Genome Biology and Evolution

SMBE

RESEARCH ARTICLE

Evolutionarily Stable Association of Intronic snoRNAs and microRNAs with Their Host Genes

Marc P. Hoeppner, Simon White,† Daniel C. Jeffares,†‡ and Anthony M. Poole*§*

*Department of Molecular Biology and Functional Genomics, Stockholm University, SE-106 91 Stockholm, Sweden; †Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, United Kingdom; ‡Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, United Kingdom; and §School of Biological Sciences, University of Canterbury, Private Bag 4800, Christchurch 8140, New Zealand

Small nucleolar RNAs (snoRNAs) and microRNAs (miRNAs) are integral to a range of processes, including ribosome biogenesis and gene regulation. Some are intron encoded, and this organization may facilitate coordinated coexpression of host gene and RNA. However, snoRNAs and miRNAs are known to be mobile, so intron-RNA associations may not be evolutionarily stable. We have used genome alignments across 11 mammals plus chicken to examine positional orthology of snoRNAs and miRNAs and report that 21% of annotated snoRNAs and 11% of miRNAs are positionally conserved across mammals. Among RNAs traceable to the bird–mammal common ancestor, 98% of snoRNAs and 76% of miRNAs are intronic. Comparison of the most evolutionarily stable mammalian intronic snoRNAs with those positionally conserved among primates reveals that the former are more overrepresented among host genes involved in translation or ribosome biogenesis and are more broadly and highly expressed. This stability is likely attributable to a requirement for overlap between host gene and intronic snoRNA expression profiles, consistent with an ancestral role in ribosome biogenesis. In contrast, whereas miRNA positional conservation is comparable to that observed for snoRNAs, intronic miRNAs show no obvious association with host genes of a particular functional category, and no statistically significant differences in host gene expression are found between those traceable to mammalian or primate ancestors. Our results indicate evolutionarily stable associations of numerous intronic snoRNAs and miRNAs and their host genes, with probable continued diversification of snoRNA function from an ancestral role in ribosome biogenesis.

Introduction

Noncoding RNAs (ncRNAs) are known to have a diverse range of roles in eukaryotes (Eddy 2001; Mattick 2003; Stefani and Slack 2008). Among the numerous groups of ncRNA described, several abundant classes of small ncRNA with a broad phylogenetic distribution are known, including C/D and H/ACA box small nucleolar RNAs (snoRNAs) and microRNAs (miRNAs). SnoRNAs have well-documented roles in cleavage-based processing and modification, primarily of rRNAs (Kiss 2002), but have also been documented to modify other RNA targets including small nuclear RNAs of the spliceosome (Ganot et al. 1999; Jády and Kiss 2001; Bachellerie et al. 2002; Darzacq et al. 2002). More recently, a role in regulation of alternative splicing of mRNA has been described (Kishore and Stamm 2006). MiRNAs, on the other hand, have well-documented roles in gene regulation across a broad range of species and biological processes. They act to repress gene expression posttranscriptionally through direct pairing to a target mRNA (Bartel 2009; Carthew and Sontheimer 2009). The genomic arrangement of both classes of RNA is varied and includes independent transcripts, genomic clusters consisting of multiple RNAs and residence within the introns of protein-coding genes (Weinstein and Steitz 1999; Mattick 2003; Brown et al. 2008; Royo and Cavaillé 2008).

The intronic location of ncRNAs is interesting in that it represents a situation where two distinct gene products may be expressed from the same transcript. Expression of intronic ncRNAs is largely (though not exclusively) splicing dependent (Hirose et al. 2003; Baskerville and Bartel 2005; Brown et al. 2008). Assuming that expression profiles of both intronic ncRNA and host gene are subject to natural selection, one may envisage several explanations for this arrangement. One is that ncRNAs in introns primarily emerge de novo (Lu et al. 2008) and that a given intronic ncRNA is retained by selection on the basis of it performing some selectively advantageous function within the scope of the host gene expression profile. Another model builds upon the observation that ncRNAs, including snoRNAs, have been documented to be mobile (Weber 2006; Zemann et al. 2006; Schmitz et al. 2008) and may move between genomic locations over evolutionary time via reverse transcription (Volff and Brosius 2007). Mobility may result in a copy of an existing ncRNA becoming intronically located (from some other position, either intronic or not) and being retained at that site because overlap of ncRNA and host gene expression is beneficial. Under both models, which are not mutually exclusive, coexpression of host gene and intronic ncRNA may result in some optimal expression profile for both products, with maximum overlap and minimum trade-off. This might potentially be achieved by switching from one host gene to another (Enerly et al. 2003). Note that the mobility model results in ncRNA duplication (via segmental duplication or retrotransposition), which may in some cases lead to functional divergence of the copies (Volff and Brosius 2007).

Anecdotal observations support evolutionarily stable ncRNA–host gene relationships (Cervelli et al. 2002), mobility (Weber 2006; Schmitz et al. 2008), and segmental duplication (Zemann et al. 2006; Nahkuri et al. 2008). However, short lengths and limited sequence conservation among small RNAs make it nontrivial to distinguish orthology and paralogy. Genome alignments make assignment of

Key words: snoRNA, miRNA, intron, evolution.

E-mail: anthony.poole@canterbury.ac.nz.

Genome. Biol. Evol. 1:420–428.
doi:10.1093/gbe/evp045
Advance Access publication November 5, 2009

orthology between small ncRNAs more reliable than by simple sequence similarity alone, and within this framework it is possible to systematically examine the association between ncRNAs and their host genes (Tanaka-Fujita et al. 2007). We therefore made use of available multispecies whole-genome alignments (Hubbard et al. 2009) to examine the degree to which intron occupancy by miRNAs and snoRNAs is stable across mammalian genomes. For both classes of ncRNA, around 50% of all annotated ncRNAs appear to be intronic in the genomes we studied, and we report a high degree of evolutionary conservation between intronic ncRNAs and their host genes across mammals. Out of the several hundred snoRNAs and miRNAs annotated in the respective genomes (e.g., 717 snoRNAs and 1664 miRNAs in humans), 87 snoRNAs and 103 miRNAs are traceable to the mammalian ancestor using synteny established from genome alignments. Of these, almost all snoRNAs (87/89) and the majority of miRNAs (61/80) are intronic.

