[Frontiers in Bioscience E5, 814-822, June 1, 2013]

Massive microRNA sequence conservation and prevalence in human and chimpanzee introns

Aubrey E. Hill[1], Eric J. Sorscher[2,3]

[1]Department of Computer Science, University of Alabama at Birmingham,1530 3rd Avenue South, MCLM 796, Birmingham, AL 35294-0005, USA [2]Department of Medicine, University of Alabama at Birmingham,1530 3rd Avenue South, MCLM 796, Birmingham, AL 35294-0005, USA, Gregory Fleming James Cystic Fibrosis Center, University of Alabama at Birmingham, 1530 3rd Avenue South, MCLM 796, Birmingham, AL 35294-0005, USA

TABLE OF CONTENTS

1. ABSTRACT

Human and chimpanzee introns contain numerous sequences strongly related to known microRNA hairpin structures. The relative frequency is precisely maintained across all chromosomes, suggesting the possible co-evolution of gene networks dependent upon microRNA regulation and with origins corresponding to the advent of primate transposable elements (TEs). While the motifs are known to be derived from transposable elements, the most common are far more numerous than expected from the number of TEs and their paralogous sequences, and exhibit striking conservation in comparison to the surrounding TE sequence context. Several of these motifs also exhibit structural complimentarity to each other, suggesting a pairing function at the level of DNA or RNA. These “pseudomicroRNAs,” in semblance to pseudogenes, include hundreds of thousands of vestigial paralogs of primate microRNAs, many of which may have functioned historically or remain active today.

2. INTRODUCTION

Mattick (*1-4*) and colleagues have postulated that introns constitute the basis for genetic regulatory networks. We previously demonstrated that complete intronic sequences transfected into human cells result in intron-specific gene expression patterns (*5*). However, expression of conventional microRNAs, small 22 nt sequences classically cleaved by Drosha from larger, hairpin-containing DNA motifs, could not account for these earlier observations. MicroRNA expression clearly contributes to gene network organization, but the extent and impact of microRNAs on mRNA expression, primate development and evolution remain to be determined. For example, environmental perturbations such as hypoxia dramatically alter genome-wide microRNA profiles, but the changes show limited concordance (and poorly predict) overall shifts in the mRNA transcriptome (*6*).

Rodriguez *et al* (*7*) noted that among a well described cohort of 232 mammalian microRNAs, more than half were found in introns. These same investigators observed that expression of many mammalian microRNAs may be directly linked to transcription of the surrounding DNA sequences (both protein-coding and noncoding RNAs). Others (*8*) have suggested that introns are the result of the propagation of replicative transposons or retrotransposons and that introns are, in fact, "broken" transposons. This latter assertion is compatible with the finding that 60% of transposable elements in both human and mouse are found in introns (*9*), and that a number of human microRNAs have originated from transposable elements (*10*). However, the bioinformatic connections between intronic DNA, the prevalence of microRNA hairpin structures in both human and nonhuman primates, and the relationships of these microRNA signatures with regard each other or to the transposable elements from which they derive, have not been assessed in detail.

Ruby *et al* (*11*) identified sequences from introns that resemble microRNA hairpins and showed that these "mirtrons" can sometimes be processed in the absence of Drosha, the enzyme that typically cleaves the microRNA hairpin at its base. In their model, splicing defines pre-microRNAs in a fashion similar to the classical means of cleaving microRNA hairpins from a larger RNA sequence. In the present work, we provideevidence for many hundreds of thousands of genomic sequences that may in fact be functioning (or have functioned evolutionarily) as microRNAs. The finding of a massive number of microRNA related hairpins ("pseudomicroRNAs") suggests a critical role during establishment of primate evolutionary diversity, and may represent the same sort of historical record attributable to pseudogenes. In this report, we describe a relative frequency profile of the 10-12 most common of these sequences in human and chimpanzee. This profile is essentially the same in both species and is strikingly maintained across all chromosomes irrespective of chromosomal size. The two most common pseudomicroRNAs have been maintained at an approximate 1:1 ratio despite a twenty-one-fold difference in the number of copies of transposable elements from which the sequences were derived.

3. MATERIALS AND METHODS

All intronic sequences from human and chimpanzee genomes were obtained from the UCSC Table Browser using hg18/NCBI Build 36.1 (*32)*. Intronic sequences were individually used as queries in a BLAST (version 2.2.16) (*33)* search against both the microRNA hairpins and the mature microRNAs from the miRBase database (*12*). We then developed Java servlets to convert results of the BLAST searches into a relational database using Microsoft Access. The resulting database was filtered by using a number of SQL queries. The results of the initial BLAST search were filtered to yield only those matches in which there were at least 18 identical nucleotides and a maximum E-value of 1 x 10^{-4}.

4. RESULTS

4.1. Genome-wide analysis of DNA hairpin sequences

The complete sets of known human and chimpanzee intronic sequences were used as queries against the Sanger microRNA hairpin database and the mature microRNA database (*12*). The result of the initial BLAST search for microRNA hairpins was filtered to yield only those matches in which there were at least 18 identical nucleotides and an E-value $\leq 1 \times 10^{-4}$. There were 698,069 hairpin hits that matched these criteria across the human genome and 762,787 across the chimpanzee genome. The distribution of hairpins by chromosome is depicted in Table 1 for human (and Table 2 for chimpanzee). Human and chimpanzee E-values ranged from 10^{-4} to 4 x 10^{-67} and 1 x 10^{-4} to 9 x 10^{-67}, respectively.

Table 3 depicts the percentages of significant BLAST hits against a subset of the most common microRNA hairpins for each human chromosome, and for the entire genome (Table 4 provides the corresponding data in chimpanzees). The relative proportion of hits for each of the hairpins is remarkably similar and strongly conserved across every chromosome. Notably, each of the microRNAs presented in Table 3 is listed as occurring only once in the Sanger hairpin database (i.e. at a single chromosomal location) when a standard microRNA search was conducted, despite the presence of hundreds of thousands of closely related sequences throughout both primate genomes. The authentic frequency of these related sequences, therefore, could be missed using standard methodologies to query genomic data repositories.

4.2. Genome-wide mature microRNA analysis

A BLAST search against the Sanger mature microRNA database yielded 1,661 exact microRNA matches (E-value <= 0.0001) in the human genome and 2,003 in the chimpanzee genome. This represents a more conventional accounting of microRNAs than the hairpin-based search shown above. None of the most numerous mature microRNAs corresponded to the frequent hairpin sequences described in Table 1, and are therefore likely to represent functionally and/or evolutionarily distinct entities.

