RESEARCH ARTICLE

Retrotransposon-centered analysis of piRNA targeting shows a shift from active to passive retrotransposon transcription in developing mouse testes

Tobias Mourier

Abstract

Background: Piwi-associated RNAs (piRNAs) bind transcripts from retrotransposable elements (RTE) in mouse germline cells and seemingly act as guides for genomic methylation, thereby repressing the activity of RTEs. It is currently unknown if and how Piwi proteins distinguish RTE transcripts from other cellular RNAs. During germline development, the main target of piRNAs switch between different types of RTEs. Using the piRNA targeting of RTEs as an indicator of RTE activity, and considering the entire population of genomic RTE loci along with their age and location, this study aims at further elucidating the dynamics of RTE activity during mouse germline development.

Results: Due to the inherent sequence redundancy between RTE loci, assigning piRNA targeting to specific loci is problematic. This limits the analysis, although certain features of piRNA targeting of RTE loci are apparent. As expected, young RTEs display a much higher level of piRNA targeting than old RTEs. Further, irrespective of age, RTE loci near protein-coding coding genes are targeted to a greater extent than RTE loci far from genes. During development, a shift in piRNA targeting is observed, with a clear increase in the relative piRNA targeting of RTEs residing within boundaries of protein-coding gene transcripts.

Conclusions: Reanalyzing published piRNA sequences and taking into account the features of individual RTE loci provide novel insight into the activity of RTEs during development. The obtained results are consistent with some degree of proportionality between what transcripts become substrates for Piwi protein complexes and the level by which the transcripts are present in the cell. A transition from active transcription of RTEs to passive co-transcription of RTE sequences residing within protein-coding transcripts appears to take place in postnatal development. Hence, the previously reported increase in piRNA targeting of SINEs in postnatal testis development does not necessitate widespread active transcription of SINEs, but may simply be explained by the prevalence of SINEs residing in introns.

Background

Retrotransposable elements (RTE) constitute a significant proportion of mammalian genomes. The RTEs proliferate through an RNA stage that is subsequently reverse transcribed back to genomic DNA [1]. The high level of divergence in RTE insertions between closely related organisms [2-5] and the link between RTE insertions and diseases [6-8] witness the ongoing activity of

RTEs in mammalian genomes. Several genomic mechanisms are devised to minimize the proliferation of RTEs acting both at pre- and post-transcriptional levels [9-11].

Mouse retrotransposable elements

Around 40 percent of the mouse genome consists of RTE sequence, slightly lower than observed for the human genome, although this presumably is a result of the higher substitution rate in mouse, limiting the identification of old RTE sequence [12,13]. RTEs are divided according to the presence or absence of long terminal repeats (LTR). Mammalian LTR elements consist mainly

Correspondence: tmourier@snm.ku.dk
Centre for GeoGenetics, Natural History Museum, University of Copenhagen, Øster Voldgade 5-7, DK-1350 Copenhagen K, Denmark

of endogenous retroviruses (ERV) that at some point during evolution have been inserted in the germline and fixed. Although the amount of sequence occupied by LTR elements is comparable between human and mouse, the level of *de novo* mutations caused by LTR element activity is extensively higher in mouse than in human [8,14]. The most abundant ERV class in the mouse genome (~5.5%) is the Class III ERVs, which in the RepeatMasker [15] annotation - upon which this study is based - is broadly divided in two groups, the ERVL and MaLR elements. The latter is a non-autonomous transposon, meaning that the elements do not encode the enzymatic machinery required for its own transposition. The Class II ERVs (~4% of the genome), annotated as ERVK in RepeatMasker, is believed to be younger than Class III ERVs [16] and consists of a broad range of clades, including the IAP elements (Intracisternal A-type Particles). Class I ERVs (ERV1 in RepeatMasker) cover less than 1% of the mouse genome. Through ectopic recombination between the flanking LTR sequences, solitary LTR sequences may be formed. In RepeatMasker, terminal LTR sequences and the internal sequences (residing between the terminal LTRs in a complete LTR retrotransposon) are annotated independently. Although the terminal and internal sequences may in many cases be determined to form a single LTR retrotransposon, for simplicity, the two annotations (termed 'LTRter' and 'LTRint', respectively) are analysed independently in this study.

Non-LTR retrotransposons are divided into LINEs (Long INterspersed Elements) that are autonomous, and SINEs (Short INterspersed Elements) that are non-autonomous. LINEs occupy roughly 20% of the mouse genome. The majority of mouse LINE elements belong to the L1 superfamily, which contains sub-families that are still active [17-19]. Despite the comparable levels of genome occupied by LINE sequences in human and mouse [12,13], there are more than 15 times as many full-length L1 elements with intact open reading frames in the mouse genome [20]. Almost 1.5 million SINE elements are present in the mouse genome, making up approximately 8% of the total genome size. Unlike the human genome where a single SINE, the Alus, is dominating [21], the mouse genome harbours two successful superfamilies of SINEs, Alu and B2 that are present in equal numbers [12]. The evolutionary histories of the mouse SINEs are truly different; Alus are derived from a 7SL RNA, whereas B2s evolved from a tRNA sequence [21,22].

Piwi proteins and small RNAs

Piwi-associated RNAs (piRNAs) are small (24-30 nucleotides long) RNAs that bind Piwi proteins of the Argonaute family [23,24]. The mouse genome encodes 3 Piwi proteins, MILI, MIWI and MIWI2 that all binds piRNAs

in the male germline [25,26]. Initially, piRNAs from adult mouse testis were found to contain less RTE sequence than would be expected from the genomic content of RTEs, suggesting that piRNAs were not specifically targeting RTEs [27,28]. However, a later study on piRNAs from an earlier (pre-pachytene) stage showed a high content of RTE sequence in piRNAs [29]. Further evidence for the involvement of mouse piRNAs in controlling RTE activity came with the finding that knockout of Mili and Miwi2 resulted in reduced piRNA levels and increased RTE transcription [29,30]. Knockout mice further showed decreased DNA methylation levels at RTE loci [31,32]. As the temporal expression of Piwi proteins in developing mouse testis coincides with the resetting of genomic methylation [33], it is hypothesised that piRNAs act as guides for the methylation machinery [29,31,32].

By analysing the piRNAs bound to MIWI2 and MILI, Aravin and colleagues [31] suggested the following scenario: In prenatal development (16.5 days postcoitum, dpc), transcripts from full-length active RTEs are the main substrates for piRNAs that primarily associate with MILI (and to a lesser extent to MIWI2). Available transcripts containing antisense RTE sequence bind this complex and antisense RTE piRNAs are formed which in turn associate primarily with MIWI2 (and MILI, respectively). Both complexes may bind complementary RTE transcripts, entering the so-called ping-pong amplification cycle of piRNAs, in which Piwi-bound piRNAs pair with complementary transcripts that are subsequently cut into new piRNAs having a 10 nucleotide overlap with the template piRNAs [31,34]. In prenatal development, piRNAs are primarily targeting L1 and IAP RTEs, for which activity has been reported at this stage [35,36]. In postnatal development (10 days postpartum, dpp) MIWI2 is no longer detectable, whereas MILI is present throughout germline development [31,37,38]. The overall level of piRNA targeting of RTE sequences drops at 10 dpp, but interestingly, a relative increase in the piRNAs targeting B1 SINEs (members of the Alu superfamily) was observed [31].

