BIOINFORMATICS DISCOVERY NOTE
Vol. 26 no. 8 2010, pages 992–995
doi:10.1093/bioinformatics/btq077

Genome analysis
Advance Access publication February 24, 2010

Biased hosting of intronic microRNA genes

David Golan[1,†], Carmit Levy[2,†], Brad Friedman[3] and Noam Shomron[1,*]
[1]Department of Cell and Developmental Biology, Sackler Faculty of Medicine, Tel Aviv University, Tel Aviv 69978, Israel, [2]Cutaneous Biology Research Center, Harvard Medical School, Charlestown, MA 02129 and [3]Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138, USA

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: MicroRNAs (miRNAs) are involved in an abundant class of post-transcriptional regulation activated through binding to the 3′-untranslated region (UTR) of mRNAs. The current wealth of mammalian miRNA genes results mostly from genomic duplication events. Many of these events are located within introns of transcriptional units. In order to better understand the genomic expansion of miRNA genes, we investigated the distribution of intronic miRNAs.

Results: We observe that miRNA genes are hosted within introns of short genes much larger than expected by chance.

Implementation: We explore several explanations for this phenomenon and conclude that miRNA integration into short genes might be evolutionary favorable due to interaction with the pre-mRNA splicing mechanism.

Contact: nshomron@post.tau.ac.il

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on 16 September 2009; revised on February 14, 2010; accepted on February 22, 2010

1 INTRODUCTION

MicroRNAs (miRNAs) are short ~22 nt RNAs that bind to complementary sequences in the mRNA's 3′-untranslated region (UTR) in order to regulate gene expression through translational silencing, mRNA degradation or cleavage (Ambros, 2004; Bartel, 2004; Filipowicz *et al.*, 2005; Plasterk, 2006). miRNA regulation is an abundant mode of post-transcriptional regulation as each miRNA targets a multitude of mRNAs (Friedman *et al.*, 2009; Lewis *et al.*, 2005; Nielsen *et al.*, 2007). miRNAs were found to be involved in many cellular processes such as growth, differentiation and cell death (Bushati and Cohen, 2007). The current wealth of miRNA genes results mostly from genomic episodes dominated by large duplication events (Hertel *et al.*, 2006) in addition to local expansions (e.g. see Bentwich *et al.*, 2005; Shomron *et al.*, 2009). These events can be located in either intergenic or intronic regions, namely, within introns of transcriptional units (Kim and Kim, 2007; Rodriguez *et al.*, 2004; Shomron and Levy, 2009; Weber, 2005). In order to better understand the genomic expansion of miRNA genes, we investigated the genome-wide distribution of intronic miRNAs. We observe, like others, that miRNA genes are on average more frequently hosted in introns of long genes, though we also notice that their presence within introns of short genes is much larger than expected by chance. We discuss several explanations for this phenomenon and link this observation to differential pre-mRNA splicing of short versus long introns.

2 METHODS

2.1 Simulating miRNA distribution

Given an miRNA density value, we simulated ~26 K genes of the same sizes as the set of human genes and randomly allocated miRNAs to those genes according to the given density. This enables us to estimate shape, mean and variance of various statistics such as the average size of an miRNA hosting gene, and the distribution of miRNA hosting genes' population.

We used this data to reject the hypothesis that the average length of miRNA hosting genes is drawn from the expected distribution given the random insertion model. Moreover, using Kolmogorv–Smirnov test, we rejected the hypothesis that the lengths of miRNA host genes are sampled from the expected distribution given the random insertion model.

We used regression analysis to show a negative correlation between the miRNA density of a gene and the gene's length (the number of hosted miRNAs divided by the gene's length).

For host-intron length conservation, we used Bartlett's statistical test in order to see whether two sets of samples have the same variance. This test is robust to departures from the normality assumption.

