CpG Islands as a Putative Source for Animal miRNAs: Evolutionary and Functional Implications

Dvir Dahary,† Reut Shalgi,† and Yitzhak Pilpel*

Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel

†These authors contributed equally to this work.

***Corresponding author:** E-mail: Pilpel@weizmann.ac.il.

Associate editor: Dan Graur

Abstract

MicroRNAs (miRs) are considered major contributors to the evolution of animal morphological complexity. Multiple bursts of novel miR families were documented throughout animal evolution, yet, their evolutionary origins are not understood. Here, we discuss two alternative genomic sources for novel miR families, namely, transposable elements, which were previously described, and a newly proposed origin: CpG islands. We show that these two origins are evolutionarily distinct and that they correspond to marked differences in several functional and genomic characteristics. Together, our results shed light on the intriguing origin of one of the major constituents of regulatory networks in animals, miRs.

Key words: microRNA, animal evolution, CpG islands.

Introduction

microRNAs (miRs) represent a principal layer of gene regulatory networks in metazoans and are hence considered to be major contributors to the evolution of animal complexity (Sempere et al. 2006; Niwa and Slack 2007; Grimson et al. 2008). miRs are short endogenous noncoding RNAs that posttranscriptionally regulate gene expression through interaction with their mRNA targets. Hundreds of animal miRs have been identified in the past decade, potentially affecting thousands of mRNA targets and playing key roles in various pivotal developmental and cellular processes. Recent studies revealed several episodes of expansion in the number of miR families, which correspond to augmentation of complexity during animal evolution (Sempere et al. 2006; Niwa and Slack 2007; Grimson et al. 2008). Remarkably though, despite the enormous potential role ascribed to miRs in evolving animal complexity, their origins, that is, the evolutionary mechanism that gave rise to them, are not yet understood.

Notably, even though there is a common ancestry to plants and animals RNAi pathways, studies suggest that the origins of plant miR genes are divergent from the origins of miR genes in animal genomes as plant and algal miR genes show structure, biogenesis, and targeting properties, which are clearly distinct from animal miRs. Taken together with the absence of miR genes in some fungal species and other intervening lineages, it was concluded that animal and plant miRs had independent origins (Grimson et al. 2008).

High-throughput analyses of small RNAs from closely related fly species elucidated features of the dynamics of miR innovation in invertebrate genome evolution but have not provided specific insights regarding the genomic material for such innovations. Novel miR genes in Drosophila genomes have been shown to emerge either by gene duplication and subsequent functionalization, similar to protein-coding genes, or de novo from random hairpins (Ruby et al. 2007; Lu et al. 2008). Functional copies of miRs that originated from the same ancestral miR gene are classified as members of the same miR family (discussed below) and thus are not pertinent to expansions in the number of miR families. Here, we focus on the origins of novel miR families that were associated with animal complexity, exemplified mainly in the radiation of vertebrates and to a larger extent the mammalian lineages.

In this study, we use the evolutionary classification of miR families and set to explore the mechanisms through which novel miRs have emerged in animal genomes. We discuss two distinct routes for the introduction of novel miRs during evolution. The first, which was reported in recent studies, involves transposable elements (TEs). We further suggest here a second potential source of miR innovations: CpG-rich regions and specifically CpG Islands (CGIs). We show that these two groups of miRs differ in several genomic and functional features and suggest that both routes still serve as an active source for the birth of novel miR families.

Dating of miR families, that is, estimating the evolutionary lineage in which they were introduced into animal genomes, might provide preliminary information pertinent to their potential origin. As a basis for the analyses described below, we classified 670 human miRs from miRbase13.0 (Griffiths-Jones et al. 2006) utilizing three distinct approaches. First, individual miRs with a likely common ancestor were grouped into miR families. Independently, miRs that reside close to each other in the human genome were grouped into miR clusters (as in Shalgi et al. 2007). Note that the family

