Mobile Genetic Elements 2:3, 184-192; May/June, 2011; © 2011 Landes Bioscience

OrbId
Origin-based identification of microRNA targets

Teresa J. Filshtein[1,†] and Craig O. Mackenzie,[1,†] Maurice D. Dale,[1] Paul S. Dela-Cruz,[1] Dale M. Ernst,[1] Edward A. Frankenberger,[1] Chunyan He,[1] Kaylee L. Heath,[1] Andria S. Jones,[1] Daniel K. Jones,[1] Edward R. King,[1] Maggie B. Maher,[1] Travis J. Mitchell,[1] Rachel R. Morgan,[1] Sirisha Sirobhushanam,[1] Scott D. Halkyard,[1] Kiran B. Tiwari,[1] David A. Rubin,[1] Glen M. Borchert[1,2,*] and Erik D. Larson[1]

[1]School of Biological Sciences; Illinois State University; Normal, IL USA; [2]Department of Biological Sciences; University of South Alabama; Mobile, AL USA

[†]These authors contributed equally to this work.

Keywords: Alu, LINE, microRNA, miR, repetitive, target prediction, TE, transposable, UTR

Abbreviations: bp, basepair; LINE, long interspersed repeated element; LTR, long terminal repeat; miR, microRNA; mRNA, messenger RNAs; nt, nucleotide; ORF, open reading frame; Pol III, RNA polymerase III; pre-miR, miR hairpin; pri-miR, initial miR transcript; RISC, RNA induced silencing complex; RNAi, RNA interference; SINE, short interspersed repeated elements; siRNA, small interfering RNA; TE, transposable element; tRNA, transfer RNA; UTR, untranslated region

microRNAs coordinate networks of mRNAs, but predicting specific sites of interactions is complicated by the very few bases of complementarity needed for regulation. Although efforts to characterize the specific requirements for microRNA (miR) regulation have made some advances, no general model of target recognition has been widely accepted. In this work, we describe an entirely novel approach to miR target identification. The genomic events responsible for the creation of individual miR loci have now been described with many miRs now known to have been initially formed from transposable element (TE) sequences. In light of this, we propose that limiting miR target searches to transcripts containing a miR's progenitor TE can facilitate accurate target identification. In this report we outline the methodology behind OrbId (Origin-based identification of microRNA targets). In stark contrast to the principal miR target algorithms (which rely heavily on target site conservation across species and are therefore most effective at predicting targets for older miRs), we find OrbId is particularly efficacious at predicting the mRNA targets of miRs formed more recently in evolutionary time. After defining the TE origins of > 200 human miRs, OrbId successfully generated likely target sets for 191 predominately primate-specific human miR loci. While only a handful of the loci examined were well enough conserved to have been previously evaluated by existing algorithms, we find ~80% of the targets for the oldest miR (miR-28) in our analysis contained within the principal Diana and TargetScan prediction sets. More importantly, four of the 15 OrbId miR-28 putative targets have been previously verified experimentally. In light of OrbId proving best-suited for predicting targets for more recently formed miRs, we suggest OrbId makes a logical complement to existing, conservation based, miR target algorithms.

Introduction

During the latter half of the 20th century one of the greatest achievements in genetic research was the meticulous cataloging of epistatic relationships between genetic loci. While new relationships brought new insights, they also created massive networks of seemingly endlessly interacting genetic pathways. In 1993, however, Lee et al. described an entirely new short noncoding RNA that, despite its size, would ultimately be recognized as an important player in deciphering complex genetic interactions.[1] These small microRNAs (miRs) are only ~20 nts in length (**Fig. 1A**) and are capable of coordinating the expressions of networks of mRNAs (mRNAs) through complementary basepairing.[1] Strikingly, over 1,900 unique human miRs have been cloned[2] since the first were discovered in 2001.[3-5] As such, it is of little surprise that miR research has seen a recent explosion of interest, especially considering that a single miR has the potential to control expression of dozens of genes and miR misregulations are commonly associated with oncogenesis (recently reviewed in ref. 6).

Whereas novel miR discovery has been forthcoming, progress in deciphering miR regulations has proven exceptionally challenging. This is largely due to miRs requiring very little sequence complementarity to the mRNAs they coordinate. In contrast to siRNAs which depend upon almost perfect complementarity to direct message degradation, miR target recognition and

*Correspondence to: Glen M. Borchert; Email: borchert@southalabama.edu
Submitted: 06/04/12; Revised: 07/20/12; Accepted: 07/24/12
http://dx.doi.org/10.4161/mge.21617