At the same time, many snoRNAs and miRNAs are restricted to specific lineages within the mammalian tree, suggesting either ancestral losses or a more recent evolutionary origin. In the case of miRNAs, the latter is generally assumed given the well-documented role this class of ncRNA plays in gene regulation. Although data are emerging to support a broader regulatory role for snoRNAs (Kishore and Stamm 2006; Royo and Cavaillé 2008), such snoRNAs are found in clusters and are generally not intronic (though some may have evolved from intronic snoRNAs [Nahkuri et al. 2008], and some are found in the introns of nontranslated mRNAs [Tycowski et al. 1996]).

We compared the functions of mammalian genes carrying intronic ncRNAs whose intronic positions are stable and ancient (conserved across all 11 mammalian genomes in our data set) with the functions of those that have been in their current location more recently (restricted to primates). For the stable ancient snoRNAs, there appears to be significant overrepresentation of host genes involved in protein synthesis and ribosome biogenesis, whereas no functions are significantly overrepresented among the less stable lineage-specific snoRNAs. Against the backdrop of stable association between ncRNAs and their host genes, this may suggest that snoRNAs have taken on additional roles during the diversification of mammals, in line with suggestions that mammals (and vertebrates, see Heimberg et al. 2008) employ extensive RNA regulatory networks for fine-tuning function and gene expression (Mattick 2001, 2009).

Materials and Methods

Data set

A precompiled genomic alignment of 11 mammals and one bird was retrieved from release 54 of the Ensembl Compara database ("12 amniota vertebrates Pecan," id 338), comprising the following species: *Homo sapiens* (Human), *Pan troglodytes* (Chimpanzee), *Pongo pygmaeus* (Orangutan), *Macaca mulatta* (Macaque), *Rattus norvegicus* (Rat), *Mus musculus* (Mouse), *Canis familiaris* (Dog), *Equus caballus* (Horse), *Bos taurus* (Cow), and *Gallus gallus* (Chicken). We examined the conservation of annotated snoRNAs and miRNAs across this data set. Annotated snoRNAs and miRNAs in release 54 are derived from Rfam (Griffiths-Jones et al. 2003) and miRBase (Griffiths-Jones 2006) databases. The annotation pipelines employ primary, manually curated seed sequences from these databases, as follows. Seed sequences are used in Blast searches against each genome to identify putative ncRNA genes. Because both classes of ncRNA possess secondary structure motifs, in silico folding of sequences is subsequently performed to check for characteristic structural motifs as means to ascertain functionality using covariance models (snoRNAs, see [Nawrocki et al. 2009] or stem-loop folding miRNAs, [Hofacker et al. 1994]), respectively. A description of the Ensembl release 54 annotation pipeline can be found in the FAQs at www.ensembl.org.

Assigning Orthology to ncRNAs using Synteny

SnoRNA and miRNA orthology across species were established using two criteria. The first is simple assignment of homology based on common Rfam and miRBase IDs. IDs in these databases are assigned to ncRNAs based on similarity to the covariance model or seed alignment describing each "family" (each family corresponds to a particular id).

Next, genomic locations of ncRNAs were overlaid onto the genome alignment to identify cases of positional conservation. We draw a distinction between candidate ncRNAs that fall within aligned regions and those that fall outside identified syntenic regions; only the former are used in our analyses (table 1) on the grounds that it is nontrivial to assign orthology for the latter group.

To account for slight positional variations in ncRNA predictions, and minor inaccuracies, gaps and small indels in the genomic alignments, we only infer orthology among similar ncRNA sequences across the genome alignment where the alignment falls within a range of ± 80 nucleotides for snoRNAs and ± 40 nucleotides for (the generally shorter) miRNAs across the entire alignment. In both cases, we stayed under the total length of individual genes to avoid unwanted overlap with adjacent paralogues within an RNA cluster. Although these range constraints may result in the loss of data (i.e., false negatives), larger ranges may result in inclusion of false positives in our data set; the latter is of greater concern than the former. Manual vetting of the data indicate that for most cases of positional conservation across the genome alignment, the range is considerably smaller (<5 nt).

Analysis of ncRNA Conservation across the Mammalian Tree

Relationships between the 11 mammals used in our analysis were established from a recently published mammalian supertree (Bininda-Emonds et al. 2007); chicken was added manually as an out-group by assuming a divergence time of 310 Myr (Hedges 2002). We inferred the ncRNA status of each internal node in the tree with maximum parsimony using DolloP from the PHYLIP package (Felsenstein 2004), using the subset of ncRNAs where positional conservation could be established between at least two genomes (table 1). Maximum likelihood was

Table 1
Number of Annotated ncRNAs per Genome Versus Aligned Regions[a]

	Total (Mb)	snoRNAs	miRNAs	Alignment (bp)	snoRNAs	miRNAs	Genome%	snoRNA%	miRNA%
Bos taurus	2918	586 (267)	565 (185)	1160	348 (232)	288 (154)	39.75	59.39	50.97
Canis familiaris	2385	490 (225)	628 (249)	1232	280 (194)	366 (200)	51.65	57.14	58.28
Equus caballus	2429	406 (182)	612 (221)	1219	221 (166)	348 (182)	50.18	54.43	56.86
Gallus gallus	1051	148 (118)	560 (256)	498	106 (94)	287 (183)	47.38	71.62	51.25
Homo sapiens	3253	717 (360)	1664 (740)	1834	450 (306)	1046 (570)	56.37	62.76	62.86
Macaca mulatta	3094	715 (280)	1208 (422)	1589	383 (240)	610 (338)	51.36	53.57	50.50
Monodelphis domestica	3502	221 (137)	375 (121)	1605	166 (124)	168 (88)	45.85	44.27	44.80
Mus musculus	3421	949 (351)	1081 (504)	1427	475 (289)	677 (421)	41.72	50.05	62.63
Ornithorhynchus anatinus	1918	2342 (259)	605 (123)	554	276 (177)	154 (97)	28.88	11.78	25.45
Pan troglodytes	2929	716 (333)	1464 (550)	1836	432 (282)	889 (440)	62.69	60.34	60.72
Pongo pygmaeus	3109	763 (280)	1485 (451)	1691	389 (228)	763 (347)	54.4	50.98	51.38
Rattus norvegicus	2507	1023 (373)	760 (290)	1376	511 (319)	446 (250)	54.88	49.95	58.68

[a] The 12 genome alignment for our study was obtained from the Ensembl database. The average proportion of each genome sequence present in synteny blocks is approximately 50% (genome% column). Approximately 50% of annotated snoRNAs (snoRNA% column) and miRNAs (miRNA% column) were included in synteny blocks and thus used in this study. Numbers in parentheses indicate intronically encoded RNAs.

not used owing to the absence of an accurate evolutionary model to statistically describe the gain and loss of ncRNAs.