Letter

and cluster affiliations do not necessarily overlap, that is, miRs from a given family may reside in different genomic clusters, and miRs in a cluster might belong to distinct families. We further classified both families and clusters into four distinct groups by their probable evolutionary age as inferred by their identification in various species throughout the animal kingdom (using 64 available metazoan genomes; see Methods). We designated primate-, mammal-, and vertebrate-specific and older miRs (the ones that originated before vertebrates radiation) as *PRIM*, *MAMM*, *VERT*, and *OLD*, respectively (see Supplementary Material online and Sempere et al. 2006). We used an inclusive approach in determining the age of both miR families and miR clusters, that is, the oldest miR in a family or in a cluster defines the age of its entire class. For instance, a family or a cluster is considered *VERT* if its oldest miR member has its most remote homolog in a nonmammalian vertebrate. We used our final data set, summarized in supplementary table S1 (Supplementary Material online), to gain new insights into the origins of miRs.

Recently, a few studies implied on the TE-associated origin of numerous mammal- and primate-specific miRs. Smalheiser and Torvik (2005) reported 11 instances of presumably TE-derived mammalian miRs. These miRs showed sequence complementarity with many mRNAs that harbor copies of these TEs in their 3′ UTRs. A later work by Piriyapongsa et al. (2007) identified dozens of miRs overlapping TEs in the human genome comprising ~12% of human miRs in their data set. These miRs reside within TE copies of all four major TE classes including short interspersed repetitive elements (SINEs), long interspersed repetitive elements (LINEs), long terminal repeats (LTRs), and DNA transposons, suggesting that the formation of novel miRs from these elements has occurred in several events during the human genome evolution. Furthermore, Lehnert et al. (2009) showed evidence that *Alu*-derived miRs target *Alu* sequences in the human genome implying on the functional role of these miRs in the repression of *Alu* elements activity.

Examining the genomic locations of miRs in the human genome, we first set to identify the ones that completely overlap TE-derived genomic repeats and hence are likely to have originated from ones. All in all, 147 of the 670 miRs in our compilation were found to overlap genomic repeats in the human genome, including repeats from the four main classes—SINEs, LINEs, LTRs, and DNA transposons—as mapped in the UCSC genome browser (Karolchik et al. 2003), applying RepeatMasker (Smit 1996) for the identification and classification of repeats (see supplementary table S2, Supplementary Material online). These TE-derived miRs comprise 22% of the current collection of documented human miRs, exceeding the 12% previously reported (Lehnert et al. 2009). We further regarded each miR family as TE derived if at least one representative of the family overlapped a genomic repeat. We found that 31% of the primate-specific and 21% of mammal-specific miR families showed association with TEs (summarized in supplementary table S3, Supplementary Material online and fig. 1*B*). In accordance with the assumption that the sequence relics of the TE in the genome decay with evolutionary time, very few vertebrate miR families are associated with TEs (see Supplementary Material online) and none of the *OLD* families.

Our findings are in agreement with recent reports regarding the association between miRs and TEs. However, this mode of miR innovation accounts for the origin of ~30% of the evolutionarily young miR families and 22% of the whole human miR collection. The ambiguity regarding the origins and evolution of other miR families and particularly ancient miRs thus remains.

We now set to examine the hundreds of miRs that were not associated with genomic repeats, searching for other sequence characteristics that might imply on their origin. To this aim, we analyzed the nucleotide content of all human miR precursors and their flanking genomic regions. As might be expected from the constraint on the stability of RNA secondary structure, we found that the average GC content in miR hairpins is considerably higher than the average GC content of the entire genome (~50% compared with 41%; Lander et al. 2001).

We further examined the profiles of dinucleotides in the miR flanking regions (±2 kb, masked for TEs and exons, excluding the miR sequence) and compared them with the genomic averages (Simmen 2008). Interestingly, for only one specific dinucleotide—CpG, we found a very unique pattern of enrichment close to the miR and decay in such enrichment as a function of distance from it as demonstrated in figure 1*A* (for all dinucleotide profiles, see supplementary fig. S1, Supplementary Material online). In a window of ~300 bp upstream and downstream the miR position, the observed-to-expected ratio of CpG is significantly higher than the genome average. Recent studies reported unique patterns of methylation in human tumors of miRs residing in proximity to CGIs (Weber et al. 2007; Lujambio et al. 2008). Concurrently, we would like to hypothesize here that some miRs were actually originated from these unique CpG-rich regions.