This raises two fundamental questions. Firstly, do Piwi proteins discriminate between transcripts and how is RTE sequences then distinguished from other transcripts? The finding of piRNAs targeting supported a scenario with limited discrimination [31]. Secondly, what lies behind the apparent shift in RTEs being targeted by piRNAs during development in male mouse germline? By analysing to extent to which genomic RTE loci are targeted by piRNAs in developing mouse testes, the present study aims at assessing the transcriptional dynamics of RTE during development, and consider the relationship between RTE activity and piRNA generation further.

The data for such analysis should meet a range of criteria. Although numerous mouse RNA libraries are

available, only a small subset is derived from wild-type developing mouse testes [39]. Further, as the prevalent transcription of RTEs will results in a large population of fragmented transcripts of sizes similar to piRNAs, analysis should be restricted to libraries of RNAs associated with Piwi proteins. This limits the available data to libraries from the above-mentioned study by Aravin and colleagues [31].

Results and Discussion

Theoretical piRNA coverage of individual RTE loci

Three libraries of small RNAs bound to MIWI2 and MILI proteins in mouse male germline [31] were reanalysed; one library with MIWI2-bound RNAs from 16.5 dpc (henceforth referred to as 'MIWI2 early') and two libraries with MILI-bound RNAs from 16.5 dpc and 10 dpp ('MILI early' and 'MILI late', respectively) (Table 1).

To analyse in detail the theoretical piRNA coverage of each individual RTE loci, reads from each RNA library were mapped onto the mouse genome. Only perfectly mapping reads were considered. Due to the inherent redundancy of RTE sequences, reads mapping uniquely to RTE loci are scarce and therefore all mapping reads were considered. The number of reads mapping to each RTE loci was weighted by the uniqueness of the reads, so that the count from each read was simply divided by the number of times that particular read mapped to the entire genome. Further, read coverage was weighted by library size and length of the locus, in an attempt to allow direct comparisons between libraries and RTE types. As seen from Figure 1, differences in this theoretical piRNA coverage differ markedly between RTE families and between libraries. When only considering reads that are exclusively mapping within a particular superfamily of RTEs (*e.g.* L1 LINEs) the overall patterns of coverage still remain (middle column in Figure 1). For LTR elements (both internal sequence and terminal repeats), most major trends are still observable when only considering reads mapping entirely within a single family of RTE (*e.g.* internal sequence of IAP-d elements) (right column in Figure 1). Unless otherwise noted, in the following all reads are used for analysis, and only RTE families with a least a thousand genomic

members are considered (see Additional File 1, Table S1 for a list of these 318 RTE families). The previously reported decrease in piRNA coverage of IAP LTRs and L1 LINEs and the corresponding increase in SINE coverage during development [31] are clearly evident from Figure 1 (left column). Also consistent with earlier findings [31], the median MILI early piRNA coverage of individual RTE families is highly correlated with the median coverage from MIWI2 early piRNAs, but not with the coverage from MILI late piRNAs (Additional File 1, Figure S1).

Higher piRNA coverage of younger elements

The RepBase database of repeated sequences [40] contain consensus sequences for individual RTE families, and RepeatMasker [15] annotation is based on sequence similarity to these consensus sequences. As the vast majority of RTE loci are under no negative selection (but see, for example [41-43]) the level of divergence between genomic loci and the RepBase consensus sequence can be taken as a proxy for the age of the RTE family. When plotting median piRNA coverage of RTE families against their median divergence, a clear trend of highly covered RTEs being relatively young is observed irrespective of RTE type (Figure 2). For all types of RTEs, the average piRNA coverage of younger elements is significantly higher than coverage of older elements (Additional File 1, Table S2). Also, the high level of variation in piRNA coverage between individual RTE loci is evident from the percentiles shown in Figure 2.
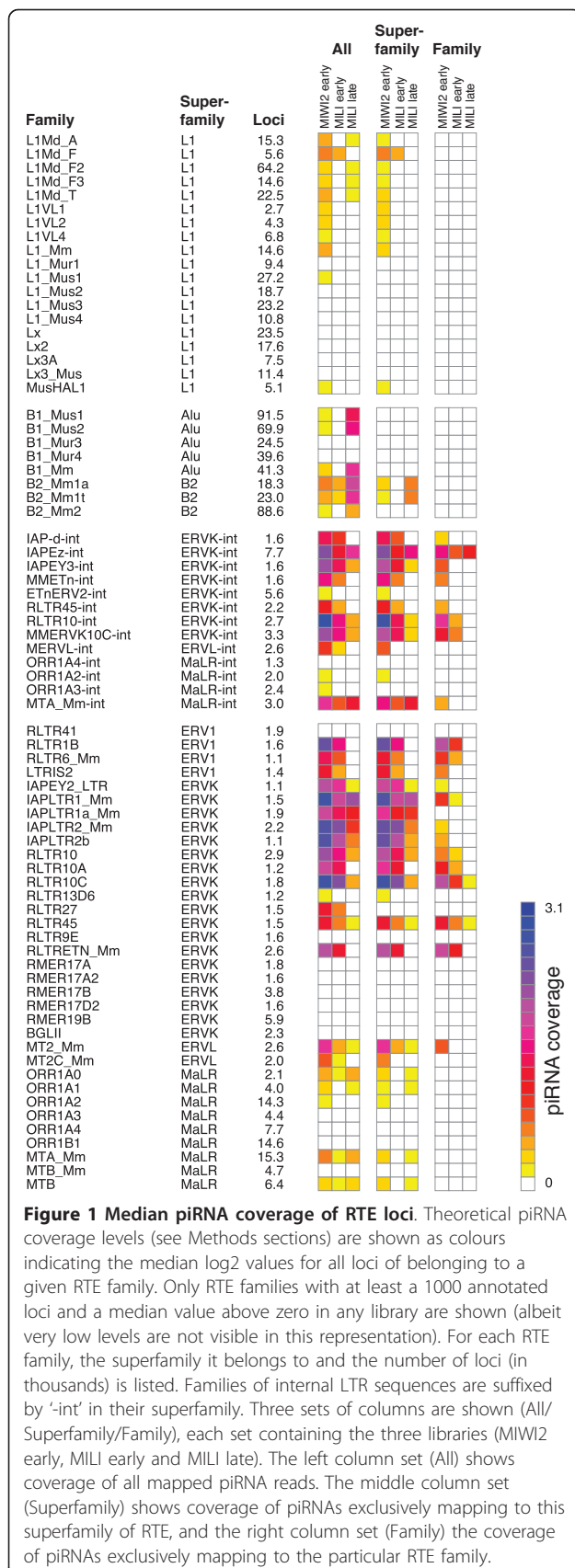
Gene expression levels and piRNA coverage