4.3. Analysis of microRNA-related hairpins in the human and chimpanzee genomes

The most common microRNA hairpin database sequences resulting from BLAST hits against the Sanger database are shown in Figure 1. There was no strong preference for the plus or minus strand among the intronic sequences. A subset of the frequently occurring human intronic hairpin sequences (colored) versus complete microRNA hairpin sequences obtained from the Sanger database is depicted, with representative alignments of known microRNAs to numerous pseudomicroRNAs. The sequences corresponding to mature microRNAs are highlighted in yellow. Figure 2 represents a summary of multiple sequence alignments (*13*) of twenty-five hsa-mir-566 homologs with the most significant BLAST E-values

Table 1. Distribution of microRNA hairpin BLAST hits across human chromosomes

Chromosome	Mbp	% genome bp	No. genes	% genes	BLAST HITS	% BLAST HITS	OBS/EXP (% bp) BLAST HITS*	OBS/EXP (% genes) BLAST HITS*	No. unique microRNA hairpins†
1	247	8.02	3186	9.69	46511	6.66	0.83	0.69	150
2	243	7.89	2093	6.36	49282	7.06	0.89	1.11	106
3	200	6.5	1638	4.98	41555	5.95	0.92	1.2	153
4	191	6.2	1300	3.95	27192	3.9	0.63	0.99	117
5	181	5.88	1448	4.4	33556	4.81	0.82	1.09	140
6	171	5.55	1843	5.6	31602	4.53	0.82	0.81	62
7	159	5.16	1722	5.24	41466	5.94	1.15	1.13	144
8	146	4.74	1162	3.53	25699	3.68	0.78	1.04	47
9	140	4.55	1394	4.24	30173	4.32	0.95	1.02	303
10	135	4.38	1259	3.83	38813	5.56	1.27	1.45	89
11	134	4.35	2000	6.08	30904	4.43	1.02	0.73	82
12	132	4.29	1509	4.59	36777	5.27	1.23	1.15	76
13	114	3.7	611	1.86	12519	1.79	0.48	0.96	20
14	106	3.44	1420	4.32	25285	3.62	1.05	0.84	42
15	100	3.25	1143	3.48	26836	3.84	1.18	1.1	71
16	89	2.89	1270	3.86	28495	4.08	1.41	1.06	48
17	79	2.57	1650	5.02	42490	6.09	2.37	1.21	101
18	76	2.47	480	1.46	12950	1.86	0.75	1.27	84
19	64	2.08	1861	5.66	35475	5.08	2.44	0.9	138
20	62	2.01	824	2.51	20880	2.99	1.49	1.19	126
21	47	1.53	386	1.17	9657	1.38	0.9	1.18	93
22	50	1.62	812	2.47	18684	2.68	1.65	1.08	61
X	155	5.03	1529	4.65	29173	4.18	0.83	0.9	140
Y	58	1.88	344	1.05	2095	0.3	0.16	0.29	11
genome	3079		32884		698069				

*Frequency of BLAST hit observations normalized to the expected number if all BLAST hits were distributed evenly among all chromosomes or known genes. † Total number of unique microRNA related hairpins (with distinct mi croRNA identification (ID) numbers) matched by BLAST search

($\leq$1 x 10^{-30}), and indicates a much higher degree of conservation among hsa-mir-566 hairpin homologs versus either upstream (5') or downstream (3') DNA.

A high degree of similarity between the most common intronic hairpin sequences and corresponding, previously established microRNAs was observed. The two most prevalent, but previously unappreciated microRNA-related motifs represent over 280,000 occurrences of hsa-mir-566 or hsa-mir-619 paralogs in the human genome, and more than 290,000 in chimpanzee.

As an additional test of the prevalence and conservation of a specific microRNA hairpin, the first sequence of Figure 1 (sub-sequence of hsa-mir-566; occurring 19,521 times) was used as a query to retrieve common and related motifs. The fifteen most frequent among 10,861 resulting human sequences are shown in Table 5. All of these are essentially the same except for minor base substitutions. The most prevalent (which occurs 362 times) was aligned with hsa-mir-566 (Figure 3). The region includes a sub-sequence that differs from the mature hsa-mir-566 by only three bases. When the same procedure was used to retrieve common intronic sequences similar to hsa-mir-619, the fifteen most common (of 783 intronic sequences) again exhibited only minor base substitutions (not shown).

5. DISCUSSION

Although once viewed as irrelevant, intronic DNA is presently known to serve a number of crucial functions (*14-25*). Among these is a role for introns and their imbedded transposable elements as a repository for DNA variation (*7,9,10*). For example, the microRNA hsa-mir-566 is derived from Alu Sg and found on human chromosome 3. Hsa-mir-619 occurs on chromosome 12, and is derived from the more ancient Alu Sx and the LINE, L1MC4. Alus Sx and J, the earliest ancestors of all Alus, had their period of greatest amplification around 55 million years ago (mya) at which time approximately 850,000 thousand copies are thought to have appeared during the initial stages of the primate radiation. Approximately 40,000 copies of Alu Sg1, a descendant of Alus Sx and J, appeared later - around 35 mya (*26*). Although an approximately twenty-one fold difference exists in the amplification of these two Alus during the primate radiation, the ratio of microRNA-related hairpin sequences derived from these retrotransposons (hsamir-619 and hsa-mir-566) has been maintained at a constant ratio (1.1:1.0) across every chromosome of both the human and chimpanzee genome. The explanation for this remarkable degree of conservation among hsa-mir-619 and -566 is not known (Table 3 and Table 4). However, the ratio of sequences similar to known microRNAs is clearly a function of both the invasive success of the transposons,

Table 2. Distribution of microRNA hairpin BLAST hits across chimpanzee chromosomes