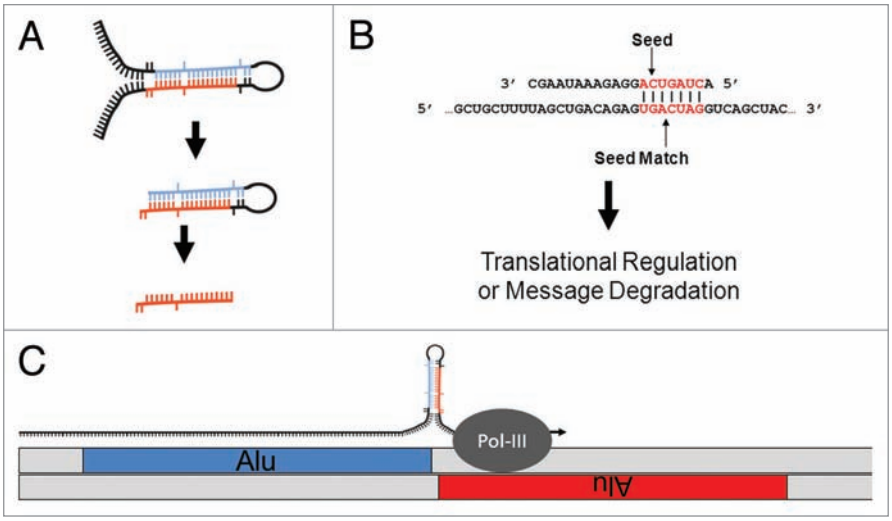

Figure 1. miR biology and origins. (A) miR generation. miRs can occur inter- or intragenically and be transcribed by either RNA Polymerase II or III.[24] Following transcription, the "pre-miR" hairpin (middle) is excised from the initial transcript (or pri-miR) (top) by Drosha. Once in the cytoplasm, the hairpin or stem loop is cleaved and denatured by Dicer to excise the ~20 nt mature miR (bottom). (B) miR seeds. A seed match between a miR (top) and target mRNA (bottom) is illustrated. The nucleotides in a miR generally referred to as a "seed" (nts 2 through 8) and a "seed match" in a mRNA are depicted in red. Basepairing is indicated by vertical lines. (C) Cartoon depicting the molecular origin of many miR loci. miRs were initially formed by the neighboring insertions of related TEs. A pri-miR is depicted just above the genome with an arrow indicating readthrough Pol-III transcription from a (+) strand Alu SINE into a neighboring (-) strand Alu. As illustrated, transcriptional readthrough would generate a RNA stem loop whose stems (loaded into the RISC machinery if processed) would correspond to the terminal nucleotides of the neighboring Alus. Figure adapted from reference 23.

consequent repression can be mediated through as few as 7 bps of complementarity. Generally thought to most frequently occur in the 5' miR sequence, these seven participating nts are typically referred to as the miR "seed" and the complement in a mRNA as the "seed match"[7-9] (**Fig. 1B**). The recurrent observation of complementarity between seed and seed match in a few initially characterized miR-target interactions lead to the majority of miR target recognition algorithms basing target searches on perfect seed matches. Following this, most algorithms differ primarily by the significance they attribute to seed match conservation between species, the presence of multiple seed matches in a given mRNA target and the extent of complementarity between the proposed target and remainder of the miR (recently reviewed refs. 10-14). While algorithms have been developed that do not require target site conservation across species (focusing instead on thermodynamic stability and target site secondary structure (e.g., PITA[15] and rna22[16]), the principal, most widely accepted target prediction algorithms (DIANA-microT,[17] miRanda,[18] PicTar,[19,20] and TargetScan[21]) each incorporate target site conservation into their prediction methodologies. Although efforts to characterize the specific requirements for miR target recognition continue to advance, to date the principal target algorithms typically suggest several hundred putative mRNA targets for each individual miR. As such is the case, no model of miR target prediction has been widely accepted.

Similar in rationale to the principal miR target prediction algorithms (although not requiring target site conservation across species), we have developed an entirely novel approach to miR target identification. First suggested by Smalheiser and Torvik,[22] the molecular events responsible for the genomic formation of many miR loci from transposable element (TE) sequences have now been described[22-28] (**Fig. 1C**). Having recently performed a series of detailed genomic analyses describing the TE origins of ~2,400 distinct miRs,[23] we hypothesized that a miR and its mRNA target sites might actually be formed in parallel by the ongoing colonization of a common ancestral transposable element (**Fig. 2**). In light of this, we propose that limiting miR target searches to mRNAs containing the TE initially giving rise to a miR can significantly hone accurate target identification. In this work we outline the methodology behind, and initial findings for, a novel miR target prediction strategy: OrbId (**Or**igin-**b**ased **Id**entification of microRNA targets). In all, we have successfully generated target sets for 191 unique miRs after applying OrbId to a set of 208 distinct human miRs of defined TE origin.[23] While the majority of OrbId putative targets were for recently formed miR loci, we did generate targets for the evolutionarily older miR-28 family and find our results largely in agreement with both traditional target prediction strategies[17,21] and existing experimental evidence.[29] Thus, the mRNA targets of a given miR can largely be predicted based on shared transposable element origins.

Results

Targets predicted for 92% of human miRs with defined TE origins. OrbId operates under the premise that a miR and its mRNA target sites were formed in parallel by the colonization of a common progenitor transposable element (**Fig. 2**). Utilizing this premise, we have successfully generated putative target sets for 191 of 208 human miRs with defined TE origins.[23] In stark contrast to the principal miR target algorithms currently utilized (which typically predict several hundred putative mRNA targets for individual miRs[17-19,21]), we find OrbId predicts significantly fewer mRNA targets per miR (average 7.9, median 3) (**Table 1**). In all, 59 produced a single mRNA target, 120 distinct miRs were suggested to have between 2 and 25 target mRNAs and 12 were predicted to target > 25 mRNAs (max = 94, putative targets for miR-574) (**Table S1**). In order to ensure strict adherence to the OrbId operating methodology, sequence alignments of unique mRNA target sites, miRs and progenitor TEs were independently verified (**Fig. 3**, **Table S2**).