Analysis of Host Gene Function

The analysis of host gene function was based upon GO terms from the Gene Ontology project (Ashburner et al. 2000). Given that many GO terms are assigned on the basis of sequence similarity to experimentally characterized homologs, we restricted our analysis to genes from *H. sapiens* as one of the better studied genomes. The graphical representation (fig. 4) was created from data from the Gene Ontology Term Mapper (http://go.princeton.edu). Statistical support was computed using the GoStat web server (Beissbarth and Speed 2004), employing a stringent cutoff of $P \leq 0.001$ and Benjamini correction for false positive detection. Expression data for a statistical comparison of Shannon entropy and strength of expression (approximated as the sum across all tested tissues) were obtained from the human transcriptome atlas (Su et al. 2004). Data were obtained from Array Express, accession E-TABM-145. To estimate the expression level of each gene, we calculated the median array signal for all tissues (removing duplicates, such as brain subsamples). To estimate the expression breadth, we calculated the Shannon entropy as $S = -\text{sum}(Pi \times \ln(Pi))$. Where the total expression T is the sum of all expression values for tissues (1..i), Ei is the expression of the gene in tissue i and the proportion of expression in tissue(i) is $Pi = Ei/T$.

Results and Discussion

Extensive Conservation of snoRNAs and miRNAs across the Mammalian Tree

We made use of available whole-genome alignments across 12 vertebrates (11 mammals plus chicken; Hubbard et al. 2009) and evolutionary conservation of annotated snoRNAs and miRNAs from Rfam (Griffiths-Jones et al. 2003) and miRBase (Griffiths-Jones 2006) to examine positional conservation of orthologous ncRNAs across the mammalian tree. Only snoRNAs and miRNAs located in syntenic regions were considered, yielding a set of 3041 unique miRNA and 1648 snoRNA groups in the 6.5 gigabase pair long alignment. Out of these, 648 snoRNAs and 964 miRNAs were present in more than one genome and formed the basis for our analysis (table 1). Given genome alignments and the evolutionary relationships between mammalian groups, we performed a parsimony-based analysis of conservation of miRNAs and snoRNAs using DolloP from the PHYLIP package to establish the ncRNA content at different stages during mammalian evolution (as represented by internal nodes in the tree, see fig. 1).

Our results indicate that a considerable number of snoRNAs and miRNAs can be traced back to the mammalian ancestor on the basis of genome alignment aided orthology assignment, 135 snoRNAs (135/648 = 21%) and 103 miRNAs (103/964 = 11%; see Node 2, fig. 1). We refer to these as ancestral positionally conserved (APC) RNAs, indicating that we can be confident of an ancestral conserved location for these ncRNAs. Other snoRNAs and miRNAs are present in a more limited number of nodes. Because this analysis cannot distinguish between a genuine de novo origin of a particular ncRNA within a particular lineage and an earlier origin with mobility or loss in deeper branching lineages, we collectively refer to these as novel location (NL) RNAs.

Clearly, there will be false discoveries and false negatives with automated ncRNA predictions (Griffiths-Jones 2007), and this may impact our results. Likewise, assembly errors in individual genomes are likewise a potential source of either missing or duplicated data, though overall these problems are likely to have a smaller overall impact than ncRNA annotation. The risk of including false positive ncRNA annotations will be higher for NL ncRNAs because inferences rely upon sequence data from only a few species. For deeper divergences, false positives become less likely because sequence conservation and consistent spurious ncRNA prediction is less likely. However, the greater sequence divergence between ancient ncRNAs may mean that the initial Blast-based screens fail to identify a putative ncRNA in the fist place (see Materials and Methods). Therefore, we probably "underestimate" the true number of APC ncRNAs and "overestimate" the true number of NL ncRNAs. The result is that our predictions for the percentages of APC ncRNAs (21%