Chromosome	Mbp	% genome bp	No. genes	% genes	BLAST HITS	% BLAST HITS	OBS/EXP (% bp) BLAST HITS*	OBS/EXP (% genes) BLAST HITS*	No. unique microRNA hairpins†
1	230	7.25	2771	9.82	70777	9.28	1.28	0.94	320
2A	224	3.59	962	3.41	27838	3.65	1.02	1.07	64
2B	250	7.88	951	3.37	27175	3.56	0.45	1.06	115
3	204	6.43	1606	5.69	44871	5.88	0.91	1.03	216
4	195	6.14	1220	4.32	28434	3.73	0.61	0.86	129
5	184	5.80	1320	4.68	37180	4.87	0.84	1.04	193
6	174	5.48	1542	5.47	33809	4.43	0.81	0.81	90
7	160	5.04	1471	5.21	42844	5.62	1.11	1.08	175
8	145	4.57	1027	3.64	27782	3.64	0.80	1.00	82
9	139	4.38	1159	4.11	31331	4.11	0.94	1.00	346
10	135	4.25	1118	3.96	40126	5.26	1.24	1.33	106
11	134	4.22	1591	5.64	31724	4.16	0.99	0.74	102
12	135	4.25	1436	5.09	42097	5.52	1.30	1.08	115
13	116	3.65	522	1.85	11523	1.51	0.41	0.82	26
14	107	3.37	925	3.28	25414	3.33	0.99	1.02	64
15	100	3.15	877	3.11	28515	3.74	1.19	1.20	93
16	91	2.87	1115	3.95	31406	4.12	1.43	1.04	75
17	83	2.61	1490	5.28	50315	6.60	2.53	1.25	129
18	77	2.43	483	1.71	13590	1.78	0.73	1.04	107
19	64	2.02	1629	5.77	37419	4.91	2.43	0.85	112
20	62	1.95	728	2.58	21553	2.83	1.45	1.10	156
21	46	1.45	319	1.13	10476	1.37	0.95	1.22	105
22	50	1.58	586	2.08	19692	2.58	1.63	1.24	75
X	155	4.88	1198	4.25	25760	3.38	0.69	0.79	178
Y	24	0.76	165	0.58	1136	0.15	0.20	0.26	8
genome	3174		28211		762787				

*Frequency of BLAST hit observations normalized to the expected number if all BLAST hits were distributed evenly among all chromosomes or known genes. †Total number of unique microRNA related hairpins (with distinct microRNA identification (ID) numbers) matched by BLAST search

and the degree to which the associated microRNAs have diverged from a common founder sequence. In some cases (e.g. hsa-mir-566), the sequence of a mature microRNA has been highly conserved within numerous introns. In others, such as those related to hsa-mir-619, the transposon-derived intronic sequences do not resemble known mature microRNAs, but maintain strong homology to a microRNA hairpin, and have their ancestral basis in a microRNA paradigm. Notably, despite common origins within transposable elements, the hairpins themselves are much more highly conserved than surrounding DNA sequence, a finding that supports evolutionarily preservation specifically of the hairpin related motifs, rather than nearby sequences within the ancestral Alu or other repetitive elements.

The finding that hundreds of thousands of microRNA sequences have expanded in concert with the primate radiation, and persist as a predominant feature of the genome, may have significant implications with regard to the evolution of human DNA. For example, it is reasonable to imagine that a diaspora of transposable elements encoding microRNAs capable of gene network regulation might promote substantial organism diversity (and complexity), despite a comparatively modest number of discrete genes. High level sequence conservation (Table 3, Figure 1) and stoichiometry (Table 3) among microRNA hairpins suggest a recondite selective advantage important to maintenance or evolution of the primate genome. Stem-loop models of the most prevalent hairpins support the notion that hsa-mir-566 and hsa-mir-619 form structures that expose complementary bases to each other (as well as to their respective reverse complements; an example is provided for hsa-mir-566 in Figure 4). Such structures could promote RNA binding interactions, such as an adaptive role silencing transposable element expression (*10*). On the other hand, this sort of stem-loop complementarity might instead contribute to chromosomal pairing during meiosis as proposed earlier by Forsdyke (*15*, although the circumstances under which this might occur are not known with certainty). The relative paucity (by 3-15 fold; Table 1) of microRNA-related hairpin sequences on the Y chromosome (which does not extensively pair during meiosis) could be compatible with the latter interpretation. In either case, the observation that legions of microRNA-related hairpins occur in the primate genome, including tens of thousands that encode a mature microRNA, will need to be reconciled with current models of transcriptional regulation and specificity. This is especially true given recent evidence for nonconventional pathways that process microRNAs, such as those that do not

Table 3. Percent of total BLAST hits among hairpins commonly observed in human introns