To test our hypothesis, we looked for miRs (pre-miR sequences) that physically overlap annotated CGIs in the human genome. Remarkably, we found 65 human miRs overlapping CGIs (59 of these are fully contained within CGIs), none of them were TE associated. Thus, more than 12% of the non-TE miRs reside within CGIs. Notably, CGIs occupy less than 1% of the human genome, setting the observed level of overlap between miR genes and CGIs highly significant (P value $= 8 \times 10^{-57}$; see Supplementary Material online). Controlling for possible biases as miRs are often transcribed as polycistrons and found in close proximity to each other on the genome, we still observed that more than 12% of miR clusters are CGI associated (P value $= 2 \times 10^{-49}$; supplementary table S2, Supplementary Material online). This overrepresentation remains highly significant even when comparing it with a variety of different background models (as opposed to the entire genome), which take into account other constraints that may occur on miR genomic localization (see Supplementary Material online). In particular, focusing on the sequence character-

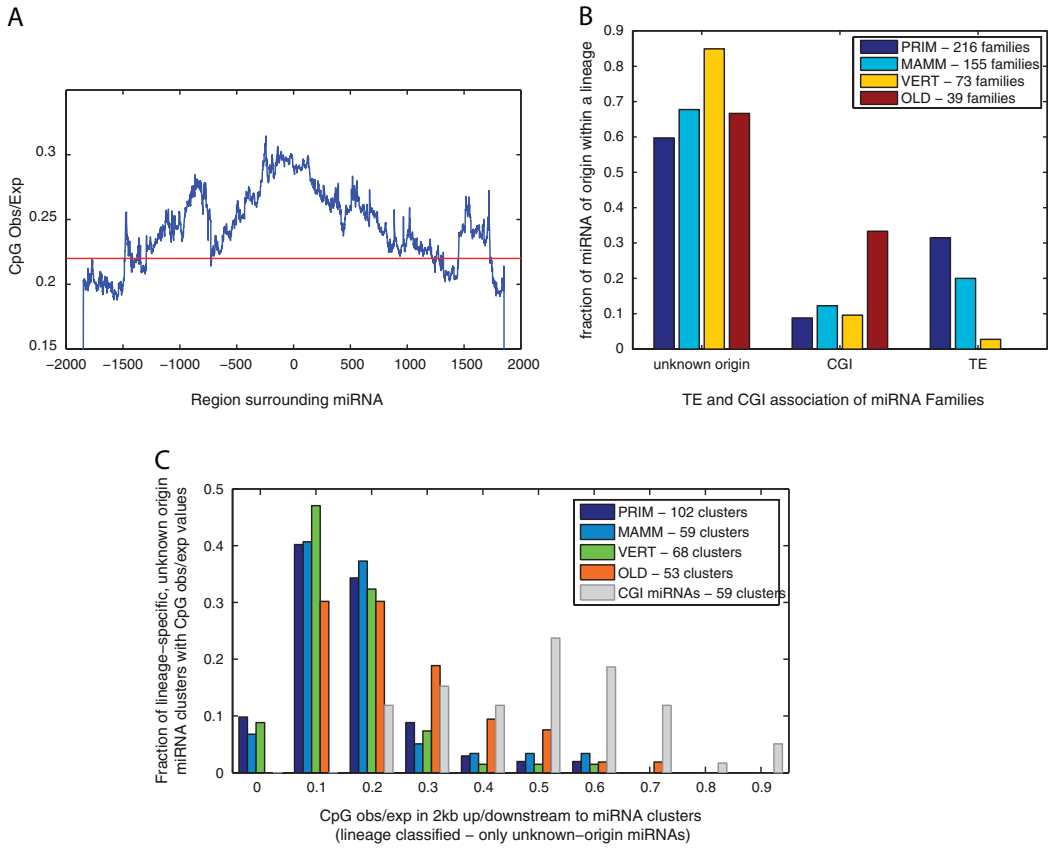

FIG. 1 Association of miRs with CpG-rich regions. (*A*) Averaged observed-to-expected (obs/exp) CpG values, calculated in a running window of 300 bp in ±2-kb flanking regions of human miRs. (*B*) Fractions of miR families from the four lineages that are CGI associated, TE associated, or of unknown origin. Fractions are within each lineage. The number of miR families of each lineage is given in the legend. (*C*) Distribution of observed-to-expected CpG values in ±2-kb flanking regions of miR clusters with unknown origin from each lineage. The distribution of CGI-associated miR clusters is given (light gray) as a reference. The number of miR clusters in each lineage is given in the legend.

istics that might be unique to introns, we compared the subset of intronic miRs with their own genomic context, that is, the introns in their hosting genes, therefore controlling for local GC content, expression potential, and other genomic characteristics that might be locally related, and our hypothesis still holds highly significant. We thus demonstrate that the observed phenomenon is significant by itself and not as a by-product of other confounding genomic features.

Examining these findings in the context of miR families and their evolutionary age, the results are striking. Whereas approximately 10% of *PRIM*, *MAMM*, and *VERT* miR families are CGI associated, one third of the *OLD*, that is, the most conserved miR families that are common to species from across the animal kingdom, are found within CGIs (fig. 1*B*; supplementary table S3, Supplementary Material online). To support the idea that CGIs were the source for these 13 *OLD* miR families, we examined their location in the chicken genome. We found representatives of 11 of these families in the chicken genome, 6 of them fully overlapped with annotated CGIs, further implying on CpG-rich regions as the origins of these miRs.