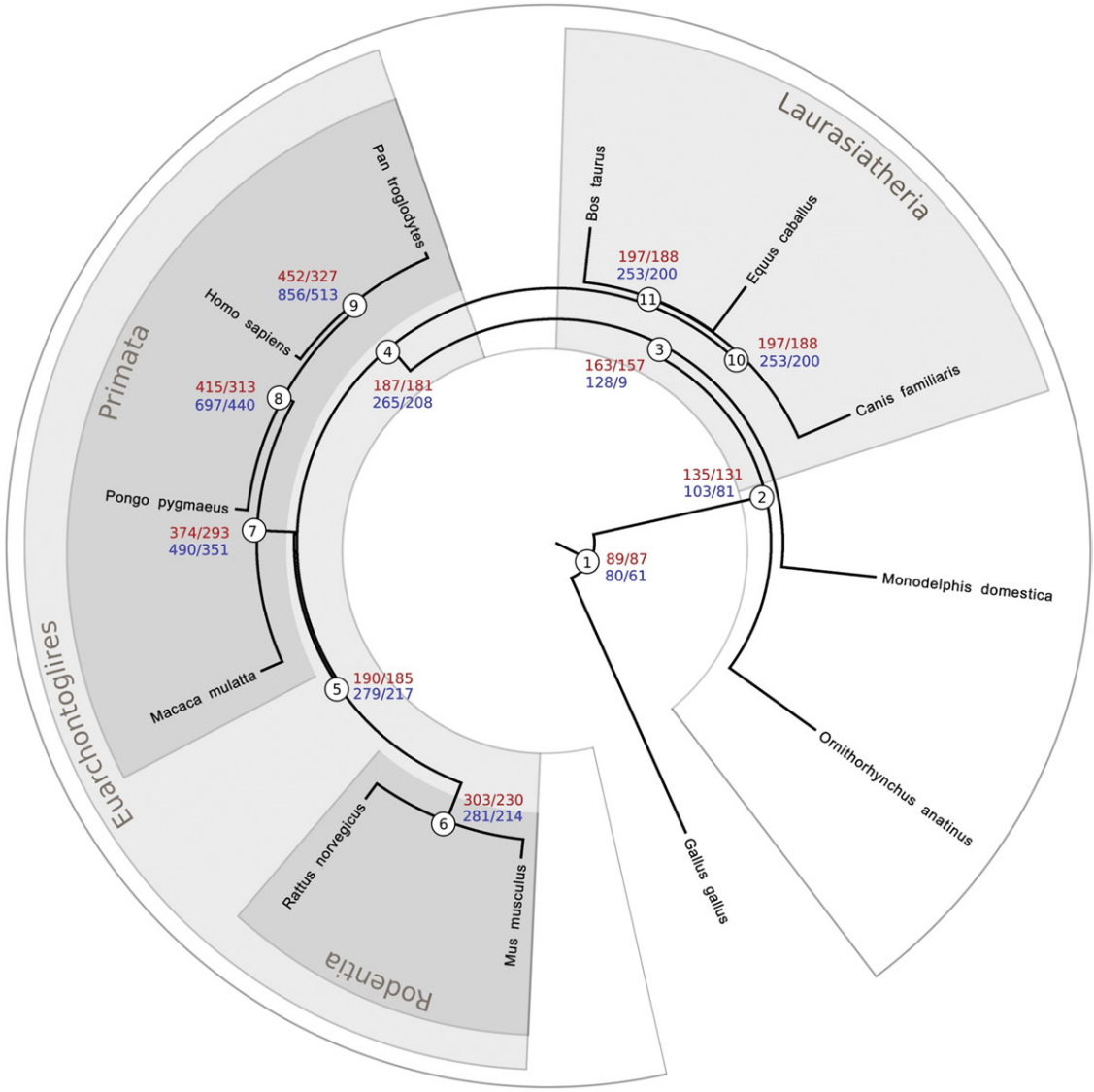

FIG. 1.—Reconstruction of ancestral states for positionally conserved snoRNAs and miRNAs across 11 mammalian genomes. The tree is based on the phylogeny reported by Bininda-Emonds et al. (2007), with modifications as described in Materials and Methods. Counts of positionally conserved ncRNAs are derived from a maximum parsimony analysis using DolloP from the PHYLIP package (see Materials and Methods). The numbers of ncRNAs inferred from synteny to be present at each internal node are listed (red: snoRNAs, blue: miRNAs). The first number indicates the total count of orthologous mi/snoRNAs inferred to be present at a given node, followed by the number of intronic ncRNAs inferred to be present at a given node (a subset of the first value).

of snoRNAs and 11% of miRNAs) are expected to be conservative.

Our analysis includes only the ncRNAs in aligned regions of the genomes, which were predicted in at least two species. Consequently, our data set includes only approximately 50% of all the annotated ncRNAs for these genomes (table 1). To examine how representative our analysis is of ncRNA gene paralogs, we reconstructed the ancestral states for individual snoRNA and miRNA families (as defined by Rfam and miRBase) based on their presence or absence in individual genomes (without reference to aligned regions and not taking into account copy numbers or location). We then compared these numbers with those obtained from our data set. The results (fig. 2) indicate that our analysis has good coverage of ncRNA families: approximately 80% of snoRNA families and 70% of miRNA families conserved across the entire 12 genome data set are included (node 2, fig. 2). This suggests that the remaining 20–30% are either mobile or located in regions too divergent to be alignable across larger evolutionary distances.

To confirm that our APC ncRNAs are more conserved than NL ncRNAs, we calculated percentage identity and median genomic evolutionary rate profiling (GERP) scores (GERP method, Cooper et al. 2005) from the genomic alignment. APC RNAs showed significantly greater percentage identity (Mann–Whitney U test $P = 2.2 \times 10e^{-16}$) and significantly higher median GERP scores (Mann–Whitney U test $P = 3.486 \times 10e^{-14}$) than NL RNAs inferred to have emerged along the branches leading to primates.

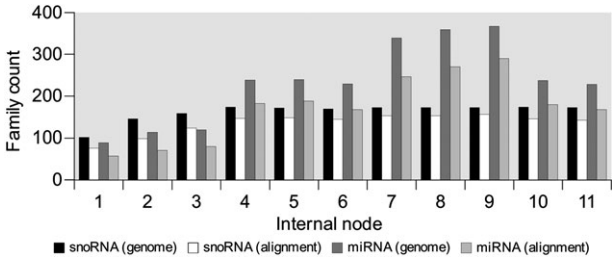

FIG. 2.—Representation of snoRNA and miRNA families among the subset of positionally conserved ncRNAs in this study. Our analysis is based on genomic alignments and thus excludes approximately 50% of annotated snoRNAs and miRNAs located in nonsyntenic regions of the respective genomes (table 1). To estimate the coverage of snoRNA and miRNA families (as defined by Rfam and miRBase) in our analysis, we performed a per-node reconstruction of family presence/absence irrespective of copy number or positional conservation (genome) and compared these numbers (family count) with the family representation in our analysis (alignment). The results indicate that our study set provides good coverage (between 70% and 90%) of snoRNA and miRNA families. The remaining ncRNA families are likely located in regions not alignable across genomes.

We observe that the vast majority of APC ncRNAs are intronic in all mammalian lineages represented in our study; 97% (131/135) of snoRNAs and 79% (81/103) of miRNAs traceable to the mammalian ancestor are intronic. This trend extends to the common ancestor of birds and mammals (Node 1, fig. 1). Thus, many ncRNAs appear to be stably associated with the same intron of the same host gene over considerable evolutionary timescales, possibly indicating a selective advantage for this arrangement over an intergenic location.

To examine patterns of intronic and intergenic ncRNA conservation across mammalian evolution, we compared the ncRNA inventory of the mammalian ancestor (node 2, fig. 1) with numbers obtained for the primate ancestor. We used the ncRNAs from the primate ancestor (node 4, fig. 1), rather than a data from specific species because elements present across several species will have a lower false positive rate. We find that 104 of the 374 primate snoRNAs (28%) are also present in the mammalian ancestor (table 2). Similarly, 92 of the 490 primate miRNAs (19%) are also present in the mammalian ancestor. The majority of these are intronic; 102 snoRNAs (102/104—98%) and 72 miRNAs (72/92—78%). Intronic snoRNAs ($\chi^2 = 22.33$; $P << 0.001$) are thus significantly more positionally stable than intergenic elements, whereas no such trend was found for miRNAs ($\chi^2 = 1.71$; $P = 0.19$).

The majority of ncRNAs used in our analysis (table 1) are specific to a particular mammalian order. Numbers in Laurasiatherians (horse, cow, and dog) are likely low on account of limited experimental study of ncRNAs among the Laurasiatherian genomes included in this study. The intensive experimental focus on human ncRNA (e.g., Fejes-Toth et al. 2009), particularly for miRNA identification (e.g., Bar et al. 2008; Wyman et al. 2009), is likely to be responsible for inflation of the numbers of annotated ncRNAs among primates. Given strong miRBase growth (Griffiths-Jones et al. 2008; see supplementary fig. S1, Supplementary Material online), it would be premature to conclude that the observed higher number of miRNAs in primates (fig. 1 and fig. 2) is due to a corresponding jump in miRNA disparity (*sensu* Heimberg et al. 2008) within this group; in the current analysis, we cannot exclude the possibility that this is an artifact of greater experimental focus on miRNAs in *H. sapiens* (supplementary fig. S1, Supplementary Material online). Analysis of reported expression profiles of miRNA host genes (Su et al. 2004) failed to detect correlation with a particular tissue (data not shown); a significant correlation might have been expected if newly emerging miRNAs were predominantly involved in, for example, brain development. This should not be taken as evidence against a general correlation between the evolution of the human brain and miRNA genesis—our conclusion is limited to intronic miRNAs present in the primate ancestor.