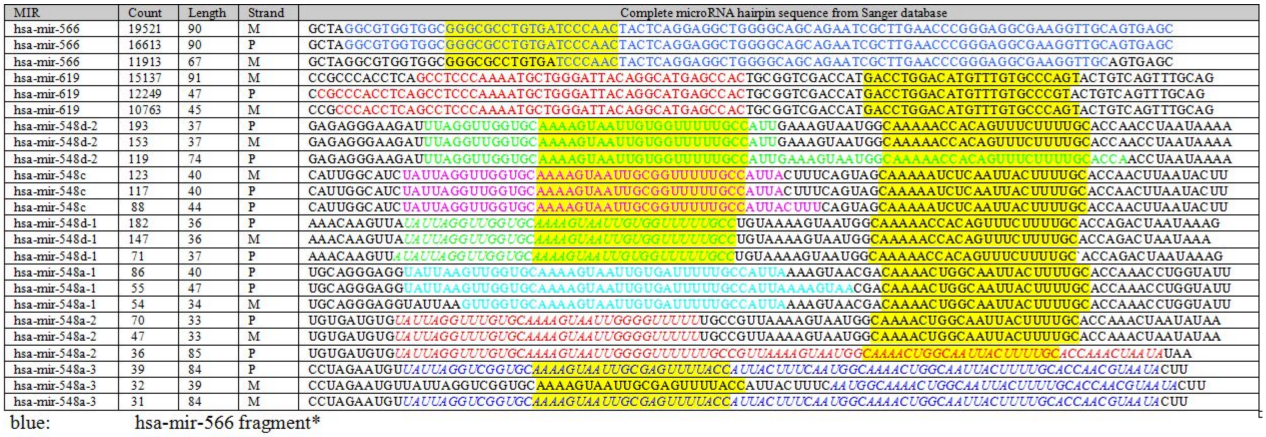
Chromosome	hsa-mir-619	hsa-mir-566	hsa-mir-548d-2	hsa-mir-548c	hsa-mir-548d-1	hsa-mir-548a-1	hsa-mir-548a-2	hsa-mir-548a-3	hsa-mir-645	hsa-mir-548b	hsa-mir-603	hsa-mir-570
1	52.68	45.31	0.40	0.32	0.28	0.15	0.14	0.12	0.06	0.05	0.03	0.02
2	51.72	44.98	0.67	0.61	0.50	0.31	0.20	0.24	0.06	0.11	0.07	0.09
3	51.36	44.98	0.66	0.66	0.56	0.29	0.22	0.22	0.05	0.13	0.07	0.09
4	50.88	45.13	0.75	0.72	0.57	0.33	0.22	0.31	0.07	0.13	0.11	0.07
5	51.62	45.33	0.62	0.63	0.45	0.27	0.19	0.20	0.04	0.10	0.04	0.07
6	51.45	45.11	0.68	0.65	0.52	0.32	0.28	0.25	0.05	0.10	0.07	0.11
7	50.96	45.41	0.69	0.63	0.55	0.34	0.23	0.22	0.07	0.09	0.03	0.05
8	51.38	44.59	0.85	0.79	0.67	0.35	0.29	0.34	0.09	0.12	0.07	0.09
9	50.48	44.44	0.63	0.63	0.52	0.28	0.15	0.18	0.04	0.05	0.03	0.10
10	51.79	44.74	0.66	0.68	0.54	0.20	0.24	0.27	0.04	0.15	0.10	0.10
11	51.85	45.19	0.53	0.50	0.42	0.24	0.20	0.17	0.06	0.08	0.05	0.07
12	52.53	44.97	0.54	0.45	0.40	0.27	0.16	0.15	0.07	0.07	0.08	0.05
13	51.04	45.08	0.80	0.83	0.57	0.40	0.35	0.35	0.06	0.10	0.14	0.18
14	52.56	44.98	0.53	0.54	0.40	0.24	0.17	0.20	0.03	0.02	0.05	0.09
15	52.29	45.60	0.41	0.44	0.33	0.18	0.15	0.12	0.03	0.06	0.04	0.04
16	52.86	45.20	0.37	0.43	0.28	0.16	0.13	0.12	0.05	0.06	0.05	0.04
17	51.33	46.86	0.33	0.32	0.26	0.13	0.01	0.11	0.03	0.05	0.02	0.03
18	50.75	44.14	0.97	1.00	0.85	0.52	0.36	0.35	0.06	0.16	0.10	0.09
19	52.22	46.15	0.18	0.15	0.14	0.10	0.10	0.09	0.01	0.06	----	-----
20	51.91	45.31	0.44	0.40	0.34	0.16	0.13	0.08	0.03	0.12	----	0.03
21	48.45	46.44	0.61	0.68	0.50	0.24	0.17	0.22	-----	0.12	0.10	0.19
22	51.95	46.71	0.21	0.18	0.13	0.08	0.07	0.04	0.04	-----	------	0.03
X	49.53	45.94	0.82	0.78	0.62	0.37	0.25	0.30	0.03	0.10	0.13	0.15
Y	51.65	47.06	0.29	0.29	0.24	0.14	0.14	0.05	0.05	-----	-----	0.05
genome avg.	51.59	45.34	0.56	0.54	0.44	0.25	0.19	0.19	0.05	0.09	0.07	0.06

Table 4. Percent of total BLAST hits among hairpins commonly observed in chimpanzee introns

	hsa-mir-619	hsa-mir-566	hsa-mir-548d-2	hsa-mir-548c	hsa-mir-548d-1	hsa-mir-548a-1	hsa-mir-548a-2	hsa-mir-548a-3	hsa-mir-548b	hsa-mir-603
chromosome										
1	53.64	43.40	0.46	0.42	0.35	0.20	0.21	0.16	0.09	0.05
2A	52.92	44.55	0.50	0.44	0.34	0.21	0.19	0.19	0.07	0.06
2B	52.69	43.10	0.77	0.72	0.57	0.38	0.32	0.30	0.10	0.09
3	52.51	43.28	0.66	0.69	0.55	0.29	0.27	0.23	0.11	0.09
4	52.05	43.83	0.75	0.68	0.58	0.33	0.23	0.26	0.11	0.12
5	52.58	44.14	0.58	0.60	0.42	0.28	0.19	0.18	0.12	0.07
6	52.34	43.62	0.67	0.69	0.53	0.33	0.36	0.28	0.14	0.10
7	52.14	43.95	0.74	0.67	0.59	0.30	0.23	0.22	0.09	0.05
8	52.51	43.02	0.84	0.82	0.68	0.39	0.33	0.38	0.12	0.14
9	52.01	42.63	0.69	0.63	0.54	0.29	0.22	0.21	0.08	0.06
10	53.18	43.68	0.57	0.61	0.46	0.20	0.17	0.20	0.08	0.05
11	53.07	43.54	0.51	0.54	0.41	0.27	0.24	0.28	0.10	0.09
12	53.12	43.92	0.46	0.46	0.41	0.24	0.19	0.22	0.09	0.10
13	51.98	43.77	0.86	0.75	0.64	0.39	0.41	0.36	0.15	0.21
14	53.71	43.3	0.52	0.50	0.40	0.27	0.27	0.29	0.07	0.09
15	53.65	43.82	0.42	0.40	0.30	0.17	0.20	0.17	0.08	0.04
16	54.15	43.74	0.37	0.36	0.29	0.12	0.13	0.12	0.04	0.04
17	52.03	46.05	0.28	0.31	0.23	0.11	0.12	0.11	0.05	0.04
18	51.64	42.88	0.87	0.94	0.95	0.50	0.35	0.43	0.18	0.18
19	53.45	44.75	0.19	0.19	0.16	0.11	0.14	0.13	0.07	0.02
20	52.88	43.82	0.45	0.41	0.37	0.17	0.16	0.10	0.10	0.01
21	49.74	44.76	0.59	0.73	0.45	0.26	0.26	0.30	0.14	0.16
22	52.66	45.53	0.18	0.19	0.14	0.07	0.09	0.11	0.06	0.03
X	50.27	44.23	0.83	0.09	0.62	0.36	0.27	0.32	0.19	0.12
Y	55.11	43.05	0.44	0.44	0.35	0.26	0.18	0.18	0.00	0.00
genome avg.	52.70	43.90	0.55	0.54	0.43	0.25	0.22	0.21	0.10	0.07

Table 5. Common human intronic sequences with similarities to hsa-mir-566 hairpin