CGIs are essentially short regions comprising about 1% of the sequence of mammalian genomes that show elevated levels of observed-to-expected CpG frequencies comparing with the rest of the genome. The extent of CpG depletion is correlated with the levels of cytosine methylation. Mammalian genomes, which are globally methylated, show average observed-to-expected CpG values of around 0.2 (Simmen 2008), and CGIs are defined, among other parameters, by values higher than 0.6 (i.e., CGIs are strongly enriched for CpG dinucleotides compared with the rest of the genome; however, even there the observed CpG rate is usually lower than expected simply by GC content). Nonmammalian vertebrates genomes show moderate depletion of CpGs, presumably due to lower levels of methylation, and CpG depletion is negligible in arthropods, for instance, where cytosine methylation is barely detectable (Simmen 2008). In fact, the most common pattern in invertebrates, and therefore in animals, is of "mosaic methylation"- featuring domains of heavily methylated DNA interspersed with methylation free domains (Suzuki and Bird 2008). Several invertebrate genomes analyzed so far show that methylated and unmethylated domains coexist in similar proportions in these genomes. It is therefore postulated that mosaic methylation was ancestral to vertebrate global methylation. Empirical sequencing results (bisulfite sequencing that is sensitive to methylation) show that methylated domains had significantly lower observed-to-expected CpG values when

comparing with unmethylated domains (Suzuki et al. 2007). These observations imply that ancestral animal genomes had long stretches of unmethylated and therefore CpG-rich DNA. Global genome methylation seems to have been introduced during vertebrate radiation, resulting in the pattern most clearly observed in mammalian genomes.

Therefore, the association between miRs and CGIs raises two possible scenarios, namely, either CpG-rich regions serve as genomic material for miRs to emerge or miRs are preserving these regions from their natural decay by methylation and deamination. Actually, these scenarios are not mutually exclusive—it is possible that some miRs were born in CpG-rich regions and once formed could protect these regions from natural decay resulting in a CGI. Importantly though, CGI-associated miRs currently reside in nonmethylated regions (otherwise, these regions would be CpG poor) within CGIs, and this implies that they originated in such regions as we are not aware of a mechanism to elevate CpG richness but rather simply keeping them CpG rich by some kind of protection from methylation. Together, this further supports the strong association between CpG-rich regions and the origin of miRs.