Target sites are generally not preferentially located in 3'UTRs. miRs have now been conclusively shown to regulate target mRNAs through interactions with 5' untranslated

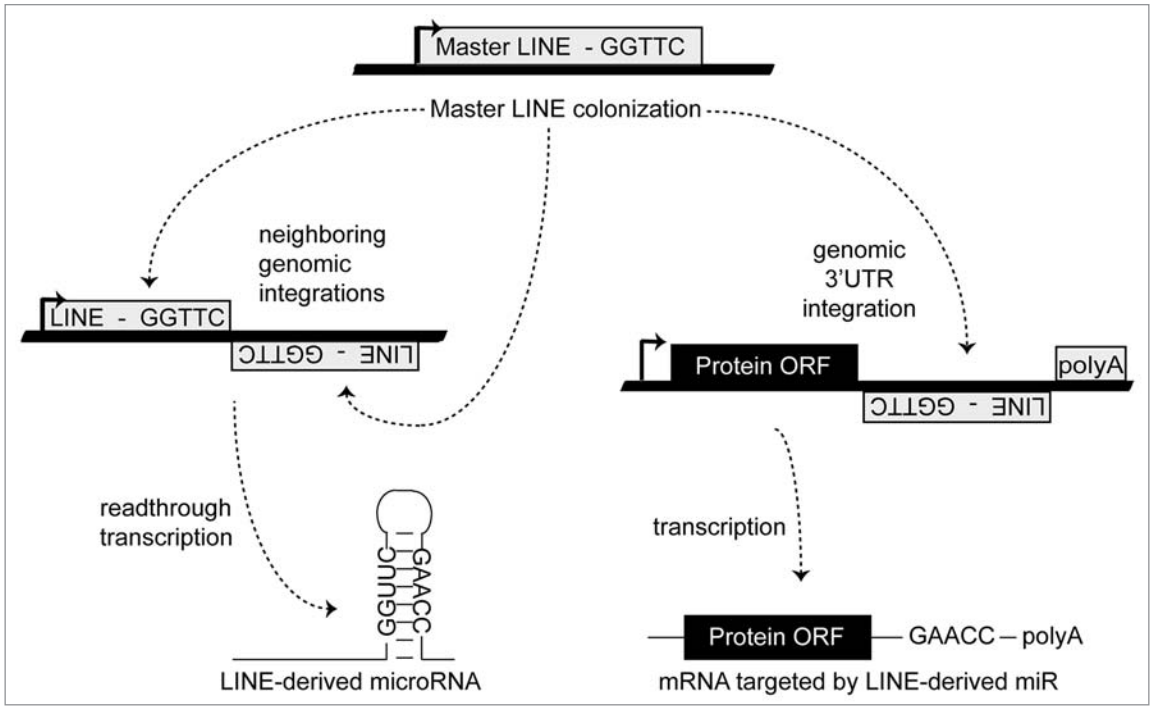

Figure 2. Establishing a miR regulatory network. miR regulatory networks are formed when an advantageous regulation arises from a series of random TE insertions into expressed genomic loci, and the formation of a TE juxtaposition by the positive and negative strand insertions of related TEs. Thick lines indicate genomic DNA and thin lines denote RNA. Figure adapted from reference 23.

regions (UTR) and open reading frame (ORF) sequences similar to 3' UTR interactions.[30-35] As such, it is somewhat surprising that most target prediction algorithms[16-18,20] predominately screen mRNA 3' UTRs for miR regulatory sites. In this analysis, we assessed all publically available human mRNA sequence regardless of functional annotation.[36,37] Strikingly, not only did we find strong evidence supporting 5' UTR and ORF regulations, we did not observe a general bias for 3' UTR target sites (**Table 1**). In all, of 1,529 unique predicted regulations, 970, 410 and 149 are located within 3' UTR, 5' UTR and ORF sequences respectively. When the average lengths of human 3' UTR (386 nt), 5' UTR (117 nt) and ORF (647 nt) sequences are taken in consideration, we find no significant bias for targeting to occur in either UTR preferentially. However, we did find target sites were approximately 12 times as likely to occur in noncoding UTR sequences than in ORF coding regions. Importantly, while we observed no general bias for miR targeting of 3' UTR, 5' UTR or ORF sequences, individual miR families showed significant targeting preferences. Of note, the targets of the principal mariner transposon derived miR family (miR-548) were located almost exclusively (> 99%) within 3' UTR sequences, the targets of the principal LINE derived miR family (miR-28) were similarly biased to occur within 3' UTR sequences (> 96%), but, in sharp contrast, the targets of a novel Alu SINE derived miR family (Alu-miR) were located predominately (> 81%) within 5' UTR sequences (**Table 2**; **Fig. S1**). Additionally, while less than 10% of putative targets were predicted to occur in ORFs [despite ORFs accounting for > 56% of the total transcript sequence examined (**Table 1**)], we identify two miRs, miR-544 and miR-301a-5p which are predicted to preferentially (> 90%) target ORF sequences (**Table 2**, **Fig. S1**).

