Genome Analysis

Human cis natural antisense transcripts initiated by transposable elements

Andrew B. Conley[1], Wolfgang J. Miller[2] and I. King Jordan[1]

[1] School of Biology, Georgia Institute of Technology, Atlanta, GA 30306, USA
[2] Laboratories of Genome Dynamics, Center of Anatomy and Cell Biology, Medical University of Vienna, Vienna, Austria

The capacity of human transposable elements (TEs) to promote cis natural antisense transcripts (cis-NATs) is revealed by the discovery of 48 718 human gene antisense transcriptional start sites (TSSs) within TE sequences. TSSs that yield cis-NATs are overrepresented among TE sequences, and TE-initiated cis-NATs are more abundant close to the 3′ ends of genes. The TE sequences that promote antisense transcription within human genes are relatively ancient, suggesting that selection has acted to conserve their function.

Cis natural antisense transcripts and transposable elements

Cis natural antisense transcripts (cis-NATs) are RNAs that are transcribed from the antisense strand of a gene locus, which are thus complementary to the RNA transcribed from the sense strand. It is becoming increasingly apparent that cis-NATs are used to regulate the expression of human genes [1–3]. Cis-NATs may regulate expression at the transcriptional level through the avoidance of transcriptional collisions [4] or post-transcriptionally through any one of the number of double-stranded RNA (dsRNA)-induced regulatory pathways collectively known as RNA interference [5].

Transposable elements (TEs) have been shown to contribute a variety of noncoding RNAs that act as dsRNA regulators of gene expression across diverse eukaryotic species, including humans. Indeed, TEs encode several distinct classes of regulatory RNAs including short interfering RNAs [6–8], microRNAs [9–11], repeat-associated small interfering RNAs [12] and piwi-interacting RNAs [13]. It seems that multiple distinct RNA interference mechanisms have evolved independently as genome defense mechanisms against TEs, only to be later coopted to regulate host genes [10]. There are three reasons to believe that TEs may represent a potentially rich source of cis-NATs that can regulate human gene expression: (i) the abundance of TEs in the human genome [14], (ii) the ability of TE sequences to promote transcription [15] and (iii) the relationship between TEs and RNA interference. To explore this possibility, we conducted a genome-scale survey of the ability of TE sequences to contribute cis-NATs to human genes.

Genome-scale identification of cis-NATs

To evaluate the capacity of TEs to contribute cis-NATs to the human genome, we took advantage of a relatively new technology – cap analysis of gene expression (CAGE) [16] – to define the location of transcriptional start sites (TSSs) in the human genome. CAGE relies on the isolation of full-length cDNAs using biotinylated mRNA caps. Linkers are ligated to the 5′ ends of the full-length cDNAs, and the first 20 bp of the cDNAs is cleaved with restriction enzymes. The resulting fragments are amplified, concatamerized and sequenced, allowing for the high-throughput characterization of the 5′ ends of mRNAs. Mapping of the 5′ mRNA end CAGE sequence tags to the genome unambiguously identifies TSSs. The location and orientation of human TSSs can be compared with gene and TE annotations to assess the relationship between cis antisense transcription and TEs.

A library of $>1.5 \times 10^6$ human CAGE sequence tags are available for download from the Japanese National Institute of Genetics website. We mapped these CAGE sequence tags to transcriptional units (TUs) in the human genome and compared their locations to those of TEs to assess whether TEs provide cis-NATs. A TU is defined here as a single protein-coding locus, with a characteristic strand orientation, bounded by the most 5′ and 3′ transcription start and termination sites, respectively. The UCSC Genome Browser database [17] KnownGenes annotations were used to locate TUs on the human genome sequence. KnownGenes annotations were chosen because they are supported by multiple sources of information including SWISS-PROT, TREMBL and GenBank mRNAs. A single TU may cover alternative transcript variants in the same orientation, and more than one TU may overlap at a single genomic locus.