To test the piRNA coverage in the genomic context of protein-coding genes, all known genes with at least 20 kb (kilo base pairs) to the nearest neighbouring gene (in both directions) and with available Affymetrix expression data from testis tissue were retrieved [44]. These 3307 genes were grouped into highly expressed genes (25% highest expression signals, 827 genes), lowly expressed genes (25% lowest, 829 genes), and medium expressed genes (rest, 1651 genes). For each gene set, the piRNA coverage of RTE residing within 10 kb upstream of their annotated transcription start sites to within 10 kb of their

Table 1 Piwi-RNA sequence read libraries

Library	Stage		Raw reads (× 1000)	Mapped reads (× 1000)	Percent of mapped reads covering:			
					LINE (19.6)	SINE (7.6)	LTRint (3.0)	LTRter (7.4)
MIWI2 early	16.5 dpc	prenatal	1,940	1,934	30.2	5.2	34.8	17.0
MILI early	16.5 dpc	prenatal	472	470	19.2	5.3	16.3	11.9
MILI late	10 dpp	postnatal	1,327	1,324	7.7	19.9	15.9	13.3

Three piRNA libraries from reference [31] were used for this study (accession numbers: MIWI2 early, GSM319957; MILI early, GSM319956; MILI late, GSM319953). Mapped reads denote the number of RNA reads mapping perfectly at least once to the mouse genome. The percentage of mapped reads mapping to loci of annotated RTEs are shown to the right. In parentheses under the RTE type the fractions of the mouse genome (mm9) occupied by the RTE sequences according to the RepeatMasker [15] annotation at the UCSC Genome Browser [67] are shown.

Figure 1 Median piRNA coverage of RTE loci. Theoretical piRNA coverage levels (see Methods sections) are shown as colours indicating the median log2 values for all loci of belonging to a given RTE family. Only RTE families with at least a 1000 annotated loci and a median value above zero in any library are shown (albeit very low levels are not visible in this representation). For each RTE family, the superfamily it belongs to and the number of loci (in thousands) is listed. Families of internal LTR sequences are suffixed by '-int' in their superfamily. Three sets of columns are shown (All/Superfamily/Family), each set containing the three libraries (MIWI2 early, MILI early and MILI late). The left column set (All) shows coverage of all mapped piRNA reads. The middle column set (Superfamily) shows coverage of piRNAs exclusively mapping to this superfamily of RTE, and the right column set (Family) the coverage of piRNAs exclusively mapping to the particular RTE family.
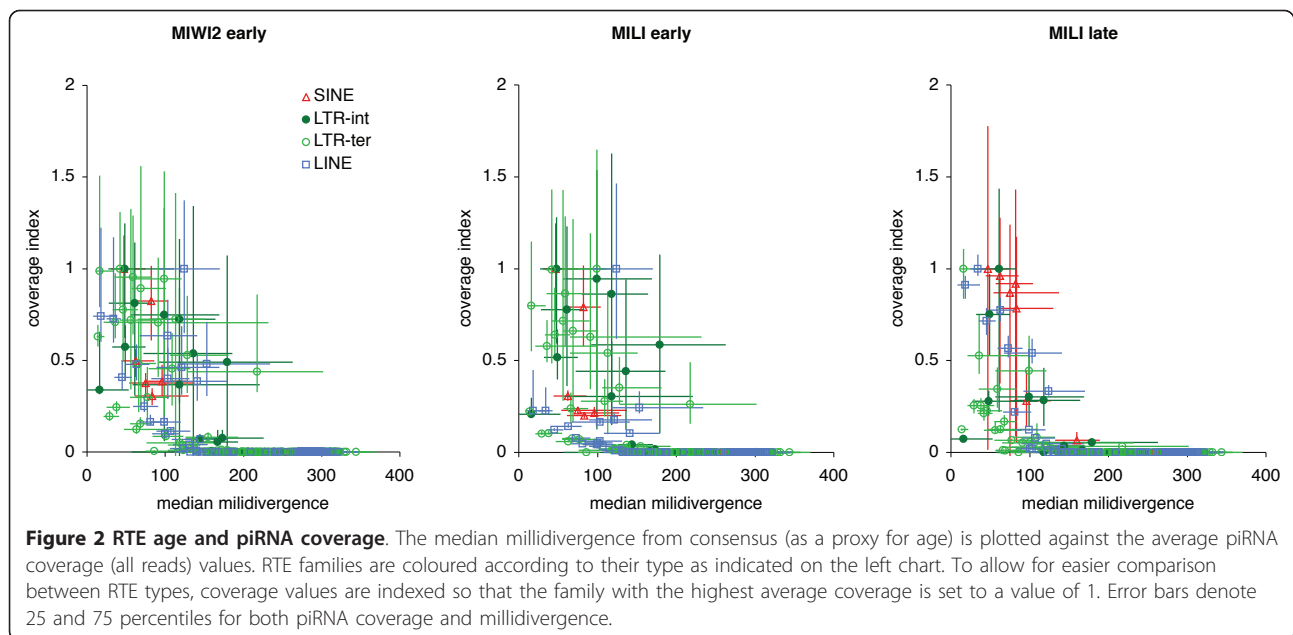
stop sites was recorded (Figure 3). For all types of RTE elements, piRNA coverage in the context of highly expressed genes was significantly higher than both medium and lowly expressed genes. This was found for all piRNA libraries (Additional File 1, Table S3).

For some RTE types, the relative number of young and old loci differ between the vicinities of highly expressed genes and lowly expressed genes, (Additional File 1, Figure S2), suggesting that the higher levels of piRNA coverage of RTE near highly expressed genes could simply be explained by the age of the RTE sequences. However, when repeating the analysis without the youngest RTE sequences, essentially similar results and significance levels are found (Additional File 1, Table S3).

Interestingly, when assessing the piRNA coverage of RTE sequence near transcription start sites (TSS), peaks are observed immediately upstream of TSS on the reverse strand, and for piRNAs not targeting RTEs, also immediately downstream of TSS on the forward strand (Additional File 1, Figure S3). Such a pattern resembles that of the recently discovered short transcripts generated around TSS (the TSS-associated RNAs) [45,46] suggesting that these piRNAs may in fact be TSS-associated RNAs. It is uncertain if this represents experimental contamination of non-piRNAs or if these TSS-associated RNAs provide the transcripts that a processed into piRNAs, although the presence of RNA reads smaller than the usual 24-30 nucleotides - especially among early MILI piRNAs - hints that a contribution from the former scenario cannot be ruled out (Additional File 1, Figure S4). Assuming all RNAs mapping within 1000 base pairs of an annotated TSS are TSS-associated RNAs and removing these from the analysis does not change any of the presented conclusions (data not shown).

piRNA coverage and distance to genes

The age of RTEs and their genomic distance to protein-coding genes is not independent [13,47]. If RTEs residing near genes are in general relatively young, one would expect these RTEs to display high levels of piRNA coverage as a result of this. To test if proximity to genes affected piRNA coverage independent of RTE age, members of each RTE family were divided into three equally sized groups based on their divergence from their consensus sequence (called 'young', 'median' and 'old' loci, respectively). Within each age-group, members were further divided into sub-groups according to genomic location; i) RTE loci residing inside the boundaries of known genes (termed 'genic'), ii) RTE loci in intergenic regions in proximity to known genes ('proximal'), and iii) RTE loci in intergenic regions distal to known genes ('distal'). The groups 'proximal' and 'distal' defined as loci closer or further away from genes, respectively, than the median distance of all non-'genic' loci from the RTE family.