Table 2
Patterns of Intronic and Intergenic ncRNA Conservation

ncRNA type	Present in	Intronic	Intergenic	Total
snoRNA	All primates	293	81	374
	Mammalian ancestor[a]	102	2	104
miRNA	All primates	351	139	490
	Mammalian ancestor[a]	72	20	92

[a] ncRNAs conserved across primates (node 7, fig. 1) that were already present in the mammalian ancestor (node 2, fig. 1).

Rfam and miRBase Families in the Common Ancestor of Mammals and Birds Are Represented by both Orthologues and Paralogues

Both snoRNAs and miRNAs are grouped into families by Rfam and miRBase, respectively, on the basis of sequence similarity. Using this information, we sought to extend our analysis to include the presence or absence as well as secondary losses of such families. Amongst primate NL ncRNAs, 130 out of 189 snoRNAs (68.78%) and 80 out of 220 miRNAs (36.46%) belong to families already present in the mammalian ancestor (supplementary tables S1 and S2, Supplementary Material online). The positionally conserved ncRNA content of the common ancestor of birds and mammals consists of both single-family representatives and cases of paralogy for both snoRNAs and miRNAs (supplementary tables S3 and S4, Supplementary Material online; this of course excludes those ncRNAs which are not positionally conserved). Thus, mobility is clearly a feature of numerous ncRNAs.

Interestingly, this also includes cases of evolutionarily conserved within-gene duplication. The most striking examples are miRNA miR-302 and box C/D snoRNA snoRD58 (of which there are four copies each; supplementary tables S3 and S4, Supplementary Material online).

MiR-302 has diversified into four distinct RNA species (miR-302a–d) as a result of "within-intron" duplication within the LARP7 gene prior to the bird–mammal split. A related fifth miRNA, miR-367 (miRBase accession: MI0000738), also conserved in this cluster (supplementary fig. S3, Supplementary Material online). Homologues of miR-302 are known in *Xenopus* (miR-427) and zebra fish (miR-430), and recent experimental data demonstrate that human miR-302a and *Xenopus* miR-427 are involved in

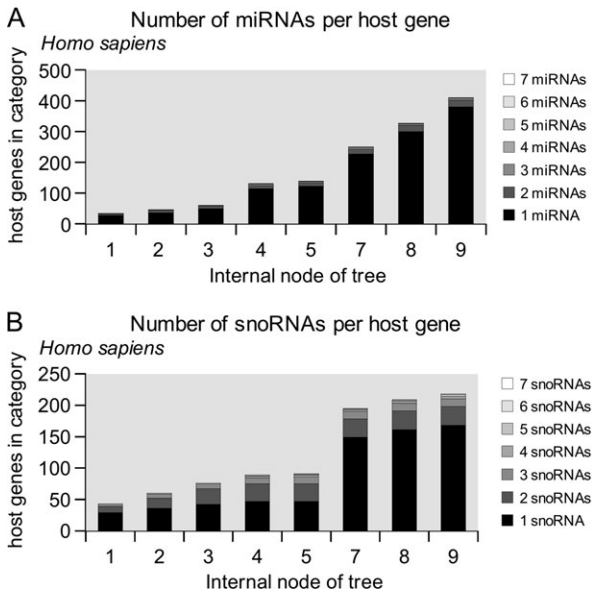

FIG. 3.—Few genes carry multiple intronic ncRNAs. A per-node (fig. 1) inspection of the number of (*A*) miRNAs and (*B*) snoRNAs per host gene in *Homo sapiens* reveals that most host genes carry only a single ncRNA (black). This suggests de novo emergence and/or transduplications (including ncRNA retroposition) of existing families are more prevalent in mammals than cis-duplication.

embryonic mesendoderm differentiation in both species through regulation of Nodal signaling (Choi et al. 2007; Rosa et al. 2009). It is unclear exactly what role the four mammalian miR-302 paralogues may have, but this broad vertebrate family appears to play numerous roles in addition to the above partially conserved functional roles (Ketting 2009). The functional significance of the association between LARP7 and the miR-302 cluster is as yet unclear; LARP7 is involved in negative regulation of RNA polymerase II genes via 7SK RNP, of which it is a constituent (He et al. 2008; Markert et al. 2008). However, we note that all miRNAs in this "intronic" cluster are coded antisense to LARP7 (supplementary fig. S3, Supplementary Material online), and, consequently, it is unclear to what extent there is overlap of expression profiles.

In the case of snoRD58 cis-duplicates, all four are found in "different" introns of the gene coding for the ribosomal protein RPL17 (supplementary fig. S2, Supplementary Material online). Two of these have been previously shown to direct 2′-*O*-methylation of 28S rRNA (snoRD58a and b; Nicoloso et al. 1996) and snoRD58c has been predicted to modify this same rRNA molecule (Yang et al. 2006). The role of snoRD58d has not been established, but its conservation across mammals and birds suggests it is not a degenerate nonfunctional copy, as has been suggested (http://www-snorna.biotoul.fr/plus.php?id=U58C; Lestrade and Weber 2006).

Such cis-duplications do not appear to be widespread across our data set; most intronic ncRNA-bearing genes in humans carry only a single snoRNA or miRNA (fig. 3). It is likely that the three processes of cis- and transduplication and de novo emergence all contribute to ncRNA evolution (Weber 2006; Zemann et al. 2006; Lu et al. 2008; Schmitz et al. 2008), however, our analysis suggests that the latter two processes may play a greater role in the evolution of snoRNA and miRNA genes in mammals.