A total of 869 085 CAGE sequence tags were mapped to 39 288 human genome TUs. The majority (639 490 or ~74%) of human TU-TSSs defined in this way correspond to sense transcripts, i.e. in the same orientation of the protein-coding mRNA. On average, each human TU has 16.3 sense TSSs. The prevalence of sense-oriented TSSs in the human genome is consistent with previous results and reflects the initiation of mRNA transcripts from multiple alternative promoters [18,19]. The relative excess of sense-oriented TSSs is also thought to be caused by selection against initiation of antisense transcription based on avoidance of collisions between the RNA transcription machinery tracking along the DNA [4]. Human TUs also have numerous antisense oriented TSSs (229 595 or ~26%) that correspond to cis-NATs. The abundance of human cis-NATs is underscored by the fact that the average human TU has 5.8 antisense TSSs.

Corresponding author: Jordan, I.K. (king.jordan@biology.gatech.edu).

TEs initiate antisense transcription

The locations of TEs in human TUs were taken from the RepeatMasker [20] annotation of the genome, and these data were used to discover transcripts that are initiated within TEs inserted into human TUs. 176 578 (~20%) of human TU TSSs were found to be initiated from within TE sequences. These data underscore the substantial capacity of TEs to promote transcription in human gene regions. TSSs that are initiated from within TEs are more likely to be found in the antisense orientation than non-TE TSSs; 48 718 of TE TSSs are found in the antisense orientation, yielding a TE antisense/sense ratio of 0.38 compared with 0.35 for non-TE TSSs. This difference, although moderate, is highly significant using a χ^2 analysis of a 2 × 2 contingency table (48 718/127.860 non-TE TSS antisense/sense = 180 877/511 630; $\chi^2 = 156.6$; $P = 6e^{-36}$). In other words, the TSSs that originate from within TE sequences are significantly enriched for cis-NATs, and this observation cannot be explained by random sampling alone. The vast majority of TEs that encode TSSs are found in introns (98.2%), which is consistent with the transcriptional collision model [4] for their mechanism of regulatory action, because these cis-NATs may not necessarily form dsRNA with mature sense transcripts.

The enrichment of cis-NATs initiated from within TE sequences is even more marked when the distribution of TSSs across TUs, from the 5′ to the 3′ ends, is observed. Human TUs were divided into 20 equal-sized bins, and antisense/sense TSS ratios were calculated for each bin. When bin-specific averages across all human TUs are plotted, the ratio of antisense/sense TE TSSs increases progressively from the 5′ to the 3′ ends of human TUs (Figure 1). The slope of this trend is positive, and the correlation is statistically significant. This enrichment suggests the possibility that antisense TE TSSs near the 3′ ends of genes are more efficacious regulators and thus favored by selection. Under the transcriptional collision model [4], the preponderance of cis-NATs initiated near the 3′ ends of genes would provide for more opportunities for collisions between RNA polymerase complexes tracking along opposite strands of the DNA and less chance for sense transcription complexes to get through to the ends of the genes. The opposite 5′ to 3′ trend in antisense/sense TSS ratios is seen for non-TE TSSs. There is a slight decrease in the antisense/sense ratio along human TUs, although this trend is far less pronounced and not statistically significant (Figure 1).

The relative excess of antisense transcripts initiated from TEs and their enrichment closer to the 3′ ends of TUs suggests the possibility that they may yield cis-NATs with biologically significant regulatory activities. If this is indeed the case, one may expect natural selection to preserve these functionally active TE-derived cis-NATs. Accordingly, TEs that initiate cis-NATs are predicted to be older than those that do not initiate transcription because of the fact that they have been preserved in the genome by selection. The age distribution of the TEs that donate cis-NATs was analyzed to evaluate this prediction. The observed proportions of elements from different TE classes/families that donate cis-NATs were compared with the expected proportions based on their relative frequencies in the genome. Consistent with the expectation, relatively ancient elements are significantly overrepresented (Figure 2). For instance, there are more antisense TU TSSs derived from the ancient L2 and MIR families than expected by their genome frequencies. Members of younger element families, such as L1 and Alu, initiate significantly fewer antisense TU TSSs than expected based on their genome frequencies.

