Pillai,R.S., Oakeley,E.J., Zavolan,M., Svoboda,P. and Filipowicz,W. (2006) Effects of Dicer and Argonaute down-regulation on mRNA levels in human HEK293 cells. *Nucleic Acids Res.*, **34**, 4801–4815.
33. 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.
34. Westfall,P.H. and Young,S.S. (1993) *Resampling-Based Multiple Testing: Examples and Methods for p-value Adjustment*. John Wiley and Sons, NY.
35. Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genome–wide studies. *PNAS*, **100**, 9440–9445.
36. Storey,J.D., Taylor,J.E. and and Siegmund,D. (2004) Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *J. R. Stat. Soc. Ser. B*, **66**, 187–205.
37. Oh,S., Ju,Y. and Park,H. (2009) A highly effective and long-lasting inhibition of miRNAs with PNA-based antisense oligonucleotides. *Mol. Cells*, **28**, 341–345.
38. Kongruttanachok,N., Phuangphairoj,C., Thongnak,A., Ponyeam,W., Rattanatanyong,P., Pornthanakasem,W. and Mutirangura,A. (2010) Replication independent DNA double-strand break retention may prevent genomic instability. *Mol. Cancer*, **9**, 70.
39. Gardiner-Garden,M. and Frommer,M. (1987) CpG islands in vertebrate genomes. *J. Mol. Biol.*, **196**, 261–282.
40. Akan,P. and Deloukas,P. (2008) DNA sequence and structural properties as predictors of human and mouse promoters. *Gene*, **410**, 165–176.
41. Janowski,B.A., Huffman,K.E., Schwartz,J.C., Ram,R., Nordsell,R., Shames,D.S., Minna,J.D. and Corey,D.R. (2006) Involvement of AGO1 and AGO2 in mammalian transcriptional silencing. *Struct. Mol. Biol.*, **13**, 787–792.
42. Landry,C.R., Wittkopp,P.J., Taubes,C.H., Ranz,J.M., Clark,A.G. and Hartl,D.L. (2005) Compensatory *cis-trans* evolution and the dysregulation of gene expression in interspecific hybrids of drosophila. *Genetics*, **171**, 1813–1822.
43. Kuo,D., Licon,K., Bandyopadhyay,S., Chuang,R., Luo,C., Catalana,J., Ravasi,T., Tan,K. and Ideker,T. (2010) Coevolution within a transcriptional network by compensatory *trans* and *cis* mutations. *Genome Res.*, **20**, 1672–1678.