Analysis of Host Gene Functions Suggests Recent Diversification of snoRNA Functions during Primate Evolution

Previous reports suggest that many of the more widely conserved snoRNAs are involved in rRNA processing (Lafontaine and Tollervey 1998; Dieci et al. 2009). Because transcriptional overlap may well be common between intronic ncRNAs and their host genes (Baskerville and Bartel 2005), we examined the difference between host gene function in our set of APC snoRNAs (mammalian ancestor) versus snoRNA groups of a putatively more recent origin (NL). To describe host gene function, we used Gene Ontology (Ashburner et al. 2000), expression level, and gene expression breadth (amongst tissues) data derived from human host genes. We reasoned that if snoRNA–host gene relationships are evolutionarily stable, host gene function and tissue distribution may provide information regarding emergent roles among intronic snoRNAs, as has been considered for miRNAs (Rodriguez et al. 2004).

Of 60 human host genes dating back to the mammalian ancestor (i.e., hosting an APC snoRNA), 21 associated with the biological process "translation" and 18 with the molecular function "RNA binding." In contrast, of the 123 host genes recruited along the branches leading to primates (i.e., hosting exclusively NL snoRNAs), only 14 associate with translation and 15 with RNA binding (fig. 4). This provides strong support ($P = 9.14 \times 10e^{-22}$, see Beissbarth and Speed 2004) for overrepresentation of human host genes involved in ribosome function traceable back to the mammalian ancestor compared with those specific to primates.

We also expected that host genes of APC snoRNAs would be expressed in a wider range of tissues than NL snoRNAs host genes, consistent with a role in more fundamental cellular processes. To test this expectation, we use the GNF/Novartis human gene expression data set, containing expression profiles for 33698 genes in 38 tissues (Su et al. 2004). As a measure of the breadth of host gene expression, we calculated the Shannon entropy for each gene expression profile. Briefly, Shannon entropy measures the degree to which a quantity is "randomly" distributed amongst categories (tissues in our case). A high entropy indicates ubiquitous expression, whereas low entropy indicates expression limited to one or a few tissues. A comparison of Shannon entropies for those host genes where expression data were available (see Materials and Methods) revealed significantly higher entropy for APC snoRNAs ($P = 4.18 \times 10e^{-6}$, Mann–Whitney *U* test) indicative of broad expression. We also considered whether APC host genes were more highly expressed than NL host genes. The median of expression levels (measured across all tissues) of genes containing an APC snoRNA were significantly higher than host genes containing NL snoRNAs ($P = 2.621e^{-7}$, Mann–Whitney *U* test). These observations indicate, in the human snoRNA data set, NL snoRNAs reside in the introns of more tissue-specific low-expression genes.

The dependence of snoRNA and host gene expression cannot be assumed if ncRNA and host gene are encoded

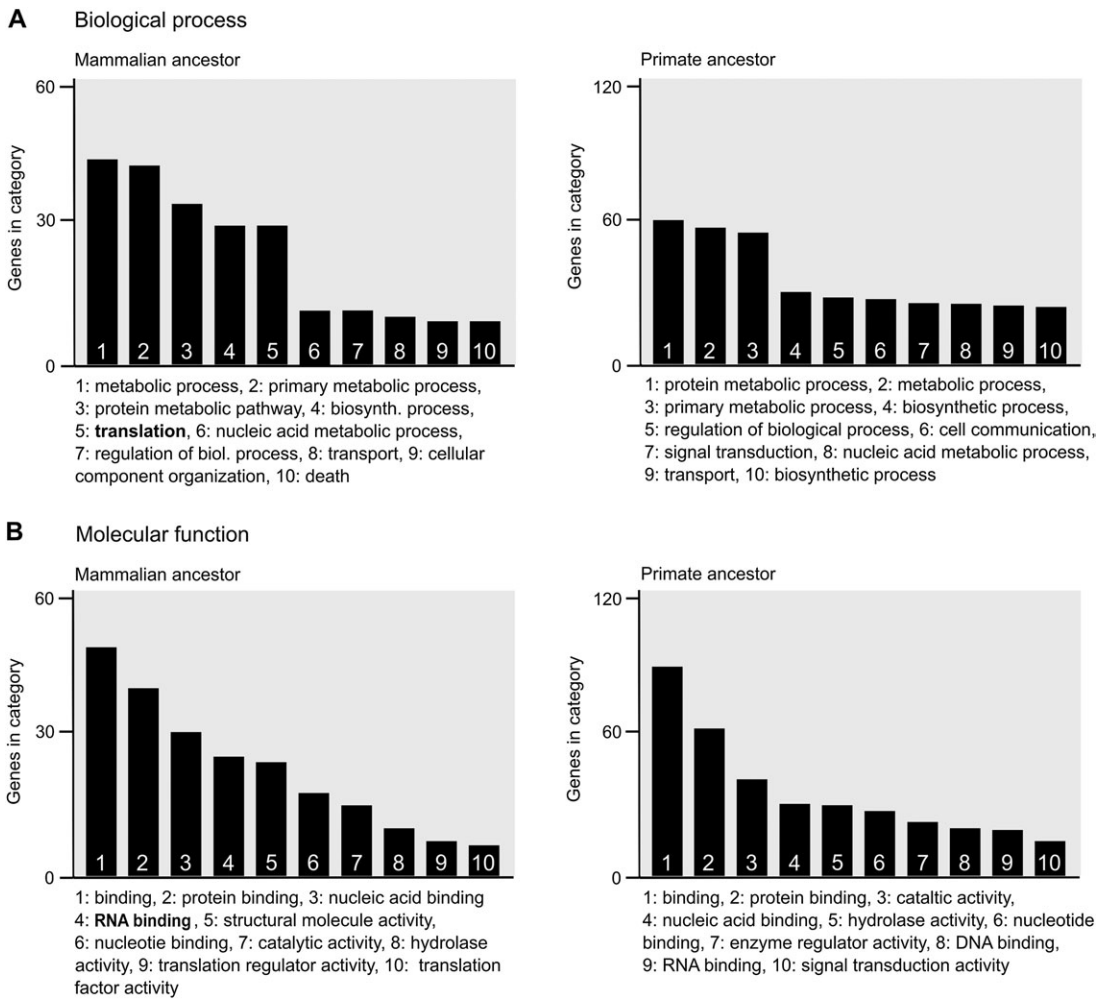

FIG. 4.—A comparison of human host gene function between the ancestor of mammals and primates suggests a diversification in the roles of snoRNA host genes in more recent evolutionary history. Whereas half of the host genes in the earliest mammals (mammalian ancestor, graphs on left) are involved in ribosome formation or protein production ([*A*] biological process: translation; [*B*] molecular function: RNA binding), no such bias can be found for snoRNA host genes traceable to the primate ancestor (minus those also in the mammalian ancestor, graphs on right). The scale on the *y* axis corresponds to the total number of genes from *Homo sapiens* used in each analysis. Only the top 10 categories are shown. *E* values for significantly overrepresented GO terms were calculated using GoStat (Beissbarth and Speed 2004).

on opposite strands. Only a fraction of intronic NL snoRNAs fall into this category (approximately, 12% in the primate ancestor), whereas all deeply conserved intronic snoRNAs (mammalian ancestor) are on the sense strand (supplementary table S5, Supplementary Material online). This finding therefore strongly supports the notion of overlapping expression profiles.