The relative ages of TEs within families can be measured by taking the divergence of the TE sequences from their family consensus sequence, because older TEs have accumulated more substitutions, on average, than younger elements [14]. For the younger TE families, relative divergence values indicate that the TEs that donate antisense TSSs are older than TEs from the same family that do not donate TU-TSSs. For instance, Alu elements that initiate antisense transcription with human TUs have significantly greater sequence divergence from their consensus sequences than Alus that do not co-locate with TSSs (average ± SD Jukes-Cantor distance for Alu-antisense-TSS = 0.15 ± 0.05, Alu-non-TSS = 0.14 ± 0.05,

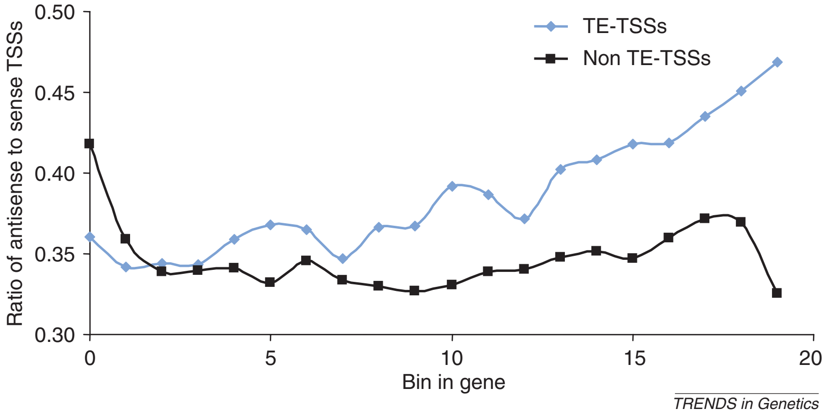

Figure 1. Ratio of antisense/sense transcriptional start sites (TSSs) along human genes. Human transcriptional units (TUs) were divided into 20 equal-sized bins, and ratios of the numbers of antisense/sense TSSs were calculated for each bin. Average bin-specific ratios are shown for TSSs initiated within transposable elements (TEs; TE TSSs in gray) and TSSs not initiated from TEs (non-TE TSSs in black). Linear regression was used to plot the slope of the antisense/sense ratio trend along bins from 5′ to 3′ gene ends, and the Spearman rank correlation coefficient (R) was used to evaluate the significance of the trends. For TE TSSs, $y = 10e^{-2}$, $R = 0.87$, $t = 7.62$, $P = 5e^{-7}$. For non-TE TSSs, $y = -3e^{-2}$, $R = -0.34$, $t = 1.54$, $P = 0.14$.

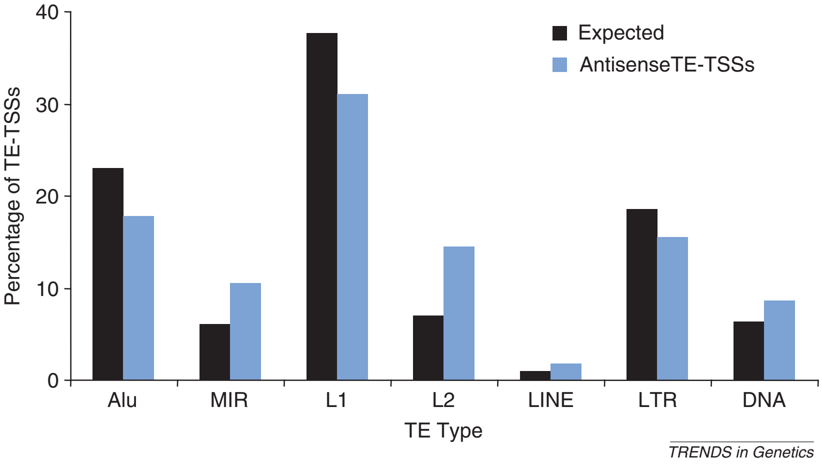

Figure 2. Relative proportions of transposable element (TE)-derived cis natural antisense transcripts (cis-NATs). Human genome TEs are broken down into seven classes/ families, and the relative percentages of TE-derived cis-NATs are shown in gray for each family. The expected percentages of TE-derived cis-NATs, based on the genome proportions of each class/family, are shown in black. A χ^2 test for goodness of fit was used to compare the observed versus expected proportions of TE-derived cis-NATs. The differences between the observed and expected distributions across classes are highly statistically significant ($\chi^2 = 7{,}671$; $P = 0$).