Intronic Sequence	Count
ggcgtagtggcgggcgcctgtagtcccagctacttgggaggctgaggcaggagaatggcgtgaacccgggaggcggagcttgcagtgagc	362
ggcgcggtggcgggcgcctgtagtcccagctactcgggaggctgaggcaggagaatggcgtgaacccgggaagcggagcttgcagtgagc	282
ggcgtggtagcgggcgcctgtagtcccagctactcgggaggctgaggcaggagaatggcgtgaacccgggaggcggagcttgcagtgagc	102
ggcgtggtggcgggcgcctgtagtcccagctactcgggaggctgaggcaggagaatggcgtgaacccgggaggcggagcttgcagtgagc	78
ggcgtagtggcgggcgcctgtagtcccagctactcgggaggctgaggcaggagaatggcgtgaacccgggaggcggagcttgcagtgagc	47
ggcgtggtggcgggcgcctgtagtcccagctacttgggaggctgaggcaggagaatggcgtgaacccgggaggcggagcttgcagtgagc	42
ggcgcggtggcgggcgcctgtagtcccagctactcgggaggctgaggcaggagaatggcgtgaacccgggaggcggagcttgcagtgagc	34
ggcgtggtggcgggcgcctgtagtcccagctactcgggaggctgaggcaggagaatggcgtgaacctgggaggcggagcttgcagtgagc	25
ggcgcggtggcgggcgcctgtagtcccagctactggggaggctgaggcaggagaatggcgtgaacccgggaagcggagcttgcagtgagc	23
ggcgtggtggcgggtgcctgtagtcccagctactcgggaggctgaggcaggagaatggcgtgaacccgggaggcggaacttgcagtgagc	21
ggcgtggtggcaggtgcctgtagtcccagccacttgggaggctgaggcaggacaatggcatgaacctgggaggcggaggttgcagtgagc	20
ggcgtgatggcaggtgcctgtaatcccagctactcaggaggctgagacaggagaatcgcttgaacccaggaggcggaggttgcagtgagc	20
ggcgtggtggcgagcacctgtcatcccagctgcttgggaggctgaggcaggagaatggcttgaaccctggaggcagaggttgtagtgagc	20
ggcgtggtggcgggtgcctgtagtcccagctacttgagaggctgaggcaggagaatggagtgaacccaggaggcggagattgccgtgagc	20
ggcgtagtggcgggcgcctgtagtcccagctacttgggaggctgaggcaggagaatggcatgaacccgggaggcggagcttgcagtgagc	19



MIR	Count	Length	Strand	Complete microRNA hairpin sequence from Sanger database
hsa-mir-566	19521	90	M	GCTAGGCGTGGTGGCGGGCGCCTGTGATCCCAACTACTCAGGAGGCTGGGGCAGCAGAATCGCTTGAACCCGGGAGGCGAAGGTTGCAGTGAGC
hsa-mir-566	16613	90	P	GCTAGGCGTGGTGGCGGGCGCCTGTGATCCCAACTACTCAGGAGGCTGGGGCAGCAGAATCGCTTGAACCCGGGAGGCGAAGGTTGCAGTGAGC
hsa-mir-566	11913	67	M	GCTAGGCGTGGTGGCGGGCGCCTGTGATCCCAACTACTCAGGAGGCTGGGGCAGCAGAATCGCTTGAACCCGGGAGGCGAAGGTTGCAGTGAGC
hsa-mir-619	15137	91	M	CCGCCCACCTCAGCCTCCCAAAATGCTGGGATTACAGGCATGAGCCACTGCGGTCGACCATGACCTGGACATGTTTGTGCCCAGTACTGTCAGTTTGCAG
hsa-mir-619	12249	47	P	CCGCCCACCTCAGCCTCCCAAAATGCTGGGATTACAGGCATGAGCCACTGCGGTCGACCATGACCTGGACATGTTTGTGCCCGTACTGTCAGTTTGCAG
hsa-mir-619	10763	45	M	CCGCCCACCTCAGCCTCCCAAAATGCTGGGATTACAGGCATGAGCCACTGCGGTCGACCATGACCTGGACATGTTTGTGCCCAGTACTGTCAGTTTGCAG
hsa-mir-548d-2	193	37	P	GAGAGGGAAGAUUUAGGUUGGUGCAAAAGUAAUUGUGGUUUUUGCCAUUGAAAGUAAUGGCAAAAACCACAGUUUCUUUUGCACCAACCUAAUAAAA
hsa-mir-548d-2	153	37	M	GAGAGGGAAGAUUUAGGUUGGUGCAAAAGUAAUUGUGGUUUUUGCCAUUGAAAGUAAUGGCAAAAACCACAGUUUCUUUUGCACCAACCUAAUAAAA
hsa-mir-548d-2	119	74	P	GAGAGGGAAGAUUUAGGUUGGUGCAAAAGUAAUUGUGGUUUUUGCCAUUGAAAGUAAUGGCAAAAACCACAGUUUCUUUUGCACCAACCUAAUAAAA
hsa-mir-548c	123	40	M	CAUUGGCAUCUAUUAGGUUGGUGCAAAAGUAAUUGCGGUUUUUGCCAUUACUUUCAGUAGCAAAAAUCUCAAUUACUUUUGCACCAACUUAAUACUU
hsa-mir-548c	117	40	P	CAUUGGCAUCUAUUAGGUUGGUGCAAAAGUAAUUGCGGUUUUUGCCAUUACUUUCAGUAGCAAAAAUCUCAAUUACUUUUGCACCAACUUAAUACUU
hsa-mir-548c	88	44	P	CAUUGGCAUCUAUUAGGUUGGUGCAAAAGUAAUUGCGGUUUUUGCCAUUACUUUCAGUAGCAAAAAUCUCAAUUACUUUUGCACCAACUUAAUACUU
hsa-mir-548d-1	182	36	P	AAACAAGUUAUAUUAGGUUGGUGCAAAAGUAAUUGUGGUUUUUGCCUGUAAAAGUAAUGGCAAAAACCACAGUUUCUUUUGCACCAGACUAAUAAAG
hsa-mir-548d-1	147	36	M	AAACAAGUUAUAUUAGGUUGGUGCAAAAGUAAUUGUGGUUUUUGCCUGUAAAAGUAAUGGCAAAAACCACAGUUUCUUUUGCACCAGACUAAUAAA
hsa-mir-548d-1	71	37	P	AAACAAGUUAUAUUAGGUUGGUGCAAAAGUAAUUGUGGUUUUUGCCUGUAAAAGUAAUGGCAAAAACCACAGUUUCUUUUGC`ACCAGACUAAUAAAG
hsa-mir-548a-1	86	40	P	UGCAGGGAGGUAUUAAGUUGGUGCAAAAGUAAUUGUGAUUUUUGCCAUUAAAAGUAACGACAAAACUGGCAAUUACUUUUGCACCAAACCUGGUAUU
hsa-mir-548a-1	55	47	P	UGCAGGGAGGUAUUAAGUUGGUGCAAAAGUAAUUGUGAUUUUUGCCAUUAAAAGUAACGACAAAACUGGCAAUUACUUUUGCACCAAACCUGGUAUU
hsa-mir-548a-1	54	34	M	UGCAGGGAGGUAUUAAGUUGGUGCAAAAGUAAUUGUGAUUUUUGCCAUUAAAAGUAACGACAAAACUGGCAAUUACUUUUGCACCAAACCUGGUAUU
hsa-mir-548a-2	70	33	P	UGUGAUGUGUAUUAGGUUUGUGCAAAAGUAAUUGGGGUUUUUGCCGUUAAAAGUAAUGGCAAAACUGGCAAUUACUUUUGCACCAAACUAAUAUAA
hsa-mir-548a-2	47	33	M	UGUGAUGUGUAUUAGGUUUGUGCAAAAGUAAUUGGGGUUUUUGCCGUUAAAAGUAAUGGCAAAACUGGCAAUUACUUUUGCACCAAACUAAUAUAA
hsa-mir-548a-2	36	85	P	UGUGAUGUGUAUUAGGUUUGUGCAAAAGUAAUUGGGGUUUUUGCCGUUAAAAGUAAUGGCAAAACUGGCAAUUACUUUUGCACCAAACUAAUAUAA
hsa-mir-548a-3	39	84	P	CCUAGAAUGUUAUUAGGUCGGUGCAAAAGUAAUUGCGAGUUUUACCAUUACUUUCAAUGGCAAAACUGGCAAUUACUUUUGCACCAACGUAAUACUU
hsa-mir-548a-3	32	39	M	CCUAGAAUGUUAUUAGGUCGGUGCAAAAGUAAUUGCGAGUUUUACCAUUACUUUCAAUGGCAAAACUGGCAAUUACUUUUGCACCAACGUAAUACUU
hsa-mir-548a-3	31	84	M	CCUAGAAUGUUAUUAGGUCGGUGCAAAAGUAAUUGCGAGUUUUACCAUUACUUUCAAUGGCAAAACUGGCAAUUACUUUUGCACCAACGUAAUACUU