We performed equivalent analyses for human miRNA host genes. We found no functional association for host genes of miRNAs, regardless of node depth (data not shown) nor an significant differences between the expression level or breadth of APC and NL miRNA host genes. There is no a priori expectation that this class of regulatory ncRNA should be associated with regulation of a specific process, and our result likely reflects the broad range of cellular processes in which miRNAs are involved (and the large number of potential target mRNAs). We also note that intronic miRNAs are more frequently housed antisense to the host gene (up to 30% per node). This may indicate that expression of a significant fraction of intronic miRNAs is not directly dependent on host gene expression.

Conclusions

We have analyzed the positional conservation of snoRNAs and miRNAs across a multiple genome alignment of 11 mammals, using the chicken genome as an out-group. We found 3041 miRNAs and 1649 snoRNAs to be present in two or more species. Of these, 169 are APC ncRNAs (89 snoRNAs and 80 miRNAs), and the vast majority (98% of snoRNAs and 76% of miRNAs) are located in the introns of protein-coding genes. Intronic snoRNAs and miRNAs are significantly more likely to be positionally stable than intergenic RNAs.

Our results thus demonstrate the utility of genome alignments for examining ncRNA orthology across considerable evolutionary timescales and complement sequence similarity guided approaches. Comparative genome analyses of ncRNAs are still in their infancy, necessitating

a conservative approach, but as ncRNA annotations and genome assemblies improve, additional questions will become tractable. The current analysis does not enable us to establish whether intronic APC ncRNAs are ancestrally intronic or whether they have migrated from other genomic locations. Among ncRNAs showing positional conservation among primates (270 snoRNAs and 398 miRNAs), some may be new RNAs that have arisen de novo in the lineage leading to primates. However, family assignments based on Rfam and miRBase classifications also indicate that numerous primate NL ncRNAs (130 snoRNAs and 82 miRNAs, supplementary tables S1 and S2, Supplementary Material online) are paralogs of families dating back to the mammalian ancestor, suggesting that ncRNAs positionally conserved among primates are likely to have inserted into their current location from elsewhere.

This indicates that only a minority of snoRNAs and miRNAs—primarily intron-encoded ncRNAs—have remained in the same location during mammalian evolution. The general patterns we observe (greater positional conservation for intronic ncRNAs, with few such locations being demonstrably ancestral) suggest that intronic location may confer an advantage but that ncRNAs only rarely arise de novo within introns.

Finally, we report that intronic APC snoRNAs are more likely to be present in the introns of genes involved in ribosome biogenesis and more likely to be broadly and highly expressed than genes containing a NL snoRNA (NL snoRNA). SnoRNAs function in ribosome biogenesis across all eukaryotes and are known to be encoded in the introns of ribosomal protein genes in species as evolutionarily distant as yeast (Bachellerie et al. 2002), and it will therefore be of interest to establish whether intronic APC snoRNAs have been ancestrally associated with these host genes or whether various intronic locations are have arisen by convergent evolution. In contrast to APC snoRNAs, miRNA host genes show no significant associations with specific biological processes or functions, and we detect no expression differences between ancestral and NL miRNAs, as measured by expression breadth across tissues or levels of expression. Interestingly, examination of host genes for NL snoRNAs reveals a pattern similar to that observed for miRNAs, suggesting that snoRNAs may have been coopted into a broader range of (possibly regulatory) roles in the course the diversification of mammals. This suggests that during the course of mammalian evolution, snoRNAs have undergone gradual diversification from their ancestral functions in translation, which may date to early stages in cellular evolution (Omer et al. 2000; Penny et al. 2009).

Supplementary Material

Supplementary figures S1–S3 and tables S1–S5 are available at *Genome Biology and Evolution* online (http://www.oxfordjournals.org/our_journals/gbe/).

Acknowledgments

A.M.P. is a Royal Swedish Academy of Sciences Research Fellow supported by a grant from the Knut and Alice Wallenberg Foundation. M.P.H. acknowledges support from the Astrobiology Graduate School at Stockholm University.

Literature Cited

Ashburner M, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 25:25–29.
Bachellerie JP, Cavaillé J, Hüttenhofer A. 2002. The expanding snoRNA world. Biochimie. 84:775–790.
Bar M, et al. 2008. MicroRNA discovery and profiling in human embryonic stem cells by deep sequencing of small RNA libraries. Stem Cells. 26:2496–2505.
Bartel DP. 2009. MicroRNAs: target recognition and regulatory functions. Cell. 136:215–233.
Baskerville S, Bartel DP. 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA. 11:241–247.
Beissbarth T, Speed TP. 2004. GOstat: find statistically overrepresented Gene Ontologies within a group of genes. Bioinformatics. 20:1464–1465.
Bininda-Emonds OR, et al. 2007. The delayed rise of present-day mammals. Nature. 446:507–512.
Brown JW, Marshall DF, Echeverria M. 2008. Intronic noncoding RNAs and splicing. Trends Plant Sci. 13:335–342.
Carthew RW, Sontheimer EJ. 2009. Origins and mechanisms of miRNAs and siRNAs. Cell. 136:642–655.
Cervelli M, et al. 2002. Comparative structure analysis of vertebrate U17 small nucleolar RNA (snoRNA). J Mol Evol. 54:166–179.
Choi WY, Giraldez AJ, Schier AF. 2007. Target protectors reveal dampening and balancing of Nodal agonist and antagonist by miR-430. Science. 318:271–274.
Cooper GM, et al. 2005. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 15:901–913.
Darzacq X, et al. 2002. Cajal body-specific small nuclear RNAs: a novel class of 2′-O-methylation and pseudouridylation guide RNAs. Embo J. 21:2746–2756.
Dieci G, Preti M, Montanini B. 2009. Eukaryotic snoRNAs: a paradigm for gene expression flexibility. Genomics. 94:83–88.
Eddy SR. 2001. Non-coding RNA genes and the modern RNA world. Nat Rev Genet. 2:919–929.
Enerly E, Mikkelsen OL, Lyamouri M, Lambertsson A. 2003. Evolutionary profiling of the U49 snoRNA gene. Hereditas. 138:73–79.
Fejes-Toth K, et al. 2009. Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs. Nature. 457:1028–1032.
Felsenstein J. 2004. PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Seattle (WA): Department of Genome Sciences, University of Washington. Available from: http://evolution.genetics.washington.edu/phylip.html.
Ganot P, Jády BE, Bortolin ML, Darzacq X, Kiss T. 1999. Nucleolar factors direct the 2′-O-ribose methylation and pseudouridylation of U6 spliceosomal RNA. Mol Cell Biol. 19:6906–6917.
Griffiths-Jones S. 2006. miRBase: the microRNA sequence database. Methods Mol Biol. 342:129–138.
Griffiths-Jones S. 2007. Annotating noncoding RNA genes. Annu Rev Genomics Hum Genet. 8:279–298.
Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. 2003. Rfam: an RNA family database. Nucleic Acids Res. 31:439–441.
Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. 2008. miRBase: tools for microRNA genomics. Nucleic Acids Res. 36:D154–D158.