$t = 19.42$, $P = 6e^{-84}$). This suggests that many antisense TE-TSSs are in fact conserved by selection and also helps to resolve a standing question as to why older elements of some classes of TEs, such as Alus, are enriched in gene regions. Alus insert more frequently into AT-rich DNA but are preferentially retained in GC-rich gene regions; this has been taken to suggest that they are conserved in gene regions by virtue of some unknown functional role that they play for those genes [14]. Our data indicate that, in the case of some Alu sequences, their functional role is related to the initiation of regulatory cis-NATs.

Another way to evaluate the relative ages of TEs is to compare their evolutionary conservation based on presence/absence patterns of orthologous insertions between related species. Using this approach, we compared the human-to-mouse conservation of TEs that encode TSSs versus those that do not; 15.3% of TE TSSs are conserved between human and mouse versus 2.8% of TEs that do not encode TSSs. This difference is statistically significant ($\chi^2 = 4 \times 10^6$; $P = 0$). The greater between-species conservation of TE-TSSs is further evidence consistent with the action of purifying selection based on function.

Concluding remarks

The ability of transposable elements (TEs) to contribute regulatory sequences to eukaryotic genomes was discovered through several case-by-case studies on individual genes [21,22]. Later, genome-scale approaches began to uncover just how widespread this phenomenon is, particularly in mammalian genomes with high TE copy numbers [23–25]. Relying on a genome-scale approach for the identification of transcriptional start sites, we showed that TEs contribute tens of thousands of cis natural antisense transcripts to human genes. The potential regulatory effects of these TE-derived antisense transcripts are substantial.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tig.2007.11.008.

References

1 Chen, J. *et al.* (2004) Over 20% of human transcripts might form sense-antisense pairs. *Nucleic Acids Res.* 32, 4812–4820
2 Lehner, B. *et al.* (2002) Antisense transcripts in the human genome. *Trends Genet.* 18, 63–65
3 Yelin, R. *et al.* (2003) Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol.* 21, 379–386
4 Osato, N. *et al.* (2007) Transcriptional interferences *in cis* natural antisense transcripts of humans and mice. *Genetics* 176, 1299–1306
5 Fire, A. (1999) RNA-triggered gene silencing. *Trends Genet.* 15, 358–363
6 Matzke, M.A. *et al.* (2000) Transgene silencing by the host genome defense: implications for the evolution of epigenetic control mechanisms in plants and vertebrates. *Plant Mol. Biol.* 43, 401–415
7 Slotkin, R.K. *et al.* (2005) Heritable transposon silencing initiated by a naturally occurring transposon inverted duplication. *Nat. Genet.* 37, 641–644
8 Vastenhouw, N.L. and Plasterk, R.H. (2004) RNAi protects the *Caenorhabditis elegans* germline against transposition. *Trends Genet.* 20, 314–319
9 Piriyapongsa, J. and Jordan, I.K. (2007) A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS ONE* 2, e203
10 Piriyapongsa, J. *et al.* (2007) Origin and evolution of human microRNAs from transposable elements. *Genetics* 176, 1323–1337
11 Smalheiser, N.R. and Torvik, V.I. (2005) Mammalian microRNAs derived from genomic repeats. *Trends Genet.* 21, 322–326
12 Vagin, V.V. *et al.* (2006) A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* 313, 320–324
13 Brennecke, J. *et al.* (2007) Discrete small RNA-Generating loci as master regulators of transposon activity in Drosophila. *Cell* 128, 1089–1103
14 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
15 Thornburg, B.G. *et al.* (2006) Transposable elements as a significant source of transcription regulating signals. *Gene* 365, 104–110
16 Shiraki, T. *et al.* (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A.* 100, 15776–15781
17 Karolchik, D. *et al.* (2003) The UCSC Genome Browser Database. *Nucleic Acids Res.* 31, 51–54
18 Carninci, P. *et al.* (2006) Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.* 38, 626–635
19 Okazaki, Y. *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420, 563–573
20 Smit, A.F.A., *et al.* (1996–2004) RepeatMasker Open-3.0

21 Britten, R.J. (1996) DNA sequence insertion and evolutionary variation in gene regulation. *Proc. Natl. Acad. Sci. U. S. A.* 93, 9374–9377

22 Britten, R.J. (1997) Mobile elements inserted in the distant past have taken on important functions. *Gene* 205, 177–182