Table 1. OrbId summary

Total human miRs analyzed	208
miRs with predicted targets	191
Average # of predicted mRNA targets	7.9
Median # of mRNA targets	3
Max # of mRNA targets	94
Min # of mRNA targets	1
Total # of human transcripts assessed	178,375
Mean transcript length	1151 nt
Total # of 3'UTR targets	970
Mean 3'UTR length	386 nt
Total # of 5'UTR targets	410
Mean 5'UTR length	117 nt
Total # of ORF targets	149
Mean ORF length	647

The full Ensembl[36] set of 178,375 unique human mRNA transcripts including 5' UTR, 3' UTR, and ORF annotations were compiled in and retrieved using the Biomart mining utility.[37] "Human miRs analyzed" correspond to the full set of human miR mature sequences identified by Borchert et al. as originating from TEs[23] and were obtained from the miR Registry miRBase.[2]

LINE L2B (miR-28) family. First identified in 2003[22] as arising from L2B LINE elements, miR-28 and miR-151 have long

Table 2. OrbId prediction set for select TE-derived human miRs

miR Name	Ensembl Gene ID	Gene Name	Diana, TS	Region
hsa-mir-28-5p	ENSG00000164136	IL15		5' UTR
	ENSG00000180957	PITPNB		5' UTR
	ENSG00000108309	RUNDC3A		3' UTR
	ENSG00000106608	URGCP	D, TS	3' UTR
	ENSG00000122741	DCAF10	D, TS	3' UTR
	ENSG00000144043	TEX261	D, TS	3' UTR
	ENSG00000152578	GRIA4	D, TS	3' UTR
	ENSG00000134046	MBD2		3' UTR
	ENSG00000117598	LPPR5.1	D, TS	3' UTR
	ENSG00000124466	LYPD3	D, TS	3' UTR
	ENSG00000102921	N4BP1	D, TS	3' UTR
	ENSG00000169016	E2F6	D, TS	3' UTR
	ENSG00000135999	EPC2	D	3' UTR
	ENSG00000123472	ATPAF1	D, TS	3' UTR
	ENSG00000116641	DOCK7		3' UTR
hsa-mir-301a5p	ENSG00000105856	HBP1	D	5' UTR
	ENSG00000175445	LPL		ORF
	ENSG00000082175	PGR		ORF
	ENSG00000166004	KIAA1731		ORF
	ENSG00000136573	BLK		ORF
hsa-mir-544a	ENSG00000144560	VGLL4	D, TS	5' UTR
	ENSG00000140632	GLYR1		ORF
	ENSG00000078018	MAP2	D, TS	ORF
	ENSG00000197279	ZNF165		ORF
	ENSG00000130066	SAT1		ORF
	ENSG00000183035	CYLC1		ORF
	ENSG00000142178	SIK1		ORF
	ENSG00000173681	CXorf23		3' UTR
hsa-mir-603	ENSG00000122692	SMU1		3' UTR
	ENSG00000102781	KATNAL1	D, TS	3' UTR
	ENSG00000116205	TCEANC2		3' UTR
	ENSG00000004468	CD38		3' UTR
	ENSG00000226264	HLA-DMB		3' UTR
	ENSG00000183908	LRRC55		3' UTR
	ENSG00000184040	FAM23B.1		3' UTR
	ENSG00000148483	TMEM236		3' UTR
	ENSG00000132623	ANKRD5		3' UTR
	ENSG00000144455	SUMF1		3' UTR
	ENSG00000215020	AL591684.1		3' UTR
	ENSG00000215033	AL603965.1		3' UTR
hsa-mir-1254-1	ENSG00000081760	AACS		5' UTR
	ENSG00000167077	MEI1		5' UTR
	ENSG00000238035	AC138035.1		5' UTR
	ENSG00000160991	ORAI2		3' UTR

"miR Name" refers to miRBase[2] annotation while "Ensembl Gene ID" and "Gene Name" were obtained using the Biomart mining utility.[37] "Diana, TS" refers to whether a predicted target is contained within publically accessible Diana (D) and TargetScan (TS) predictions.[17,21] "Region" refers to the location of a predicted target site within a given mRNA. miR-28-5p corresponds to the participating member of the miR-28 family. miR-1254-1 is a member of the Alu-miR family. miR-603 is a member of the miR-548 family.