2.2 Splicing knockdown and mRNA/miRNA profiling

HeLa cells were transfected using HiPerFect according to the manufacturer's protocol (Qiagen, Hilden, Germany) with 100 pmol siRNA oligos (against Prp8 and U2AF65 or scrambled control; Ambion, TX, USA) per well (0.5×10^6 cells). RNA was extracted 20 h later for high-throughput real-time PCR analysis on ABI miRNA TLDA arrays or using Affymetrix GeneChip Arrays 2.0 (also see Nielsen *et al.*, 2007). Splicing factors were reduced; yet the level of splicing was not drastically affected at this short time point observed by comparing the levels of all miRNA host gene transcripts. Each experiment was repeated four times. For northern blot analysis, 10 µg of total RNA were resolved in a 12% urea–polyacrylamide gel (BioRad, CA, USA) and transferred onto membrane (Hybond-N, Amersham, NJ, USA). Marker and RNA probes directed against mature let-7c, let-7g and tRNA (IDT, IA, USA) were radiolabeled (mirVana, Ambion) with ^{32}P. Prehybridization and hybridization were carried out for 1 and 12 h, respectively, at 42°C (ULTRAhyb-Oligo; Ambion). The membrane was washed for 15 min at 42°C in 2×SSC and 0.1% sodium dodecyl sulfate, exposed and scanned using a Storm PhosphorImaging system (Molecular Dynamics, NJ, USA).

3 RESULTS

3.1 Intronic miRNAs are not randomly distributed within genes

In order to address genomic distribution of miRNAs within introns, we first remapped the genomic positions of these miRNA genes,

[*]To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

obtained from the Refseq database. We identified 167 miRNAs located in introns (based on Refseq gene annotation), which represents 31.5% of the human miRNA repertoire, similar to previous observations (Kim and Kim, 2007; Rodriguez *et al.*, 2004; Weber, 2005). We note that in our study we do not segregate between protein coding and non-coding genes and we take into consideration only miRNAs within introns (ignoring other cases where, e.g., an miRNA overlaps with an exon; 1.8% in our dataset). We also ignore miRNAs which do not reside on the same strand as the hosting gene.

miRNAs mostly evolve through genomic duplication events (Heimberg *et al.*, 2008; Hertel *et al.*, 2006) in addition to some local rapid expansions (e.g. see Bentwich *et al.*, 2005). In order to identify the pattern of integration, we defined our null hypothesis to be an evenly distributed set of miRNA genes scattered among all known genes. In other words, the miRNA genomic density (the number of miRNAs divided by number of bases in a gene) should be constant. Similarly, longer genes would be prone to host more miRNA genes.

The null hypothesis would be true in the event that the mode of miRNA distribution follows a random genomic integration. For control purposes, we simulated a random miRNA integration event where the known intronic miRNAs were stochastically distributed within known genes. We then set to analyze the *bona fide* data. If the null hypothesis is true, we expect that larger genes will have more miRNAs in them in direct proportion to their length since longer genes have a higher chance of being 'hit' by an integrating miRNA. Indeed, as expected, we find that the average length of miRNA-hosting genes is almost three times larger than those of non-hosting genes: 160 kb versus 59 kb (using Refseq gene set).

Interestingly, however, when calculating miRNA densities (miRNAs per gene length) we observe higher densities of miRNAs in short genes with an average length of integration being almost half that of the expected (compare 160 kb to ∼300 kb; P-value <0.0004). This observation was cross-validated by running a linear regression model which showed that the length is negatively correlated with density (P-value <0.015). We note that there is a direct and strong correlation between short introns and short genes. Therefore, everywhere we refer to 'short introns' this could be interchangeable with 'short genes'. Moreover, miRNA hosting introns' sizes in short genes do not differ significantly from their non-hosts counterparts.

Figure 1 shows the percentage of miRNA containing genes as a function of progressive length, where the difference between the expected and observed lengths is significant (Kolmogorov–Smirnov P-value <0.001). The continuous effect observed not only emphasizes the generality of this phenomenon but also suggests that it does not depend on particular odd events. These results do not agree with a previous study (see Zhou and Lin, 2008 and Supplementary Material for further discussion). The difference remains significant even when miRNAs residing on a different strand than the hosting gene are included (data not shown; this eliminates the transcriptional dependence between the miRNA–mRNA expression).

We conclude that the presence of miRNA genes within introns of short genes is much larger than expected by chance. This apparently has a functional importance as comparative study of mice, fly and worm miRNAs gave very similar results (Fig. 1 and data not shown). Finally, we conclude that miRNA integration into introns of host genes is a non-random event.

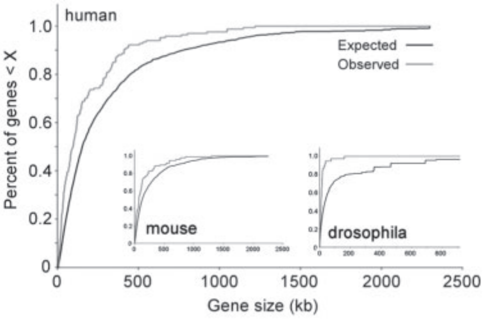

Fig. 1. The percentage of miRNA containing genes as a function of progressive gene length: black, the distribution of gene length in the human genome; Gray, the *bona fide* location of intronic miRNAs. These plots show a significant difference using a Kolmogorov–Smirnov test ($P < 0.001$). Insets show the same plots for mouse and fly miRNA genes which yield a similar P-value. The mouse plot excludes one gene (NM_177386) which hosts 27 miRNAs since it is an outlier. However, results remain significant when including this gene, too.