Figure 2 RTE age and piRNA coverage. The median millidivergence from consensus (as a proxy for age) is plotted against the average piRNA coverage (all reads) values. RTE families are coloured according to their type as indicated on the left chart. To allow for easier comparison between RTE types, coverage values are indexed so that the family with the highest average coverage is set to a value of 1. Error bars denote 25 and 75 percentiles for both piRNA coverage and millidivergence.
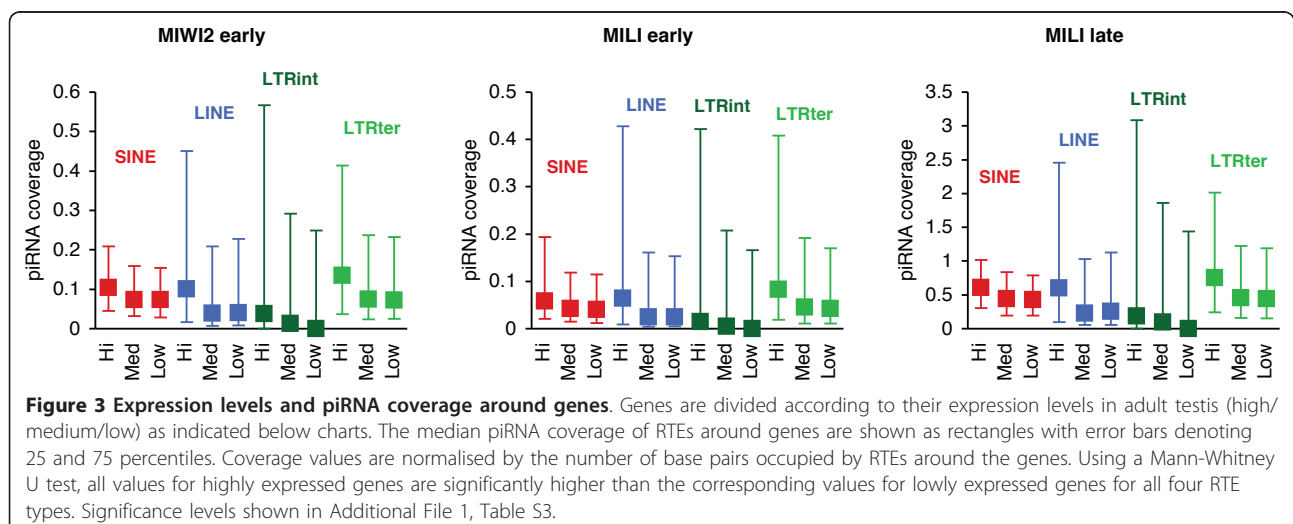
An overview of the grouping procedure is presented in Figure 4. The assumption that younger RTE members tend to reside closer to genes are confirmed by the observation that for 95% of all RTE families, the fraction of loci being proximal to genes is higher for young loci than for old loci (Wilcoxon Signed Rank, $p < 2.2 \times 10^{-16}$; values not shown).
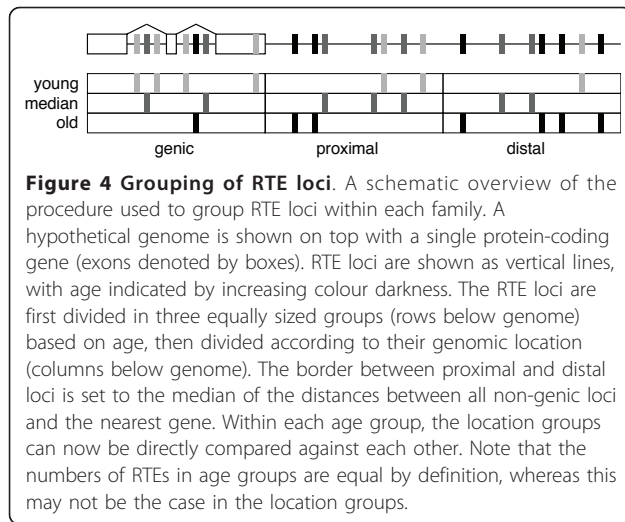
RTE loci proximal to genes have - irrespective of age - significantly higher piRNA coverage than similar RTE loci distal to genes (Figure 5). With internal LTR sequences belonging to the 'old' group as the only exception, loci proximal to genes have significantly higher coverage than genic loci in prenatal development. Interestingly, in postnatal development no RTE group

displays a significantly higher coverage for proximal loci than for genic loci, and furthermore, for all RTE groups, genic loci have significantly higher coverage than loci distal to genes (Figure 5). Thus, coverage by MILI late piRNAs is enriched in genic regions, an observation that is further supported by the fact that the total coverage of MILI late piRNAs mapping to genic regions is increased for all types of RTEs (Figure 6).

Strand bias in piRNA coverage of genic RTEs
Aravin and colleagues [31] showed that in early development, the substrate for piRNA generation is provided by actively transcribed RTE elements. Later in development, active transcription of RTEs should then be

Figure 3 Expression levels and piRNA coverage around genes. Genes are divided according to their expression levels in adult testis (high/medium/low) as indicated below charts. The median piRNA coverage of RTEs around genes are shown as rectangles with error bars denoting 25 and 75 percentiles. Coverage values are normalised by the number of base pairs occupied by RTEs around the genes. Using a Mann-Whitney U test, all values for highly expressed genes are significantly higher than the corresponding values for lowly expressed genes for all four RTE types. Significance levels shown in Additional File 1, Table S3.

Figure 4 Grouping of RTE loci. A schematic overview of the procedure used to group RTE loci within each family. A hypothetical genome is shown on top with a single protein-coding gene (exons denoted by boxes). RTE loci are shown as vertical lines, with age indicated by increasing colour darkness. The RTE loci are first divided in three equally sized groups (rows below genome) based on age, then divided according to their genomic location (columns below genome). The border between proximal and distal loci is set to the median of the distances between all non-genic loci and the nearest gene. Within each age group, the location groups can now be directly compared against each other. Note that the numbers of RTEs in age groups are equal by definition, whereas this may not be the case in the location groups.