Figure 3. miR-28 predicted target three way alignments. Alignments between OrbId predicted miR-28 target mRNAs (middle), a consensus L2B LINE (L2Plat1o) (top), and miR-28 (bottom). (*), base identity in the three aligning sequences. (^), base identity (indicating base pairing) between the miR and mRNA target only. (:), GU basepairing between miR and mRNA target. 3' UTR or 5' UTR targeting is indicated. Uracils are shown as thymines and UTRs have been reverse complemented for illustrative purposes.

been recognized as being related, and their numerous representative sequences across mammalia are collectively referred to as the miR-28 family.[2] Supporting this relationship, and despite there only being an ~10% likelihood that a given miR in this analysis would target the same mRNA as any other miR, we find ~76% of miR-28 and miR-151 proposed targets (11 of 15 and 11 of 14 respectively) common to both miRs (**Fig. 4**; **Table S1**). Our analyses also indicate the likelihood of a third, until now overlooked, member of the miR-28 family, miR-708. While initially formed from the same LINE element that gave rise to miR-28 and miR-151[23] and baring significant pre-miR homology to both miR-28 and miR-151[2] (**Fig. S2**), we find ~31% of miR-708 targets also constitute miR-28 family targets (**Fig. 4**; **Table S1**). Additionally, as the miR-28 family was the oldest in our analysis, miR-28 was one of the few miRs with publically available Diana and TargetScan predictions. Encouragingly we find ~80% of our miR-28 target predictions contained within the principal Diana and TargetScan prediction sets[16,20] (**Table 2**). Furthermore, over 25% of our putative miR-28 targets (4 of 15) have already been experimentally verified and shown to indeed regulate the mRNAs predicted by OrbId (ref. 29, data not shown).

miRs formed from Alu repeats. In contrast to the miR-548 and miR-28 families, the targets of a novel Alu SINE derived miR family (miR-566) were located predominately (> 80%) within 5' UTR sequences. While not as closely related as the miR-28 family, these Alu-derived miRs share several target relationships. While they may not constitute a traditional miR family based on common molecular origin, they could be considered to be a family in the sense of common targeting. In all, miRs -566, -1254, -1268, -1273, -1285, -1968, -1972 and -1973 appear to establish a significant network of target regulations (**Fig. S3**). Intriguingly, our findings are largely in agreement with previous reports suggesting that 3' UTR embedded Alu repeats frequently house novel, primate-specific miR target sites.[38-40]

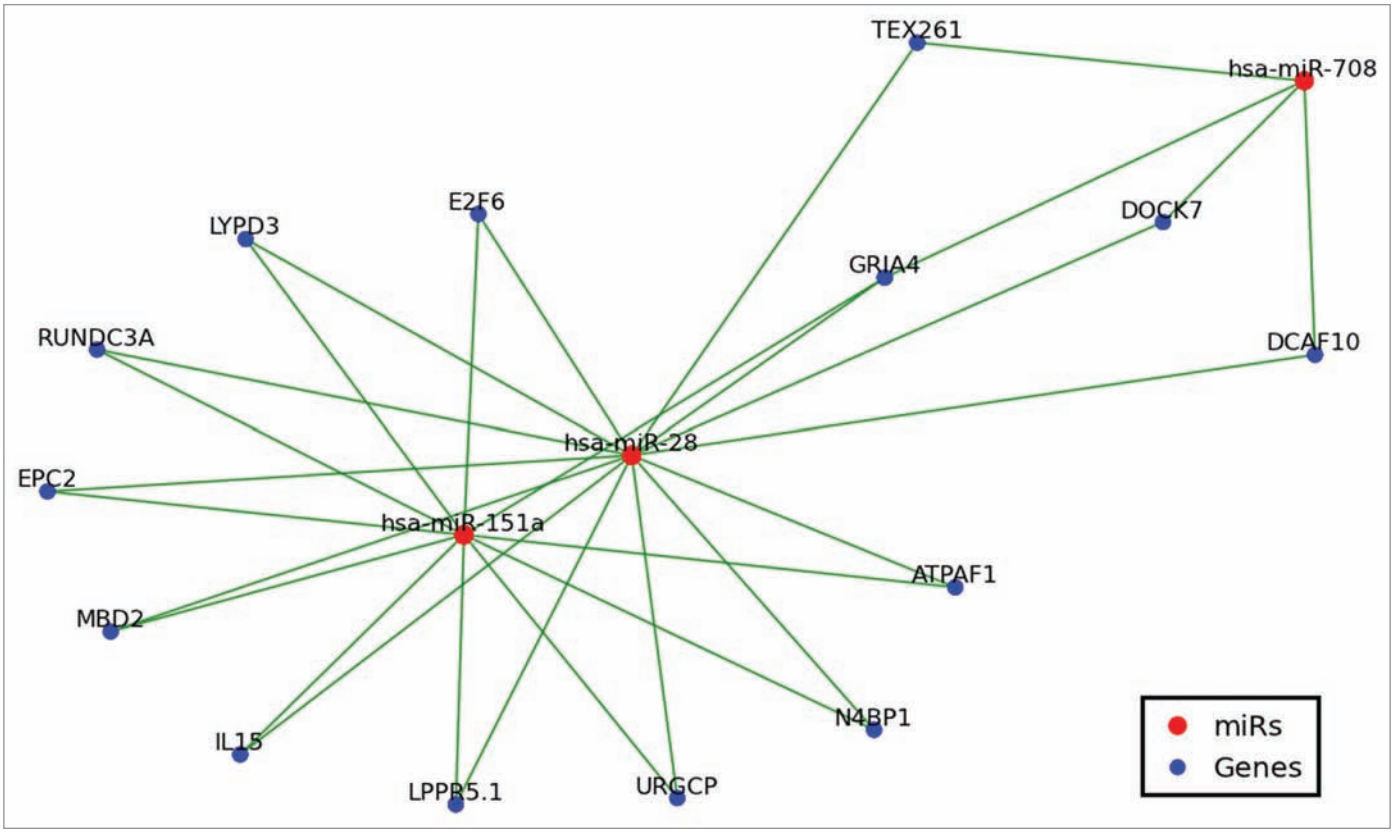

Figure 4. miR-28, miR-151 and miR-708 target network. Only shared targets are depicted including 14 of 15 miR-28-5p targets, 11 of 14 miR-151a-5p targets, and 4 of 13 miR-708 targets. Green lines indicate miR regulation.