3.2 Hosting genes' characteristics

To test if transcript expression might be associated with miRNA integration, we divided the set of intronic miRNAs into two groups: (i) those that are expressed from the same strand as the hosting gene and (ii) those that reside on the opposite strand. The P-value of significance suggests that miRNA integration into the opposite strand of DNA is much closer to random ($P = 0.08$) compared with the same strand ($P = 0.0004$). The average length of genes at the same orientation of their embedded miRNAs is 177 kb ($n = 167$) while for the opposing-orientation genes the average length is larger than 209 kb ($n = 28$). Thus, there might be a significant role of transcript expression in miRNA integration and/or maintenance, though we note that these results could be limited by the number of events (low number of miRNAs create a significantly larger standard deviation; see Supplementary Material).

Several alternative explanations were also evaluated, including a gene class preference of miRNA hosting genes. For example, house-keeping genes may provide a good candidate set because they are highly active in expression (Butte *et al*. 2001; Castillo-Davis *et al*. 2002). However, our set of 167 genes did not significantly fall under this annotation category (using GOminer and DAVID; data not shown). Neither did our gene set overlap with hundreds of widely and highly expressed genes (Eisenberg and Levanon, 2003; data not shown).

Chromatin structure plays a role in the accessibility of the genomic DNA for integration events (Quina *et al.*, 2006). We asked whether an open chromatin structure facilitates integration of miRNAs into these regions. For this purpose, we looked at up- and downstream genomic regions just adjacent to miRNA hosting genes. We reasoned that if miRNAs integrate into introns due to high genomic accessibility we will find enrichment of miRNAs in their vicinity, as well. However, there are no miRNAs adjacent to the hosting genes (<10 kb), suggesting that chromatin structure does not participate in the miRNA integration event. In fact, none of the miRNAs was found adjacent (<10 kb) to any house-keeping gene (also see Stark *et al.*, 2005).

Even though miRNA genes originate primarily from duplication events, other mechanisms, such as integration of repetitive genetic elements (Piriyapongsa *et al.*, 2007) and local duplication and mutation, may lead to miRNA 'birth' (Bentwich *et al.*, 2005; Zhang *et al.*, 2007). To exclude the possibility that miRNAs have evolved, *de novo*, in the introns of short genes, we looked into the conservation and family members of intronic miRNAs. We observe that miRNA genes of a common origin, those belonging to the same miRNA family and are known to evolve by duplication events, are prevalent in the intronic miRNA group (e.g. see intronic miR-9,15,23,26,99,103,107,218; Hertel *et al.*, 2006; Weber, 2005). Overall, >30% of the intronic miRNA genes are part of a miRNA family suggesting that most were not generated *de novo* in short introns as a consequence of accumulating mutations.

During pre-mRNA splicing, introns are removed and exons are joined in a multistep multicomponent reaction (see Kim *et al.*, 2008; Wang and Burge, 2008). It was suggested that introns of different lengths are spliced differently (Berget, 1995) and recognition of short introns might involve additional factors that facilitate intron–exon identification (Lim and Burge, 2001). We asked whether short introns are a favorable environment for miRNAs due to interaction with the pre-mRNA splicing reaction. For this purpose, we inhibited splicing by knocking down key splicing factors (see Fig. 2A and Section 2). We reasoned that if the splicing reaction plays a role in miRNA maturation, short genes will behave differently from long ones. We measured the levels of all miRNAs in the splicing inhibited cells, discarded the ones that were not expressed, divided the remaining miRNAs into two groups according to their length and then performed statistics (Kolmogorov–Smirnov test) on the two groups. Our results indicate that the short genes containing miRNAs group are significantly upregulated compared with the cohort of the long genes ($P = 0.002$). Our results were normalized to the expression level of all host genes and take into account that most of the removed miRNAs are those that have no known relevance in HeLa cells (our experimental system) and thus no detectable expression in these cells. We note here that our analysis throughout this study was carried out on mature miRNA forms. We do not think that the observed miRNA increase is due to intron retention in short genes, since we controlled for transcript levels after splicing knockout. In the event of intron retention, we would have expected many of the transcripts to be degraded leaving less rather than more miRNAs. Having said that, we cannot exclude retention of the miRNA-containing transcript in the nucleus leading to increased miRNA levels. Other tests, such as measuring the proportion of intron lengths out of the total length of the gene were not different from the average Refseq gene (see Supplementary Material). Also, tissue-specific expression did not reveal a certain miRNA gene pattern associated with short host genes.