repressed, and mRNA sequences from active full-length RTE loci are no longer widespread. This suggests that co-transcription of RTE sequences along with mRNAs from protein-coding genes (predominantly in intronic regions) could now take on a relatively larger role in providing RTE sequence transcripts. A prerequisite for the generation of piRNAs is the presence of transcripts with complementary sequences. Although active RTEs need to be transcribed from their forward strand, the RTE sequences scattered around the highly transcribed genome could produce transcripts in both orientations. But if as suggested, transcribed RTE sequences in postnatal mouse testes are mainly provided from co-transcription with genes, the transcriptional orientation of a given genic RTE loci is to a large extent determined by the orientation of the host gene. One can therefore test if the strand of piRNAs mapping uniquely to genic RTE loci corresponds to the orientation of the RTE relative to the host gene. Of course, amplification from the ping-pong cycle may generate multiple piRNAs, which may map on both strands of a genic RTE locus (although the efficiency of the ping-pong cycle may decrease in postnatal development as MIWI2 is no longer expressed [31]), potentially blurring the picture. As seen from Figure 7, a clear pattern of high sense coverage of genic RTEs in the forward orientation and high antisense coverage of genic RTEs in the reverse orientation is seen for postnatal piRNAs, but not for prenatal piRNAs.
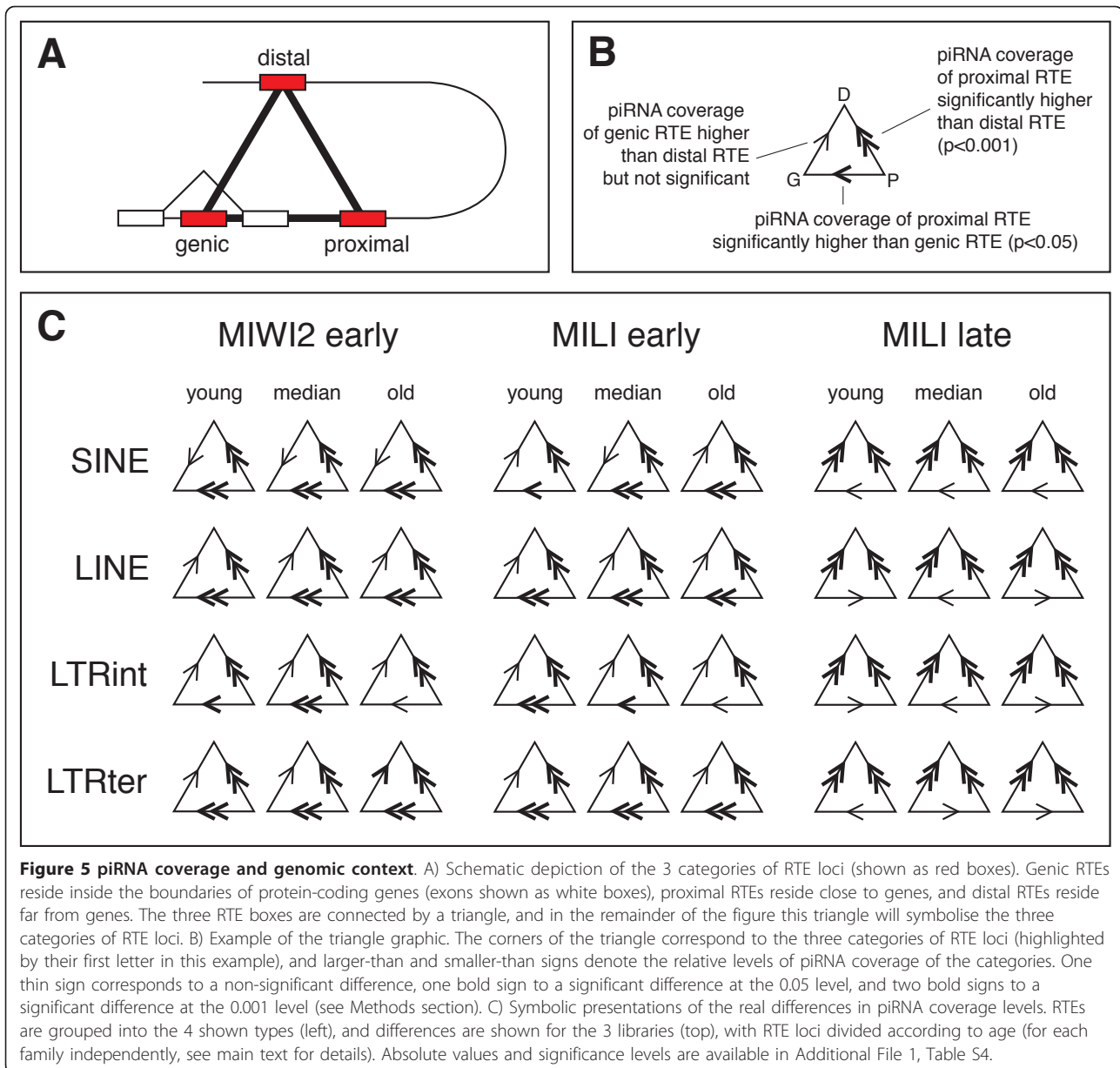
Conclusions

As reported previously, RTE families are targeted very differently by piRNAs in developing mouse testes. By focusing on the total population of RTE loci, the present reanalysis of published data reveal further differences in piRNA targeting between individual members of RTE

families. The available data for this analysis is arguably limited and the presented data relies on a single set of experiments. Although deep-sequencing techniques ideally should provide sequences from all available transcripts in a neutral fashion, biases may be introduced experimentally, especially during construction of libraries [48]. Furthermore, considerable biological differences in RTE sequences have been reported between mouse strains [49,50]. Nevertheless, the vast majority of RTE sequences will be shared among all extant mice, and the results presented here are all of a global genomic character with no predictions for individual loci, suggesting a fair generality of the findings.

Transcriptional activity is correlated between genomic regions residing near each other [51-53], and the observation that piRNA targeting of RTEs is higher around highly expressed genes, may simply reflect that transcription of RTEs is more permissible near highly expressed genes. A correlation between transcription levels of LTR sequences and their neighbouring genes has previously been reported in fission yeast [54]. This further supports the notion that RTE transcripts are not specifically recognized as RTEs by the Piwi proteins, but are largely triggering the piRNA response in a manner proportional to their presence. It should be stressed that the reported preference by MILI for sense RTE sequences and the corresponding preference by MIWI2 for antisense sequences [31] suggest some level of discrimination of transcripts.

In postnatal testis development, piRNA targeting is shifted towards loci residing in introns of protein-coding genes. If, as assumed, active transcription of RTE loci is repressed at this stage, one would expect a higher proportion of RTE sequences in the total transcriptome to be derived from co-transcription of intronic RTE loci. This observation could at least in part explain the previously observed increase in piRNAs targeting SINE elements in postnatal stages [31], as SINE elements are the most abundant RTEs in introns (Additional File 1, Figure S5). Therefore, the increased piRNA response directed at SINE sequences does not necessitate transcription of active SINE elements in postnatal development. In fact, as SINE elements are non-autonomous, presumably using the enzymatic machinery provided by LINE elements [55,56], there should be no basis for SINE proliferation in postnatal development if the prenatal silencing of LINE persists. Yet, SINE transcription may take place without subsequent transposition, and the known functional effects of mammalian SINE transcription [57-59] and the recently reported SINE RNA toxicity [60] suggest both active SINE transcription in later development, and the possible need for regulation.