He N, et al. 2008. A La-related protein modulates 7SK snRNP integrity to suppress P-TEFb-dependent transcriptional elongation and tumorigenesis. Mol Cell. 29:588–599.
Hedges SB. 2002. The origin and evolution of model organisms. Nat Rev Genet. 3:838–849.
Heimberg AM, Sempere LF, Moy VN, Donoghue PC, Peterson KJ. 2008. MicroRNAs and the advent of vertebrate morphological complexity. Proc Natl Acad Sci USA. 105:2946–2950.
Hirose T, Shu MD, Steitz JA. 2003. Splicing-dependent and -independent modes of assembly for intron-encoded box C/D snoRNPs in mammalian cells. Mol Cell. 12:113–123.
Hofacker IL, et al. 1994. Fast folding and comparison of RNA secondary structures. Monatschefte für Chemie. 125:167–188.
Hubbard TJ, et al. 2009. Ensembl 2009. Nucleic Acids Res. 37:D690–D697.
Jády BE, Kiss T. 2001. A small nucleolar guide RNA functions both in 2′-O-ribose methylation and pseudouridylation of the U5 spliceosomal RNA. Embo J. 20:541–551.
Ketting RF. 2009. Semiconserved regulation of mesendoderm differentiation by microRNAs. Dev Cell. 16:487–488.
Kishore S, Stamm S. 2006. The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. Science. 311:230–232.
Kiss T. 2002. Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. Cell. 109:145–148.
Lafontaine DL, Tollervey D. 1998. Birth of the snoRNPs: the evolution of the modification-guide snoRNAs. Trends Biochem Sci. 23:383–388.
Lestrade L, Weber MJ. 2006. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. Nucleic Acids Res. 34:D158–D162.
Lu J, et al. 2008. The birth and death of microRNA genes in Drosophila. Nat Genet. 40:351–355.
Markert A, et al. 2008. The La-related protein LARP7 is a component of the 7SK ribonucleoprotein and affects transcription of cellular and viral polymerase II genes. EMBO Rep. 9:569–575.
Mattick JS. 2001. Non-coding RNAs: the architects of eukaryotic complexity. EMBO Rep. 2:986–991.
Mattick JS. 2003. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. Bioessays. 25:930–939.
Mattick JS. 2009. The genetic signatures of noncoding RNAs. PLoS Genet. 5:e1000459.
Nahkuri S, Taft RJ, Korbie DJ, Mattick JS. 2008. Molecular evolution of the HBII-52 snoRNA cluster. J Mol Biol. 381:810–815.
Nawrocki EP, Kolbe DL, Eddy SR. 2009. Infernal 1.0: inference of RNA alignments. Bioinformatics. 25:1335–1337.
Nicoloso M, Qu LH, Michot B, Bachellerie JP. 1996. Intron-encoded, antisense small nucleolar RNAs: the characterization of nine novel species points to their direct role as guides for the 2′-O-ribose methylation of rRNAs. J Mol Biol. 260:178–195.
Omer AD, et al. 2000. Homologs of small nucleolar RNAs in Archaea. Science. 288:517–522.
Penny D, Hoeppner MP, Poole AM, Jeffares DC. Forthcoming 2009. An overview of the introns-first theory. J Mol Evol. doi: 10.1007/s00239-009-9279-5.
Rodriguez A, Griffiths-Jones S, Ashurst JL, Bradley A. 2004. Identification of mammalian microRNA host genes and transcription units. Genome Res. 14:1902–1910.
Rosa A, Spagnoli FM, Brivanlou AH. 2009. The miR-430/427/302 family controls mesendodermal fate specification via species-specific target selection. Dev Cell. 16:517–527.
Royo H, Cavaillé J. 2008. Non-coding RNAs in imprinted gene clusters. Biol Cell. 100:149–166.
Schmitz J, et al. 2008. Retroposed SNOfall—a mammalian-wide comparison of platypus snoRNAs. Genome Res. 18:1005–1010.
Stefani G, Slack FJ. 2008. Small non-coding RNAs in animal development. Nat Rev Mol Cell Biol. 9:219–230.
Su AI, et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci USA. 101:6062–6067.
Tanaka-Fujita R, Soeno Y, Satoh H, Nakamura Y, Mori S. 2007. Human and mouse protein-noncoding snoRNA host genes with dissimilar nucleotide sequences show chromosomal synteny. RNA. 13:811–816.
Tycowski KT, Shu MD, Steitz JA. 1996. A mammalian gene with introns instead of exons generating stable RNA products. Nature. 379:464–466.
Volff JN, Brosius J. 2007. Modern genomes with retro-look: retrotransposed elements, retroposition and the origin of new genes. Genome Dyn. 3:175–190.
Weber MJ. 2006. Mammalian small nucleolar RNAs are mobile genetic elements. PLoS Genet. 2:e205.
Weinstein LB, Steitz JA. 1999. Guided tours: from precursor snoRNA to functional snoRNP. Curr Opin Cell Biol. 11:378–384.
Wyman SK, et al. 2009. Repertoire of microRNAs in epithelial ovarian cancer as determined by next generation sequencing of small RNA cDNA libraries. PLoS One. 4:e5311.
Yang JH, et al. 2006. snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. Nucleic Acids Res. 34:5112–5123.
Zemann A, op de Bekke A, Kiefmann M, Brosius J, Schmitz J. 2006. Evolution of small nucleolar RNAs in nematodes. Nucleic Acids Res. 34:2676–2685.

Gertraud Burger, Associate Editor

Accepted October 31, 2009