To further characterize the miRs that could not be associated with either CGIs or TEs, we next set to examine the distribution of CpG observed-to-expected values in their flanking regions. Interestingly, whereas most of these miRs reside in genomic regions poor in CpG dinucleotides, a fraction of them show CpG observed-to-expected frequencies that are closer to the range of the CGI-associated miRs (fig. 1C). This implies that some of these miRs were also originated from CpG-rich regions, probably with values below the somewhat arbitrary thresholds of CGI annotation. Notably, many of the miRs that reside in CpG-rich regions are classified as *OLD* miR families, increasing even further the percentage of ancient miRs that could be associated with CpG-rich regions in the human genome. Altogether, these findings strongly support our hypothesis that a significant portion of animal miRs was derived from local CpG-rich regions or CGIs.

Following these findings, we further hypothesized that the different evolutionary origins of miRs might correspond to distinct functional and genomic features. To test this assumption, we now set to examine several characteristics of TE- and CGI-derived miRs. miRs can be either "intronic", that is, reside within an intron of an existing host gene and thus be transcribed along with it or "intergenic"- residing outside the boundaries of known genes being expressed independently. In less frequent cases, miRs reside within exons or on the opposite strand in introns of known genes. Examining this feature from an evolutionary perspective, we find that there seems to be a trend for newer miRs to reside more inside introns. While only ~20% of *OLD* miR clusters are intronic and more than 60% are intergenic, about 50% of the *MAMM* and *PRIM* miRs are intronic (see supplementary fig. S3, Supplementary Material online). This trend might be attributed to the distinct origins of these groups. Intriguingly, whereas 36% of the total human miR collection is intronic, only 22% of the CGI-associated miRs reside within introns of known genes compared with more than 50% of the TE-associated miRs (supplementary fig. S2, Supplementary Material online). Examining the whole human genome, we find that approximately 50% of the total sequence of both TEs and CGIs reside in introns. It thus seems that TE-derived miRs have a slight preference to reside within introns, whereas CGI-derived miRs show a marked depletion from these regions and a preference to reside in intergenic regions. These findings imply on a preference for TE-derived miRs to exploit the expression of existing genes for their function and perhaps for their generation too. Presumably, the birth of a novel TE-derived miR may be more easily facilitated when a copy of the TE resides in an intron of an already transcribed gene. This copy can then neutrally acquire the nucleotide substitutions forming a precursor miR hairpin while already being expressed together with its host mRNA. This fulfills the two basic requirements for novel miR birth—forming a hairpin secondary structure and being transcribed. CGI-associated miRs, on the other hand, might fulfill these requirements by other means. First, they reside in regions with intrinsic potential to be transcribed independently (Sandelin et al. 2007). Furthermore, their sequence is rich in CpG dinucleotides, which are self-complementary and forming three hydrogen bonds, thus increasing the stability of the potential hairpin.

We next examined the tissue expression of miRs from different lineages and different origins in 17 normal tissues using a published expression atlas of miRs (Landgraf et al. 2007). Notably, the distributions of the number of tissues in which miRs are expressed clearly differ between lineages such that the older the miR, the more broad the expression it shows across normal tissues (fig. 2A). One possible explanation is the different nature and origins of the miRs in each age group. Indeed, the expression patterns of miRs classified by their predicted origins indicate that TE-derived miRs tend to be more tissue specific than CGI-derived miRs (fig. 2B) (KS-test P value = 0.0053). The most profound difference is exemplified by the fact that more than 50% of the TE-derived miRs are not expressed in any of the normal tissues represented in the atlas comparing with only ~30% of the CGI-derived miRs. This further implies that a larger fraction of the TE-derived miRs might be nonfunctional. In order to try and differentiate between the age- and origin-related trends and to ask whether TE-derived miRs and CGI-derived miRs actually differ in their expression distribution irrespective of their age, we separately analyzed only the new miRs (*PRIM* and *MAMM*) classified by their TE versus CGI origins. Here we observed a slight, however, nonsignificant trend for broader expression of CGI-derived miRs compared with TE-derived miRs. Indeed, a larger fraction of TE-derived miRs are not expressed in any of the examined tissues, whereas more CGI-derived miRs are expressed in more than a few tissues (fig. 2C). Thus, both the age of a miR and its origin correspond with its tissue expression distribution. Interestingly, a recent study (Liang and Li 2009) showed that miRs that are not expressed at all in the examined tissues have been