blue:	hsa-mir-566 fragment*
red:	hsa-mir-619 fragment*
green:	hsa-mir-548d-2 fragment*
pink:	hsa-mir-548c fragment*
italic green:	has-mir-548d-1 fragment*
light blue:	hsa-mir-548a-1 fragment*
italic red:	hsa-mir-548a-2 fragment*
italic blue:	hsa-mir-548a-3 fragment*
yellow:	position of mature microRNA sequence

*consensus sequence obtained from Sanger hairpin database - corresponds to intronic hairpin fragment BLAST findings

Figure 1. Alignment of common human intronic hairpin fragments with complete microRNA.

require Drosha (*11*). For example, although hsa-mir-566 has been implicated as contributing to human hematopoesis, the very high incidence of paralogs shown here has not been previously considered (*27*). Similarly, hsa-mir-548 related family members are known to regulate cancer gene expression (*28)*. The specificity of this observation may need to be reevaluated in light of the thousands of genomic paralogs described by the current report.

In summary, a massive number of under-appreciated microRNA-related hairpin structures are present in human and chimpanzee genomes. It is not yet known whether these sequences played a role during human evolution and are now dormant, or retain partial activity and contribute to existing genetic networks in man. The findings suggest an expanded role for introns and transposable elements as a source of microRNA, and that variants of known regulatory RNAs may contribute to expression on a genome-wide basis.

6. ACKNOWLEDGEMENTS

We thank Ms. Jenny Mott for her role in preparing the manuscript. This work was supported by the National Institutes of Health [P30DK72482 to E.S.]; and the Cystic Fibrosis Foundation [R464-CR11 to E.S.].

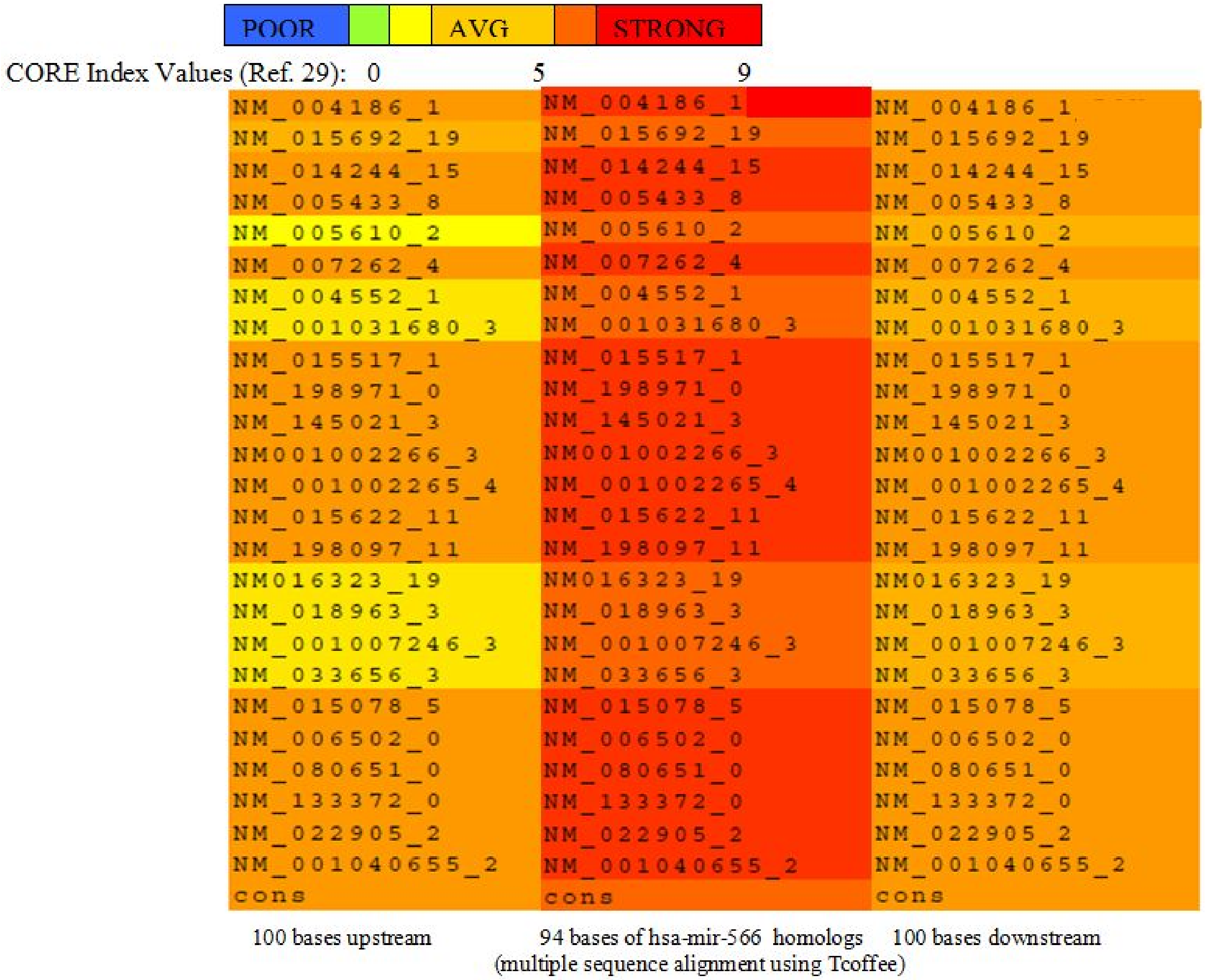
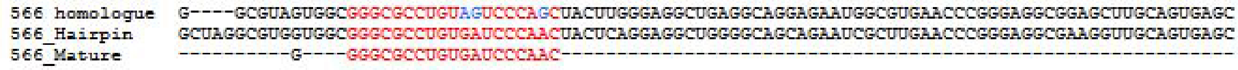