Discussion

The genomic events responsible for the initial formation of numerous miR loci have recently been described.[23] The majority of these loci appear to have initially arisen from transposable element (TE) sequences. In addition to forming miR loci, we now hypothesize that TE mobilizations also generate miR regulatory networks by simultaneously integrating into existing mRNA expression cassettes (**Fig. 2**). Thus, the principle objective of this work was to utilize common TE ancestry to facilitate accurate prediction of miR-mRNA target interactions. To accomplish this, we have developed a novel methodology titled OrbId (Origin-based Identification of microRNA targets) (**Fig. 5**). OrbId contrasts sharply with current miR target algorithms[16-18,20] as these methodologies rely heavily on target site conservation across species and have therefore been primarily effective at predicting targets for well conserved miRs. OrbId is better suited for predicting the mRNA targets of evolutionarily younger miRs for which target site conservation searches are impractical. For example, the 70 human miR loci known to have been formed from primate-specific Alu repeats,[23,24] rodent-specific miRs formed from rodent specific B1 SINEs,[23] or the marsupial-specific miRs formed from marsupial-specific transposable elements.[25]

OrbId may also prove valuable in identifying taxon-specific targets of more conserved miRs. Requiring target site conservation across species has been effective at predicting many of the targets for conserved miRs. By design, however, traditional conservation-based miR target algorithms miss any targets arising from TE mobilizations following the initial establishment of a miR regulatory network. For example, if ongoing TE colonizations occur following speciation events, separate species might well acquire distinct, novel targets for existing miRs. Although beyond the scope of this analysis, more comprehensive species wide implementations of OrbId will be needed to fully evaluate the prevalence of such events.

Future analyses will unquestionably broaden the range of OrbId utility as the existing repertoire of defined miR-TE relationships continues to expand through the ongoing characterizations of additional miR loci and novel TE sequences. Importantly, de Koning et al. recently suggested that over two-thirds of the human genome were actually formed from repetitive elements.[44] While highly intriguing, the extent of the repetitive composition of the human genome remains a significant point of debate and attempts to fully clarify this issue remain ongoing. Should the work of de Koning et al. prove largely accurate, the incorporation of this information into current OrBId methodology would clearly result in marked increases in the definable number of putative miR::target relationships. Additionally, while this would likely predominately facilitate putative target identification for evolutionarily older miRs, it would also almost certainly require increased stringency to avoid concurrent increases in false positives. As a result of electing to limit our