For additional confirmation, we chose two miRNAs residing on either a long or short gene and tested their change in expression following splicing knockdown (Fig. 2A). Only miRNA let-7g, hosted in a short gene (NM_025222; ~24 kb), but not let-7c, hosted in a long gene (NM_001005734; ~413 kb), exhibited an increase in expression in the presence of reduced splicing factors (Fig. 2B). The question raised here is in what manner does the integration of an miRNA into a short intron becomes favorable? If miRNAs in short introns are associated with lower or hindered expression, then one possibility might be integration into these introns in order to exert a mild regulation of potentially many target genes. Overtime, after target genes either gain or loose binding sites for the lowly expressed miRNA a stronger expression can evolve molding an increased regulatory effect that can accommodate high rather than low miRNA expression (see Fig. 3; also see Shomron *et al.*, 2009). Moreover, given this hypothesis one would expect to see higher conservation of the length of the miRNA hosting introns. To test this, we looked at a subset of miRNAs that are hosted in introns in both human and mouse. The estimated mean relative change in both groups was insignificantly different between the groups and insignificantly different from zero (using two-sample and one sample *t*-tests, accordingly). However, the variance of the changes in the host group was 250 times smaller than in the non-hosts group (the variance was significantly smaller with *P*-value <0.0001 using Bartlett's test, see Supplementary Material for more details). We conclude that the sizes of miRNA hosting introns are indeed more conserved, as miRNA hosting introns experience a less drastic change in size. The conservation of intron length has several evolutionary implications. For example, these introns are less likely to receive transposable elements due to detrimental effect on miRNA expression.

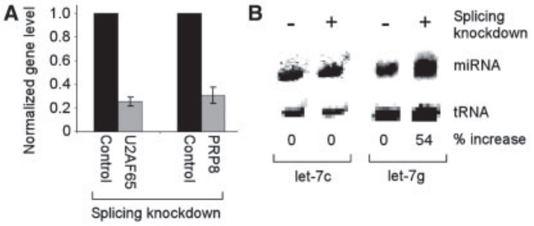

Fig. 2. Splicing knockdown increases miRNAs harbored in short but not long introns. (**A**) Pre-mRNA splicing was knocked down using siRNA directed against two central splicing factors, U2AF65 and PRP8, which led to ~70% reduction in both gene levels. A scrambled siRNA was used as control. The gene level was normalized to control transfections and to beta-actin ($n = 3$). (**B**) Northern blotting of a let-7c and let-7g, which are harbored, respectively, in long and short introns. Mature miRNA is shown. tRNA was used as a loading control. The relative intensity of the miRNA bands was calculated by comparing splicing knockdown versus mock transfection after normalizing to tRNA levels. The TINA software was used to measure band intensity (Raytest, Straubenhardt, Germany).

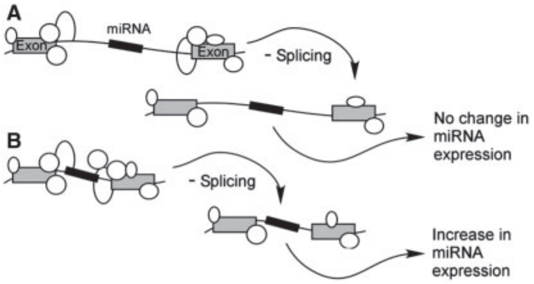

Fig. 3. A proposed model by which splicing inhibits the expression of miRNAs hosted in short introns. The splicing associated factors, present mostly on the exon–intron boundaries, do not affect long (**A**) but affect short (**B**) hosted miRNA genes possibly due to steric hindrance in the vicinity of the miRNA genes. In the absence of splicing factors (see text), there is an increase of miRNA expression only from short but not long introns. Exons are drawn as gray boxes, introns as lines, the miRNA gene as a black box and splicing factors as random sized white circles.