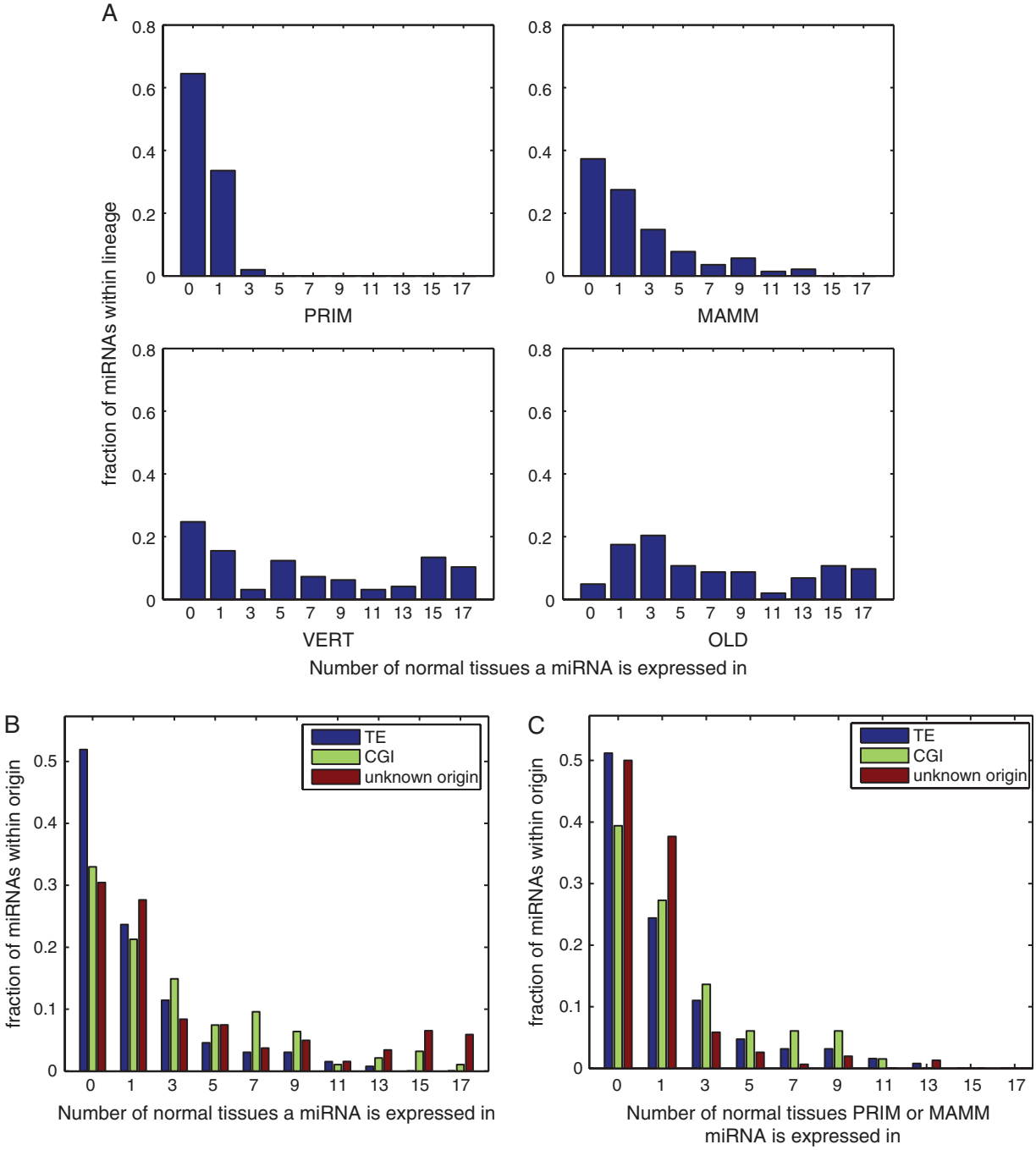

Fig. 2 Expression patterns of miRs in normal tissues. (*A*) Distributions of number of tissues miRs from each lineage are expressed in, out of 17 normal human tissues. Fractions are within each lineage. (*B*) Distributions of number of tissues of all miRs from each ascribed origin. Fractions are within each origin. (*C*) Distributions of number of tissues of primate- and mammal-specific miRs from each ascribed origin. Fractions are within each origin.