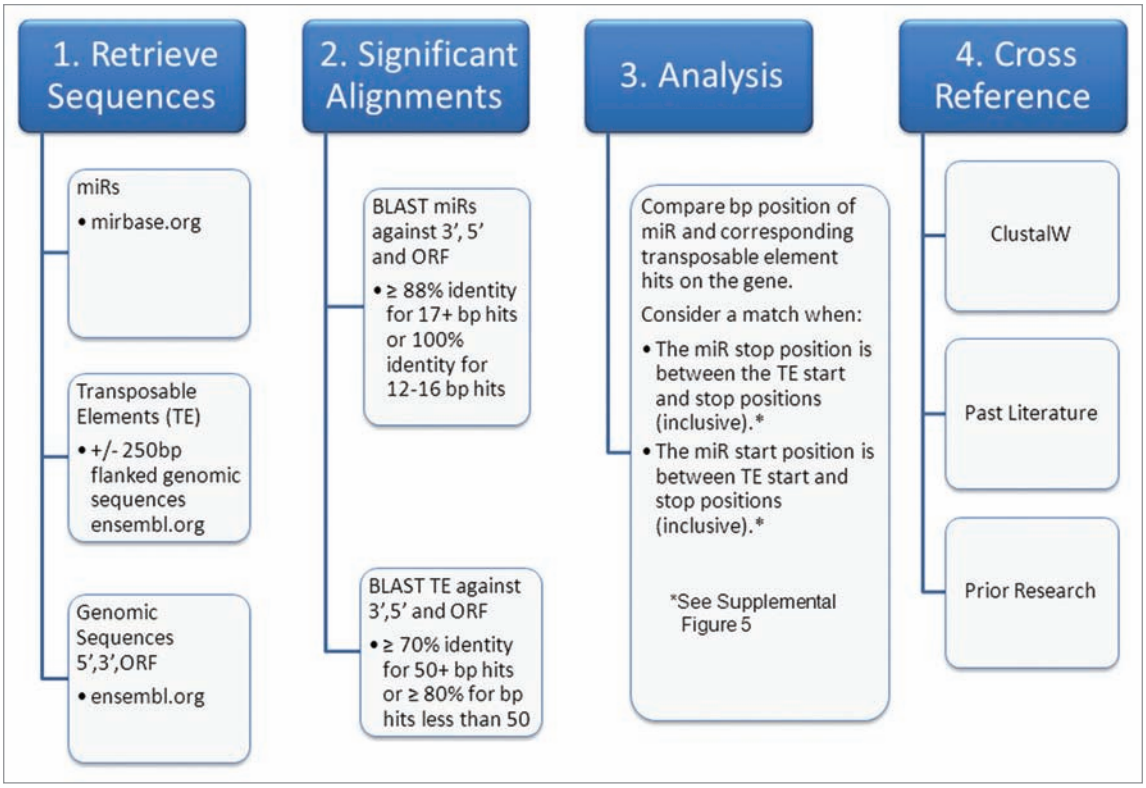

Figure 5. OrbId methodology flowchart. A high level overview of the steps taken to determine miR and transposable element concurrent alignments within the human transcriptome.

OrBId analysis to identifying the targets of miRs whose TE origins have been clearly defined[23] using RepBase annotations,[42,43] this analysis was confined to the evaluation of ~16% of currently annotated human miR loci (resulting in target predictions for 191 unique human miRs). While our OrbId analysis primarily dealt with miRs predominately unexamined by the principal miR target prediction algorithms, in striking contrast to the hundreds of putative mRNAs generally predicted by the principal algorithms,[16-18,20] OrbId averaged ~8 putative mRNA targets per unique miR. While the average number of mRNAs a typical miR regulates remains poorly defined, we suggest our predictions most likely only constitute a subset of actual miR regulations (largely due to the high degree of complementarity we required for putative target interaction). However, since OrbId target sets are derived through a rationale based on molecular origin, we suggest that the OrbId putative target lists reported here likely contain a markedly higher proportion of actual endogenous miR targets than the hundreds of predicted mRNA targets obtained through less stringent algorithms. Additionally and in terms of laboratory and clinical efforts, we suggest that a manageable number of likely endogenous relationships based on a molecular rationale is in many ways advantageous to more encompassing sets of hundreds of putative targets.

Importantly, > 95% of the miRs included in our analysis have not been examined by the principal target prediction algorithms (most likely due to either their repetitive nature or their being primate specific and not conserved across species). We do find, however, that the OrBId target predictions for the few miRs in our analysis that have previously been examined are largely in agreement with more established algorithms. For example, we find ~80% of our putative miR-28 targets are contained within the principal Diana and TargetScan predictions (**Table 2**). Excitingly, four of our 13 putative miR-28 3' UTR targets have actually previously been verified experimentally.[29] Additionally, three of these experimentally verified miR-28 targets, N4BP1, E2F6 and TEX261 are expressed alongside miR-28 in blood cell lineages and have each been speculated to contribute to myeloproliferative neoplasms.[29] While experimental corroborations such as these are encouraging, the majority of our novel OrBId miR target predictions will clearly ultimately require direct experimental validation. It is tempting to speculate, however, that experimental verification of many of our miR interactions might well be forthcoming as this work represents the first time putative target sets have been reported for the majority of the 191 distinct miRs examined in this analysis thereby constituting the first real examination of potential target interactions for ~10% of all currently characterized human miRs.

In conclusion, we report here a new approach for miR target prediction that relies on TE origins. In all probability a universal description of miR target interaction has not yet been characterized because there is no universal description of miR target interaction. Complicating factors such as GU base-pairing, nucleotide editing, target secondary structure and RNA-interacting protein effects[41] make strict thermodynamic modeling largely incapable

of honing in on actual mRNA targets. Likely a closer estimation of true mRNA regulations, OrbId predicts far fewer mRNA targets per miR than existing algorithms through employing a molecular, origin-based rationale. Importantly, incorporating logical molecular cues such as target site conservation has previously been successfully exploited to circumvent the limitations of mathematical modeling alone.[16-18,20] Similarly based on genetic rationale, this work introduces a novel consideration that helps to circumvent many of the difficulties in accurate target identification.[41] We suggest that since TEs are present in multiple copies across the genome,[36] and miRs target sequences through complementary basepairing, requiring a miR target site to occur in the same TE from which a miR was initially formed represents a logical addition to miR target prediction. In contrast to the principal miR target algorithms currently utilized[16-18,20] (which rely heavily on target site conservation across species and have therefore been primarily effective at predicting targets for well conserved miRs), OrbId has been designed to predict the mRNA targets of evolutionarily younger miRs and therefore makes a strategically logical complement to existing miR target algorithms.