On an evolutionary time scale, RTE activity has contributed hugely to the evolution of mammalian genomes
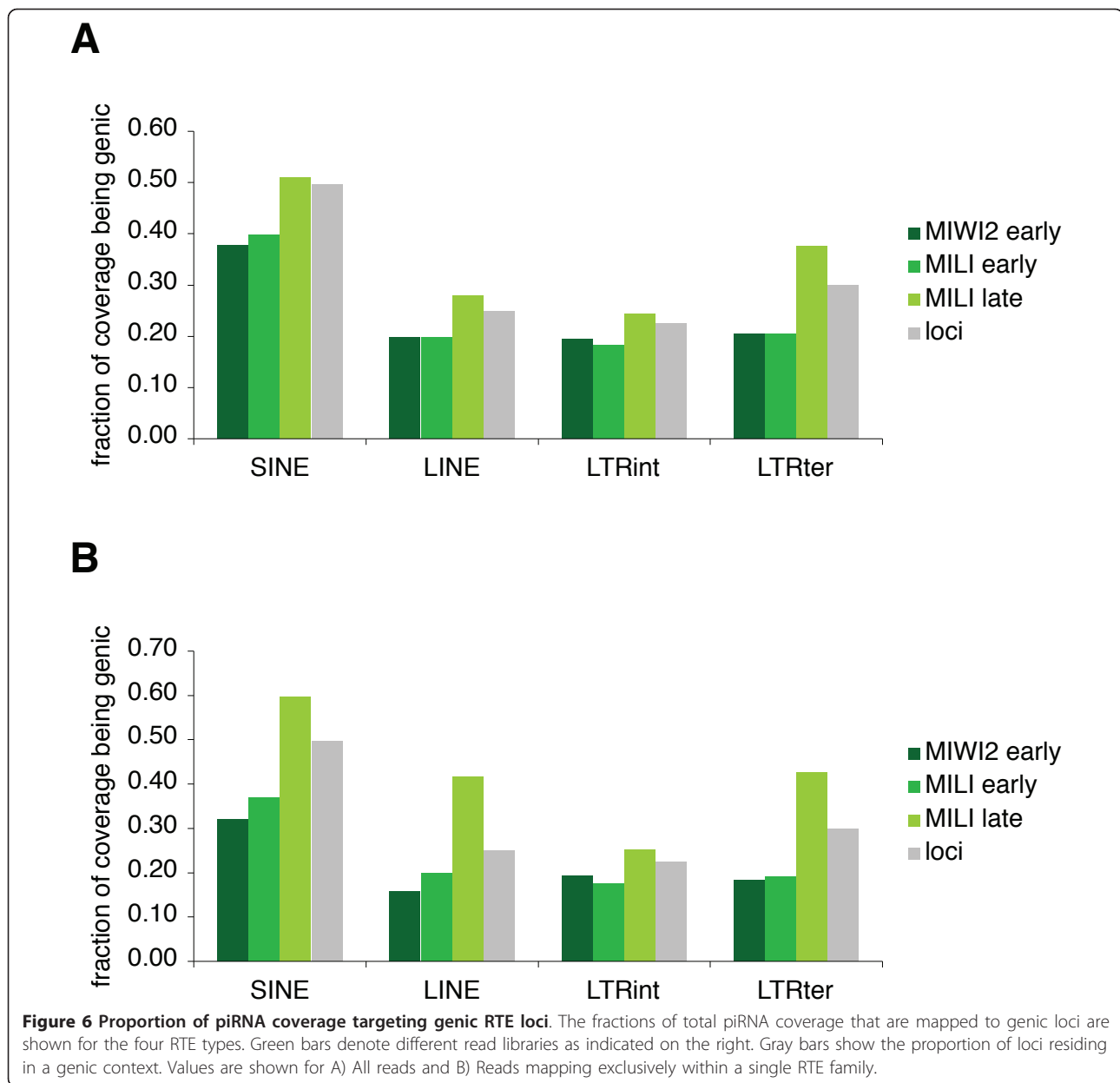
Figure 5 piRNA coverage and genomic context. A) Schematic depiction of the 3 categories of RTE loci (shown as red boxes). Genic RTEs reside inside the boundaries of protein-coding genes (exons shown as white boxes), proximal RTEs reside close to genes, and distal RTEs reside far from genes. The three RTE boxes are connected by a triangle, and in the remainder of the figure this triangle will symbolise the three categories of RTE loci. B) Example of the triangle graphic. The corners of the triangle correspond to the three categories of RTE loci (highlighted by their first letter in this example), and larger-than and smaller-than signs denote the relative levels of piRNA coverage of the categories. One thin sign corresponds to a non-significant difference, one bold sign to a significant difference at the 0.05 level, and two bold signs to a significant difference at the 0.001 level (see Methods section). C) Symbolic presentations of the real differences in piRNA coverage levels. RTEs are grouped into the 4 shown types (left), and differences are shown for the 3 libraries (top), with RTE loci divided according to age (for each family independently, see main text for details). Absolute values and significance levels are available in Additional File 1, Table S4.

[61-64], and when attempting to understand the diversity of present eukaryotic life it is essential to include the history and activity of RTEs. However, RTEs are not just silent passengers that occasionally spring into action, but have to be dealt with within each individual's life history. In this respect, the indirect approach of analysing small RNAs generated to repress RTE activity in the germline may produce further valuable knowledge on the activity of RTEs during development.

Methods
Data and Annotations
Small RNA libraries accession numbers GSM319953 (MILI late), GSM319956 (MILI early) and GSM319957 (MIWI2 early) were retrieved through DeepBase (http://deepbase.sysu.edu.cn/) [39] and mapped to the mouse genome (mm9 assembly) using bowtie [65]. Prior to mapping, reads were filtered and sequences with ambiguous base calls and low-complexity sequences were removed. The latter was done measuring the linguistic complexity [66] of the sequences in 16 bp windows, and excluding reads with an average complexity of less than 0.75. Preliminary tests showed that this would remove highly repetitive reads with very large numbers of genomic mappings (not shown). For each library, this procedure filtered out between 0.14 and 0.17% of all raw reads. RepeatMasker and known gene annotations were downloaded from the UCSC Genome Browser [15,40,67].

Figure 6 Proportion of piRNA coverage targeting genic RTE loci. The fractions of total piRNA coverage that are mapped to genic loci are shown for the four RTE types. Green bars denote different read libraries as indicated on the right. Gray bars show the proportion of loci residing in a genic context. Values are shown for A) All reads and B) Reads mapping exclusively within a single RTE family.

A set of non-overlapping TSS was selected by grouping all known genes according to their assignment to ENSEMBL genes [68]. For each ENSEMBL gene, the most abundant genomic start point was selected. If more than one point was found to have the highest abundance, the one furthest upstream of these was chosen. Gene expression levels were assessed from the 'testis' signal intensities in the Mouse GNF1M Gene Atlas from BioGPS (http://biogps.gnf.org/) [44].

Mapping and coverage
For each RTE loci the number of reads mapped within the locus were recorded. Reads were assigned a score of 1/(number of genomic mappings of read), so that only uniquely mapping reads scored 1. The read score were then divided by the size of the RTE loci (in kilo-base pairs). Finally, scores were divided by the total number (in millions) of mapped reads from the library in question.