subject to weaker selection than those that are expressed in at least one tissue. These findings, together with the above observation that more TE-associated miRs are scarcely expressed, further support the notion that the origin of a novel miR might also determine its probability to be selected for in evolution.

Our findings indicate that both the activity of TEs and the existence of CpG-rich regions account, independently, for the constant supply of novel miR families throughout animal evolution. It is noteworthy that most of the human miR families, especially the evolutionarily young ones, could not be associated neither with TEs nor with CGIs, suggesting that there are other routes of miR innovation. Nonetheless, the characteristics of these miRs bare some hints for their origins. For instance, some of these miRs reside in CpG-rich regions that escaped the annotation of CGIs (fig. 1C) and thus might belong to the CGI-derived miRs. The expression patterns of the young miRs of yet an unclassified origin (fig. 2C) were similar to that of TE-derived miRs; they tend to be expressed only in a very

few tissues or in none at all, whereas the old miRs of yet unknown origin tend to a slightly broader expression. Future analyses might reveal other sources for miR innovation along animal evolution and their implications on the functional roles of these miRs. For instance, a work by Scott et al. (2009) reported 20 mammalian miRs that show similarity to H/ACA small nucleolar RNAs (snoRNAs) implying on an evolutionary relationship between miRNAs and snoRNAs. A few of these snoRNA-associated miRs overlap TEs and therefore were classified in our analysis as TE derived; however, none of them were classified as CGI derived.

Together, our findings regarding the origins of miRs in animal genomes propose new insights not merely on how novel miRs were introduced throughout animal evolution but also on how their origins have influenced the evolution of their functions. Importantly, this result is the first to ascribe an indirect but primary role for CGIs in the evolution of animal complexity.

Methods

The genomic locations of 676 human miRNAs were examined, first excluding the four miRNAs that have multiple genomic positions and other two miRNAs that were excluded from miRNABase. miRs and their RNA hosts, CGIs, and genomic repeats data were downloaded from the UCSC annotation database for the human genome (hg18; Karolchik et al. 2003).

Genomic clusters of miRNAs were defined as neighboring miRNAs with less than 10-kb genomic distance as described in Shalgi et al. (2007). In the case of heterogenous clusters, containing different lineage miRs, further examination might be required to determine the exact age of the cluster as opposed to the inclusive approach taken here. A cluster was considered as TE or CGI associated if at least one of its miRNA members was TE or CGI associated. miR families were grouped based on miR names as in Sempere et al. (2006).

miR families were assigned an evolutionary age by their existence in various species in the animal kingdom using the available miR annotations in 64 metazoan genomes available at miRbase. Each miR family was classified as primate, mammal, or vertebrate specific or older (the ones that originated before vertebrates radiation) and designated *PRIM*, *MAMM*, *VERT*, or *OLD*, respectively.

The statistical significance of the association between miRs and CGIs was evaluated by computing a Poisson distribution for the probability of genomic segments with the average length of miRs to overlap CGIs. Several alternative backgrounds were considered and are described in detail in the Supplementary Material online, with the formula and the relevant figures.

Dinucleotide frequencies were measured in running windows of 100 bp in the 2 kb upstream and downstream to each miR excluding the pre-miR sequence and were then normalized to the product of each two individual nucleotide in the window.

Expression data were downloaded from Landgraf et al. (2007) and filtered to contain only 17 human normal tissues: hsa_Cerebellum-adult, hsa_Frontal-cortex-adult, hsa_Midbrain-adult, hsa_Hippocamp-adult, hsa_Liver, hsa_Heart, hsa_Spleen, hsa_Pituitary, hsa_Thyroid, hsa_Pancreatic-islets, hsa_USSC, hsa_Ovary, hsa_Testis, hsa_Uterus, hsa_Placenta, hsa_Epididymis, and hsa_Prostata. A miR was considered expressed in a tissue if it had one or more clones in that tissue. miRs that did not appear in the data were excluded from the tissue-counts analysis.