Figure 2. Hairpin sequence conservation of hsa-mir-566 hairpin relative to adjacent regions.



```
566 homologue   G----GCGUAGUGGCGGGCGCCUGUAGUCCCAGCUACUUGGGAGGCUGAGGCAGGAGAAUGGCGUGAACCCGGGAGGCGGAGCUUGCAGUGAGC
566_Hairpin     GCUAGGCGUGGUGGCGGGCGCCUGUGAUCCCAACUACUCAGGAGGCUGGGGCAGCAGAAUCGCUUGAACCCGGGAGGCGAAGGUUGCAGUGAGC
566_Mature      ----------G----GGGCGCCUGUGAUCCCAAC------------------------------------------------------------
```

Figure 3. Alignment of most common intronic sequence (566 homologue) with hsa-mir-566 hairpin and mature microRNA

7. REFERENCES

1. Mattick, J. S. RNA regulation: a new genetics? *Nat Rev Genet* 5, 316-323 (2004)

2. Mattick, J. S., J. Gagen. Accelerating networks in biology, engineering, and society. *Science* 307, 856-858 (2005)

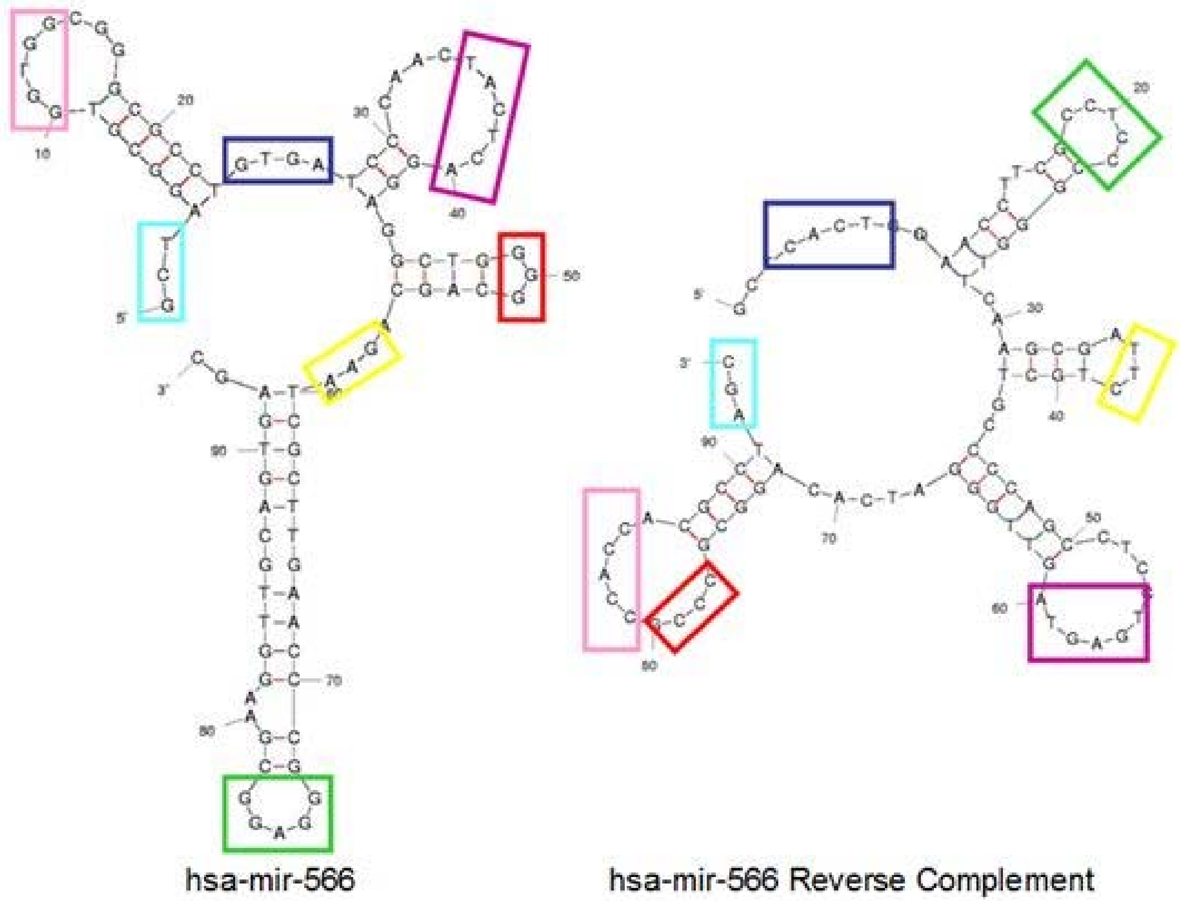
3. Mattick, J. S., M. J. Gagen. The evolution of controlled multitasked gene networks: The role of introns and other noncoding RNAs in the development of complex organisms. *Mol Biol Evol* 18, 1611-1630 (2001)