Statistical testing
To test for difference between RTE loci from different genomic regions (data presented in Figure 5), all RTE families were first split in 3 groups based on age where after members in each group were divided according to their genomic context (genic, proximal, distal). Hence,

Figure 7 Strand bias. The log2 ratio of sense and antisense coverage from uniquely mapping piRNAs shown for the 4 main RTE types. Each RTE type is divided into genic loci on the forward strand relative to the host gene ('forward'; green columns), genic loci on the reverse strand ('reverse'; orange columns), and loci residing outside genes ('nongenic'; grey columns). All values are shown for the three piRNA libraries.

for each RTE family, nine sets of loci were formed, and the average piRNA coverage for each set was recorded. To test for difference between two groups (for example, between old LINE loci being genic or distal), pairs of average values were collected for the 90 LINE families (Additional File 1, Table S1) and tested using a Wilcoxon Signed-Rank Test. Bonferroni corrections (n = 108 in Figure 5) were calculated as: $p_{corrected} = 1-(1-p)^n$. All statistical analyses were carried out using R [69].

Additional material

Additional file 1: Supplementary Material. Figures S1-S5 and Tables S1-S4. PDF format.

Authors' contributions
TM conceived of the study, analyzed the data and wrote the manuscript.

Competing interests
The authors declare that they have no competing interests.

References
1. Kazazian HH Jr: **Mobile Elements: Drivers of Genome Evolution.** *Science* 2004, **303**(5664):1626-1632.
2. Yohn CT, Jiang Z, McGrath SD, Hayden KE, Khaitovich P, Johnson ME, Eichler MY, McPherson JD, Zhao S, Paabo S, *et al*: **Lineage-specific expansions of retroviral insertions within the genomes of African great apes but not humans and orangutans.** *PLoS Biol* 2005, **3**(4):e110.
3. Akagi K, Li J, Stephens RM, Volfovsky N, Symer DE: **Extensive variation between inbred mouse strains due to endogenous L1 retrotransposition.** *Genome Research* 2008, **18**(6):869-880.
4. Bennett EA, Coleman LE, Tsui C, Pittard WS, Devine SE: **Natural genetic variation caused by transposable elements in humans.** *Genetics* 2004, **168**(2):933-951.
5. Mills RE, Bennett EA, Iskow RC, Luttig CT, Tsui C, Pittard WS, Devine SE: **Recently mobilized transposons in the human and chimpanzee genomes.** *American journal of human genetics* 2006, **78**(4):671-679.
6. Callinan PA, Batzer MA: **Retrotransposable elements and human disease.** *Genome Dyn* 2006, **1**:104-115.
7. Deininger PL, Batzer MA: **Alu repeats and human disease.** *Molecular genetics and metabolism* 1999, **67**(3):183-193.
8. Kazazian HH: **Mobile elements and disease.** *Current Opinion in Genetics & Development* 1998, **8**(3):343-350.
9. Maksakova IA, Mager DL, Reiss D: **Keeping active endogenous retroviral-like elements in check: the epigenetic perspective.** *Cell Mol Life Sci* 2008, **65**(21):3329-3347.
10. Schumann GG: **APOBEC3 proteins: major players in intracellular defence against LINE-1-mediated retrotransposition.** *Biochemical Society transactions* 2007, **35**(Pt 3):637-642.
11. Zamudio N, Bourc'his D: **Transposable elements in the mammalian germline: a comfortable niche or a deadly trap?** *Heredity* 2010, **105**(1):92-104.
12. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, *et al*: **Initial sequencing and comparative analysis of the mouse genome.** *Nature* 2002, **420**(6915):520-562.
13. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, *et al*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**(6822):860-921.
14. Maksakova IA, Romanish MT, Gagnier L, Dunn CA, van de Lagemaat LN, Mager DL: **Retroviral elements and their hosts: insertional mutagenesis in the mouse germ line.** *PLoS genetics* 2006, **2**(1):e2.
15. Smit AFA, Hubley R, Green P: **RepeatMasker Open-3.0.** 1996 [http://www.repeatmasker.org].
16. Stocking C, Kozak CA: **Murine endogenous retroviruses.** *Cell Mol Life Sci* 2008, **65**(21):3383-3398.
17. Mears ML, Hutchison CA: **The evolution of modern lineages of mouse L1 elements.** *J Mol Evol* 2001, **52**(1):51-62.
18. Goodier JL, Ostertag EM, Du K, Kazazian HH Jr: **A novel active L1 retrotransposon subfamily in the mouse.** *Genome Res* 2001, **11**(10):1677-1685.
19. Naas TP, DeBerardinis RJ, Moran JV, Ostertag EM, Kingsmore SF, Seldin MF, Hayashizaki Y, Martin SL, Kazazian HH: **An actively retrotransposing, novel subfamily of mouse L1 elements.** *Embo J* 1998, **17**(2):590-597.
20. Penzkofer T, Dandekar T, Zemojtel T: **L1Base: from functional annotation to prediction of active LINE-1 elements.** *Nucl Acids Res* 2005, **33**(suppl_1):D498-500.