Supplementary Material

Supplementary material, tables S1–S3, and figures S1–S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Moshe Oren and the Pilpel laboratory for stimulating discussions. Work in the authors' laboratories is supported by European Commission (EC) FP7 funding (ONCOMIRS, grant agreement number 201102). The EC is not liable for any use made of the information contained herein. We thank the Ben-May Foundation for continuous support.

References

Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res*. 34:D140–D144.

Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar DS, Bartel DP. 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* 455:1193–1197.

Karolchik D, Baertsch R, Diekhans M, et al. (13 co-authors). 2003. The UCSC genome browser database. *Nucleic Acids Res*. 31:51–54.

Lander ESLM, Linton B, Birren C, et al. (255 co-authors). 2001. Initial sequencing and analysis of the human genome. *Nature* 409:860–921.

Landgraf P, Rusu M, Sheridan R, et al. (50 co-authors). 2007. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*. 129:1401–1414.

Lehnert S, Van Loo P, Thilakarathne PJ, Marynen P, Verbeke G, Schuit FC. 2009. Evidence for co-evolution between human microRNAs and Alu-repeats. *PLoS One*. 4:e4456.

Liang H, Li WH. 2009. Lowly expressed human microRNA genes evolve rapidly. *Mol Biol Evol*. 26:1195–1198.

Lu J, Shen Y, Wu Q, Kumar S, He B, Shi S, Carthew RW, Wang SM, Wu CI. 2008. The birth and death of microRNA genes in Drosophila. *Nat Genet*. 40:351–355.

Lujambio A, Calin GA, Villanueva A, et al. (14 co-authors). 2008. A microRNA DNA methylation signature for human cancer metastasis. *Proc Natl Acad Sci U S A*. 105:13556–13561.

Niwa R, Slack FJ. 2007. The evolution of animal microRNA function. *Curr Opin Genet Dev*. 17:145–150.

Piriyapongsa J, Marino-Ramirez L, Jordan IK. 2007. Origin and evolution of human microRNAs from transposable elements. *Genetics* 176:1323–1337.

Ruby JG, Stark A, Johnston WK, Kellis M, Bartel DP, Lai EC. 2007. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res*. 17:1850–1864.

Sandelin A, Carninci P, Lenhard B, Ponjavic J, Hayashizaki Y, Hume DA. 2007. Mammalian RNA polymerase II core promoters: insights from genome-wide studies. *Nat Rev Genet*. 8:424–436.

Scott MS, Avolio F, Ono M, Lamond AI, Barton GJ. 2009. Human miRNA precursors with box H/ACA snoRNA features. *PLoS Comput Biol*. 5:e1000507.

Sempere LF, Cole CN, McPeek MA, Peterson KJ. 2006. The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *J Exp Zool B Mol Dev Evol*. 306:575–588.

Shalgi R, Lieber D, Oren M, Pilpel Y. 2007. Global and local architecture of the mammalian microrna-transcription factor regulatory network. *PLoS Comput Biol*. 3:e131.

Simmen MW. 2008. Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals. *Genomics* 92:33–40.

Smalheiser NR, Torvik VI. 2005. Mammalian microRNAs derived from genomic repeats. *Trends Genet*. 21:322–326.

Smit AF. 1996. The origin of interspersed repeats in the human genome. *Curr Opin Genet Dev*. 6:743–748.

Suzuki MM, Bird A. 2008. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet*. 9:465–476.

Suzuki MM, Kerr AR, De Sousa D, Bird A. 2007. CpG methylation is targeted to transcription units in an invertebrate genome. *Genome Res*. 17:625–631.

Weber B, Stresemann C, Brueckner B, Lyko F. 2007. Methylation of human microRNA genes in normal and neoplastic cells. *Cell Cycle*. 6:1001–1005.