4. Mattick, J. S. What makes a human? *The Scientist* 19, 32 (2005)

5. Hill, A. E., J. S. Hong, H. Wen, L. Teng, D. T. McPherson, S. A. McPherson, D. N. Levasseur, E. J. Sorscher. Micro-RNA-like effects of complete intronic sequences. *Front Biosci* 11, 1998-2006 (2006)

6. Guimbellot, J. S., S. W. Erickson, T. Mehta, H. Wen, G. P. Page, E. J. Sorscher, J. S. Hong. Correlation of microRNA levels during hypoxia with predicted target mRNAs through genome-wide microarray analysis. *BMC Med Genomics* 2, 15 (2009)

7. Rodriguez, A., S. Griffiths-Jones, J. L. Ashurst, A. Bradley. Identification of mammalian microRNA host genes and transcription units. *Genome Res* 14, 1902-1910 (2004)



hsa-mir-566

hsa-mir-566 Reverse Complement

Figure 4. Regions of complementarity between hsa-mir-566 and reverse complement; modeled by stem loop as in ref. 30, 31.

8. Roger, A. J., P. J. Keeling, W. F. Doolittle. Introns, the broken transposons. *Soc Gen Physiol Ser* 49, 27-37 (1994)

9. Sela, N., B. Mersch, N. Gal-Mark, G. Lev-Maor, A. Hotz-Wagenblatt, G. Ast. Comparative analysis of transposed element insertion within human and mouse genomes reveals *Alu*'s unique role in shaping the human transcriptome. *Genome Biol* 8, R127 (2007)

10. Piriyapongsa, J., L. Marino-Ramirez, I. K. Jordan. Origin and evolution of human microRNAs from transposable elements. *Genetics* 176, 1323–1337 (2007)

11. Ruby, J. G., C. H. Jan, D. P. Bartel. Intronic microRNA precursors that bypass Drosha processing. *Nature* 5, 83-86 (2007)

12. Griffiths-Jones, S., H. K. Saini, S. van Dongen, A. J. Enright. miRBase: tools for microRNA genomics. *NAR* 36 (Database Issue), D154-D158 (2008)

13. Poirot, O., E. O'Toole, C. Notredame. Tcoffee@igs: A web server for computing, evaluating and combining multiple sequence alignments. *Nucleic Acids Res* 31, 3503–3506 (2003)

14. Forsdyke, D. Are introns in-series error-detecting sequences? *J Theor Biol* 93, 861-866 (1980)

15. Forsdyke, D. A stem-loop "kissing" model for the initiation of recombination and the origin of introns. *Mol Biol Evol* 12, 949-958 (1995)

16. Forsdyke, D. An alternative way of thinking about stem-loops in DNA. A case study of the human G0S2 gene. *J Theor Biol* 192, 489-504 (1998)

17. Barrette, I. H., S. McKenna, D. R. Taylor, D. Forsdyke. Introns resolve the conflict between base order dependent stem-loop potential and the encoding of RNA or protein: further evidence from overlapping genes. *Gene* 270, 181-189 (2001)

18. Doyle, G. G. A general theory of chromosome pairing based on the palindromic DNA model of Sobell with modifications and amplifications. *J Theor Biol* 70, 171-184 (1978)

19. Ares, M., L. Grate, M. H. Pauling. A handful of intron-containing genes produce the lion's share of yeast mRNA. *RNA* 5, 1138-1139 (1999)

20. Frederickson, R. M., M. R. Micheau, A. Iwamoto, N. G. Miyamoto. 5' flanking and first intron sequences of the human beta-actin gene required for efficient promoter activity. *Nucleic Acids Res* 17, 253-270 (1989)

21. Luo, M., R. Reed. Splicing is required for rapid and efficient mRNA export in metazoans. *Proc Natl Acad Sci USA* 96, 14937-14942 (1999)

22. Fong, Y. W., Q. Zhou. Stimulatory effect of splicing factors on transcriptional elongation. *Nature* 414, 929-933 (2001)

23. Maniatis, R., R. Reed. An extensive network of coupling among gene expression machines. *Nature* 416, 499-506 (2002)

24. Hormuzdi, S., R. Penttinen, R. Jaenisch, P. Bornstein. A gene-targeting approach identifies a function for the first intron in expression of the alpha1(I) collagen gene. *Mol Cell Biol* 18, 3368-3375 (1998)

25. Beaton, M. J., T. Cavalier-Smith. Eukaryotic noncoding DNA is functional: evidence from the differential scaling of cryptomonad genomes. *Proc R Soc Lond B Biol Sci* 266, 2053-2059 (1999)

26. Roy-Engel, A. M., M. A. Batzer, P. L. Deininger. Evolution of Human Retrosequences: Alu. In *Encyclopedia of Life Sciences (ELS)* Wiley, Chichester, DOI: 10.1002/9780470015902.a0005131.pub2 (2008)

27. Kim, Y. C., Q. Wu, J. Chen, Z. Xuan, Y. C. Jung, M. Q. Zhang, J. D. Rowley, S. M. Wang. The transcriptome of human CD34+ hematopoietic stem-progenitor cells. *Proc. Natl. Acad. Sci. U S A* 106, 8278-83 (2009)

28. Pinyapongsa, J., I. K. Jordan. A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS ONE* 2, e203 (2007)

29. Notredame, C., C. Abergel. Using Multiple Alignment Methods to Assess the Quality of Genomic Data Analysis. In Andrade,M. (ed.), *Bioinformatics and Genomes: Current Perspectives*. Horizon Scientific Press, Norwich, pp. 30-50 (2003)

30. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* 31, 3406-15 (2003)

31. Mathews, D. H., J. Sabina, M. Zuker, D. H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* 288, 911-940 (1999)

32. Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, D. Haussler. The human genome browser at UCSC. *Genome Res* 12, 996-1006 (2002)

33. Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res* 25, 3389-3402 (1997)

Key Words: microRNAs, Introns, Gene Regulation, Transposons

Send correspondence to: Aubrey Hill, MCLM 794, 1530 3rd Avenue South, Birmingham, AL 35294, Tel: 205-996-4136, Fax: 205-934-5473, E-mail: ahill@uab.edu