21. Batzer MA, Deininger PL: **Alu repeats and human genomic diversity.** *Nat Rev Genet* 2002, **3(5)**:370-379.
22. Daniels GR, Deininger PL: **Repeat sequence families derived from mammalian tRNA genes.** *Nature* 1985, **317(6040)**:819-822.
23. O'Donnell KA, Boeke JD: **Mighty Piwis defend the germline against genome intruders.** *Cell* 2007, **129(1)**:37-44.
24. Moazed D: **Small RNAs in transcriptional gene silencing and genome defence.** *Nature* 2009, **457(7228)**:413-420.
25. Tang F: **Small RNAs in mammalian germline: Tiny for immortal.** *Differentiation* 2010, **79(3)**:141-146.
26. Klattenhoff C, Theurkauf W: **Biogenesis and germline functions of piRNAs.** *Development* 2008, **135(1)**:3-9.
27. Girard AI, Sachidanandam R, Hannon GJ, Carmell MA: **A germline-specific class of small RNAs binds mammalian Piwi proteins.** *Nature* 2006, **442(7099)**:199-202.
28. Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, Kingston RE: **Characterization of the piRNA complex from rat testes.** *Science* 2006, **313(5785)**:363-367.
29. Aravin AA, Sachidanandam R, Girard A, Fejes-Toth K, Hannon GJ: **Developmentally Regulated piRNA Clusters Implicate MILI in Transposon Control.** *Science* 2007, **316(5825)**:744-747.
30. Carmell MA, Girard A, van de Kant HJ, Bourc'his D, Bestor TH, de Rooij DG, Hannon GJ: **MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline.** *Dev Cell* 2007, **12(4)**:503-514.
31. Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Toth KF, Bestor T, Hannon GJ: **A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice.** *Molecular cell* 2008, **31(6)**:785-799.
32. Kuramochi-Miyagawa S, Watanabe T, Gotoh K, Totoki Y, Toyoda A, Ikawa M, Asada N, Kojima K, Yamaguchi Y, Ijiri TW, *et al*: **DNA methylation of retrotransposon genes is regulated by Piwi family members MILI and MIWI2 in murine fetal testes.** *Genes Dev* 2008, **22(7)**:908-917.
33. Weaver JR, Susiarjo M, Bartolomei MS: **Imprinting and epigenetic changes in the early embryo.** *Mamm Genome* 2009, **20(9-10)**:532-543.
34. Aravin AA, Hannon GJ, Brennecke J: **The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race.** *Science* 2007, **318(5851)**:761-764.
35. Trelogan SA, Martin SL: **Tightly regulated, developmentally specific expression of the first open reading frame from LINE-1 during mouse embryogenesis.** *Proc Natl Acad Sci USA* 1995, **92(5)**:1520-1524.
36. Dupressoir A, Heidmann T: **Germ line-specific expression of intracisternal A-particle retrotransposons in transgenic mice.** *Molecular and cellular biology* 1996, **16(8)**:4495-4503.
37. Kuramochi-Miyagawa S, Kimura T, Yomogida K, Kuroiwa A, Tadokoro Y, Fujita Y, Sato M, Matsuda Y, Nakano T: **Two mouse piwi-related genes: miwi and mili.** *Mech Dev* 2001, **108(1-2)**:121-133.
38. Kuramochi-Miyagawa S, Kimura T, Ijiri TW, Isobe T, Asada N, Fujita Y, Ikawa M, Iwai N, Okabe M, Deng W, *et al*: **Mili, a mammalian member of piwi family gene, is essential for spermatogenesis.** *Development* 2004, **131(4)**:839-849.
39. Yang JH, Shao P, Zhou H, Chen YQ, Qu LH: **deepBase: a database for deeply annotating and mining deep sequencing data.** *Nucleic Acids Res* 2010, **38(Database)**:D123-130.
40. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenet Genome Res* 2005, **110(1-4)**:462-467.
41. Lowe CB, Bejerano G, Haussler D: **Thousands of human mobile element fragments undergo strong purifying selection near developmental genes.** *Proc Natl Acad Sci USA* 2007, **104(19)**:8005-8010.
42. Tsirigos A, Rigoutsos I: **Alu and b1 repeats have been selectively retained in the upstream and intronic regions of genes of specific functional classes.** *PLoS Comput Biol* 2009, **5(12)**:e1000610.
43. Abrusan G, Giordano J, Warburton PE: **Analysis of transposon interruptions suggests selection for L1 elements on the × chromosome.** *PLoS genetics* 2008, **4(8)**:e1000172.
44. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, *et al*: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci USA* 2004, **101(16)**:6062-6067.
45. Jacquier A: **The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs.** *Nat Rev Genet* 2009, **10(12)**:833-844.
46. Seila AC, Calabrese JM, Levine SS, Yeo GW, Rahl PB, Flynn RA, Young RA, Sharp PA: **Divergent transcription from active promoters.** *Science* 2008, **322(5909)**:1849-1851.
47. Medstrand P, van de Lagemaat LN, Mager DL: **Retroelement distributions in the human genome: variations associated with age and proximity to genes.** *Genome Res* 2002, **12(10)**:1483-1495.
48. McCormick KP, Willmann MR, Meyers BC: **Experimental design, preprocessing, normalization and differential expression analysis of small RNA sequencing experiments.** *Silence* 2011, **2(1)**:2.
49. Akagi K, Li J, Stephens RM, Volfovsky N, Symer DE: **Extensive variation between inbred mouse strains due to endogenous L1 retrotransposition.** *Genome Res* 2008, **18(6)**:869-880.
50. Chang-Yeh A, Mold DE, Brilliant MH, Huang RC: **The mouse intracisternal A particle-promoted placental gene retrotransposition is mouse-strain-specific.** *Proc Natl Acad Sci USA* 1993, **90(1)**:292-296.
51. Batada NN, Hurst LD: **Evolution of chromosome organization driven by selection for reduced gene expression noise.** *Nat Genet* 2007, **39(8)**:945-949.
52. Ebisuya M, Yamamoto T, Nakajima M, Nishida E: **Ripples from neighbouring transcription.** *Nat Cell Biol* 2008, **10(9)**:1106-1113.
53. Batada NN, Urrutia AO, Hurst LD: **Chromatin remodelling is a major source of coexpression of linked genes in yeast.** *Trends in Genetics* 2007, **23(10)**:480-484.
54. Mourier T, Willerslev E: **Large-scale transcriptome data reveals transcriptional activity of fission yeast LTR retrotransposons.** *BMC genomics* 2010, **11**:167.
55. Esnault C, Maestre J, Heidmann T: **Human LINE retrotransposons generate processed pseudogenes.** *Nat Genet* 2000, **24(4)**:363-367.
56. Dewannieux M, Esnault C, Heidmann T: **LINE-mediated retrotransposition of marked Alu sequences.** *Nat Genet* 2003, **35(1)**:41-48.
57. Espinoza CA, Allen TA, Hieb AR, Kugel JF, Goodrich JA: **B2 RNA binds directly to RNA polymerase II to repress transcript synthesis.** *Nat Struct Mol Biol* 2004, **11(9)**:822-829.
58. Allen TA, Von Kaenel S, Goodrich JA, Kugel JF: **The SINE-encoded mouse B2 RNA represses mRNA transcription in response to heat shock.** *Nat Struct Mol Biol* 2004, **11(9)**:816-821.
59. Mariner PD, Walters RD, Espinoza CA, Drullinger LF, Wagner SD, Kugel JF, Goodrich JA: **Human Alu RNA Is a Modular Transacting Repressor of mRNA Transcription during Heat Shock.** *Molecular cell* 2008, **29(4)**:499-509.
60. Kaneko H, Dridi S, Tarallo V, Gelfand BD, Fowler BJ, Cho WG, Kleinman ME, Ponicsan SL, Hauswirth WW, Chiodo VA, *et al*: **DICER1 deficit induces Alu RNA toxicity in age-related macular degeneration.** *Nature* 2011, **471(7338)**:325-330.
61. Mourier T: **Reverse transcription in genome evolution.** *Cytogenet Genome Res* 2005, **110(1-4)**:56-62.
62. Krull M, Brosius J, Schmitz J: **Alu-SINE exonization: en route to protein-coding function.** *Mol Biol Evol* 2005, **22(8)**:1702-1711.
63. Medstrand P, van de Lagemaat LN, Dunn CA, Landry JR, Svenback D, Mager DL: **Impact of transposable elements on the evolution of mammalian gene regulation.** *Cytogenet Genome Res* 2005, **110(1-4)**:342-352.
64. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G: **Transposable elements have rewired the core regulatory network of human embryonic stem cells.** *Nat Genet* 2010, **42(7)**:631-634.
65. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10(3)**:R25.
66. Gabrielian A, Bolshoy A: **Sequence complexity and DNA curvature.** *Comput Chem* 1999, **23(3-4)**:263-274.
67. Kuhn RM, Karolchik D, Zweig AS, Trumbower H, Thomas DJ, Thakkapallayil A, Sugnet CW, Stanke M, Smith KE, Siepel A, *et al*: **The UCSC genome browser database: update 2007.** *Nucleic Acids Res* 2007, **35(Database)**:D668-673.
68. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, *et al*: **Ensembl 2007.** *Nucleic Acids Res* 2007, **35(Database)**:D610-617.
69. R Development Core Team: **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL. 2007 [http://www.R-project.org].