From this, we conclude that a possible mechanism assisting miRNA integration and/or maintenance into short genes is their interconnection with the pre-mRNA splicing mechanism. Along these lines, we have recently observed that intronic miRNAs may be hindered by flanking spliced exons (C. Levy *et al.*, unpublished data). We extend this observation here and show that this interference is possibly more apparent in short genes.

Overall, our results suggest that miRNAs are enriched for in shorter genes; that their integration into short genes might be due to transcript expression levels but not due to genomic accessibility; and that miRNA integration into short genes might be evolutionary favorable due to interaction with pre-mRNA splicing.

ACKNOWLEDGEMENTS

We thank E. Hornstein and the Shomron laboratory members for commenting on the manuscript.

Funding: Work in the NS laboratory is supported by the Chief Scientist Office, Ministry of Health, Israel [3-4876]; Kurz-Lion Foundation; Ori Levi Foundation for Mitochondrial research; Tel Aviv University, Sackler Faculty of Medicine, Schreiber Fellowship.

Conflict of Interest: none declared.

REFERENCES

Ambros,V. (2004) The functions of animal microRNAs. *Nature*, **431**, 350–355.

Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.

Bentwich,I. *et al.* (2005) Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.*, **37**, 766–770.

Berget, S.M. (1995) Exon recognition in vertebrate splicing. *J. Biol. Chem.*, **270**, 2411–2414.

Bushati,N. and Cohen,S.M. (2007) microRNA functions. *Annu. Rev. Cell. Dev. Biol.*, **23**, 175–205.

Butte,A.J. *et al.* (2001) Further defining housekeeping, or "maintenance," genes focus on "A compendium of gene expression in normal human tissues". *Physiol. Genomics*, **7**, 95–96.

Castillo-Davis,C.I. *et al.* (2002) Selection for short introns in highly expressed genes. *Nat. Genet.*, **31**, 415–418.

Eisenberg,E. and Levanon,E.Y. (2003) Human housekeeping genes are compact. *Trends Genet.*, **19**, 362–365.

Filipowicz,W. *et al.* (2005) Post-transcriptional gene silencing by siRNAs and miRNAs. *Curr. Opin. Struct. Biol.*, **15**, 331–341.

Friedman,R.C. *et al.* (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.

Heimberg,A.M. *et al.* (2008) MicroRNAs and the advent of vertebrate morphological complexity. *Proc. Natl Acad. Sci. USA*, **105**, 2946–2950.

Hertel,J. *et al.* (2006) The expansion of the metazoan microRNA repertoire. *BMC Genomics*, **7**, 25.

Kim,Y.K. and Kim,V.N. (2007) Processing of intronic microRNAs. *EMBO J.*, **26**, 775–783.

Kim,E. *et al.* (2008) Alternative splicing and disease. *RNA Biol.*, **5**, 17–19.

Lewis,B.P. *et al.* (2005) Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, **120**, 15–20.

Lim,L.P. and Burge,C.B. (2001) A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl Acad. Sci. USA*, **98**, 11193–11198.

Nielsen,C.B. *et al.* (2007) Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA*, **13**, 1894–1910.

Piriyapongsa,J. *et al.* (2007) Origin and evolution of human microRNAs from transposable elements. *Genetics*, **176**, 1323–1337.

Plasterk,R.H. (2006) Micro RNAs in animal development. *Cell*, **124**, 877–881.

Quina,A.S. *et al.* (2006) Chromatin structure and epigenetics. *Biochem. Pharmacol.*, **72**, 1563–1569.

Rodriguez,A. *et al.* (2004) Identification of mammalian microRNA host genes and transcription units. *Genome Res.*, **14**, 1902–1910.

Shomron,N. and Levy,C. (2009) MicroRNA processing and pre-mRNA splicing crosstalk. *J. Biomed. Biotech.*, **2009**, 594678.

Shomron,N. *et al.* (2009) An evolutionary perspective of animal microRNAs and their targets. *J. Biomed. Biotech.*, **2009**, 594738.

Stark,A. *et al.* (2005) Animal microRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. *Cell*, **123**, 1133–1146.

Wang,Z. and Burge,C.B. (2008) Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. *RNA*, **14**, 802–813.

Weber,M.J. (2005) New human and mouse microRNA genes found by homology search. *FEBS J.*, **272**, 59–73.

Zhang,R. *et al.* (2007) Rapid evolution of an X-linked microRNA cluster in primates. *Genome Res.*, **17**, 612–617.

Zhou,H. and Lin,K. (2008) Excess of microRNAs in large and very 5′ biased introns. *Biochem. Biophys. Res. Commun.*, **368**, 709–715.