23 Jordan, I.K. *et al.* (2003) Origin of a substantial fraction of human regulatory sequences from transposable elements. *Trends Genet.* 19, 68–72

24 Mariño-Ramírez, L. and Jordan, I.K. (2006) Transposable element derived DNaseI-hypersensitive sites in the human genome. *Biol. Direct* 1, 20

25 van de Lagemaat, L.N. *et al.* (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. *Trends Genet.* 19, 530–536

0168-9525/$ – see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2007.11.008 Available online 14 January 2008

Genome Analysis

Evolution of complexity in miRNA-mediated gene regulation systems

Shohei Takuno[1,2] **and Hideki Innan**[1,3]

[1] Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX 77030, USA
[2] Laboratory of Plant Breeding and Genetics, Graduate School of Agricultural Science, Tohoku University, Aoba, Sendai, Miyagi 981-8555, Japan
[3] Graduate University for Advanced Studies, Hayama, Kanagawa 240-0193, Japan

Using *Arabidopsis* microRNA (miRNA)-mediated gene regulation system as a model, we investigated how complex systems evolve with special attention to selection to maintain the systems. We found that the copy number of miRNA genes within each system is a key factor to determine the complexity of the system, indicating a crucial role of gene duplication to increase the complexity. Furthermore, we show that the mode of selection to maintain the systems depend on their complexity levels.

Definitions of a microRNA system and its complexity

Darwin's theory of evolution predicts that complex biosystems have formed by the accumulation of numerous slight adaptive changes [1], but our understanding of the evolutionary mechanisms of complex systems is limited [2–5]. Gene duplication is considered one of the major genetic sources of complexity of biological systems [6–11]. To test Darwin's theory of the step-by-step evolution of complexity by gene duplications, we focus on the microRNA (miRNA) gene regulation systems in *Arabidopsis* [12–14] (see Supplementary Material online). We define a single regulation system such that it consists of miRNAs with almost identical sequences and their target genes (Box 1). Most systems have multiple miRNAs encoded by distinct genomic regions, miRNA genes (Table S1 in Supplementary Material online). Here, the number of miRNA genes in a system is denoted by k. We hypothesize that k is positively correlated with the level of complexity because (i) miRNAs from different miRNA genes can have different expression patterns [15,16] and (ii) sequence variations between miRNAs and target genes' binding sites might affect the pattern of regulation (Box 1). If so, a gene regulatory system with larger k could potentially design more-complicated expression patterns of the target genes; therefore, k could be a crucial factor to determine the complexity of a system (i.e. complexity of the gene regulation mechanism). Although the target genes are also important, when considering the complexity level, we propose miRNAs are more significant because the miRNAs control their target genes (i.e. interaction within the system is always one way).

Evolution of complexity

Table S1 in the Supplementary Material online summarizes the 33 miRNAs with at least one target gene. Evidence suggests that several of these miRNAs have significant regulatory roles (reviewed in [17]), and this idea could be extended to a genomic scale (Box 1). This is consistent with the ancient origins of many *Arabidopsis* miRNA systems (Table S1 in Supplementary Material online). More than one half of the *Arabidopsis* miRNA systems appeared before the monocot–dicot split (~125–150 MYA) [18], and some might predate the ancestor of all land plants (>400 MYA) [19–21]. Clearly, those miRNAs with crucial functions have been conserved and subject to strong purifying selection for a long time. During this time, miRNA genes have experienced multiple rounds of duplications and subsequent losses [15,16,22,23], thereby changing k. If our hypothesis is correct, a system with a large k will be more suitable to optimize a single family of genes with similar function (i.e. multigene family). By contrast, a system with a small k will be simple, so that many kinds of functional genes (or gene families) can potentially use it. To test this prediction, we focus on the number of gene families in a system, which is denoted by t. Our hypothesis predicts a negative correlation between k and t. We used a BLASTP-based approach to estimate the numbers of gene families in the 25 *Arabidopsis* miRNA systems with more than one target gene (see Table S1 and Supplementary Material online). A gene family is defined such that all pairwise BLASTP E values among the member genes are $< 10^{-10}$ (see Supplementary Material online). As shown in Figure 1a, we found only one

Corresponding author: Innan, H. (innan_hideki@soken.ac.jp).