Materials and Methods

Retrieving miR, transposable element mRNA and genomic sequences. In 2011, Borchert et al.[23] established a connection between miRs and transposable elements (TE) providing evidence for the role of repetitive elements in miR origin. Unique TEs associated with the origins of > 200 human miRs were retrieved from the data set created from the work of Borchert et al. and used as the basis for this analysis. Single FASTA files containing the full set of human miR mature sequences were downloaded from the miR Registry housed at Sanger (www.mirbase.org).[2] Flanked genomic sequences were obtained for human miRs corresponding to genomes currently available in Ensembl (± 250 base pair flanks).[36] Unique miR accession numbers from the miR Registry were attached to the corresponding flanked genomic sequence then utilized as the origin-based TE sequence. Next, the full set of Ensembl human 5' UTR, 3' UTR and ORF sequences were compiled in and retrieved using the Biomart mining utility.[37] Of 178,375 unique human transcripts, 68,892,718 nts corresponded to 3' UTR sequence, 20,940,347 nts corresponded to 5' UTR sequence and 115,422,049 nts corresponded to ORF sequence making the average 3' UTR, 5' UTR and ORF lengths examined in this study 386, 117 and 647 nts respectively.

Correlating miR target sites with progenitor Tes. It is important to note that all alignment analyses were identically run in parallel by three independent research teams and cross examined for verification. Significant alignments between the miR and TE sequences with the human 5' UTR, 3' UTR and ORF sequences were obtained via BLAST (BLASTN 2.2.15 with -FF, -W7 flags). Beyond requiring a common molecular origin for each member of a putative miR::mRNA interaction, the majority of false positive relationships were largely avoided through requiring long, nearly perfect complementarities. Strongly agreeing with similarly stringent statistical searches for miR targets,[45] this strategy resulted in the identification of numerous long runs of perfect complementarity between putative miRs and targets and found no significant bias for that complementarity to occur near miR 5' ends or in mRNA 3' UTRs. For the miR sequences, significant alignments were strictly defined as ≥ 88% identity for ≥ 17 bp hits or 100% identity for 12_16 bps. For TE sequences, significant alignments were strictly defined as ≥ 70% identity for 50+ bp hits or ≥ 80% for bp hits less than 50. Using the proceeding search algorithm we determined alignment matches along the human 5' UTR, 3' UTR and ORF sequences between each miR and its corresponding TE. Our algorithm looked at each miR::mRNA alignment and searched for overlapping TE alignments in the same region of that transcript. If such TE alignments were found, the transcript was recorded as a target for that miR. We defined a miR as hitting the same region as its corresponding TE if either of two following criterion were satisfied: (1) The miR ending alignment position was between the TE beginning and ending alignment positions (inclusive), or (2) The miR beginning alignment position was between the TE beginning and ending alignment positions (inclusive). If at least part of the miR alignment is within the TE alignment region on a gene, then this method counted the transcript as a miR target (**Fig. S5**). Additionally, as control, we randomly generated 10 scrambled sets of matched, size appropriate miR repeat pairs to search for targets using OrBId. Importantly, we identified no putative targets for scrambled controls in the human transcriptome.

Declaration of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

This work was funded by the School of Biological Sciences, the College of Arts and Sciences at Illinois State University, USDA-NIFA-AFRI2011-67021-30114 to D.A.R. and National Institutes of Health (1R15CA137608) to E.D.L.

Supplementary Materials

Supplemental materials may be found here:
www.landesbioscience.com/journals/mge/article/21617

References

1. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell 1993; 75:843-54; PMID:8252621; http://dx.doi.org/10.1016/0092-8674(93)90529-Y
2. Kozomara A, Griffiths-Jones S. miRBase: integrating microRNA annotation and deep-sequencing data. Nucleic Acids Res 2011; 39(Database issue):D152-7; PMID:21037258; http://dx.doi.org/10.1093/nar/gkq1027
3. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. Identification of novel genes coding for small expressed RNAs. Science 2001; 294:853-8; PMID:11679670; http://dx.doi.org/10.1126/science.1064921
4. Lee RC, Ambros V. An extensive class of small RNAs in Caenorhabditis elegans. Science 2001; 294:862-4; PMID:11679672; http://dx.doi.org/10.1126/science.1065329
5. Lau NC, Lim LP, Weinstein EG, Bartel DP. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. Science 2001; 294:858-62; PMID:11679671; http://dx.doi.org/10.1126/science.1065062
6. Farazi TA, Spitzer JI, Morozov P, Tuschl T. miRNAs in human cancer. J Pathol 2011; 223:102-15; PMID:21125669; http://dx.doi.org/10.1002/path.2806
7. Hutvágner G, Zamore PD. A microRNA in a multiple-turnover RNAi enzyme complex. Science 2002; 297:2056-60; PMID:12154197; http://dx.doi.org/10.1126/science.1073827
8. Lai EC. Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. Nat Genet 2002; 30:363-4; PMID:11896390; http://dx.doi.org/10.1038/ng865
9. Zeng Y, Wagner EJ, Cullen BR. Both natural and designed micro RNAs can inhibit the expression of cognate mRNAs when expressed in human cells. Mol Cell 2002; 9:1327-33; PMID:12086629; http://dx.doi.org/10.1016/S1097-2765(02)00541-5
10. Witkos TM, Koscianska E, Krzyzosiak WJ. Practical Aspects of microRNA Target Prediction. Curr Mol Med 2011; 11:93-109; PMID:21342132; http://dx.doi.org/10.2174/156652411794859250
11. Min H, Yoon S. Got target? Computational methods for microRNA target prediction and their extension. Exp Mol Med 2010; 42:233-44; PMID:20177143; http://dx.doi.org/10.3858/emm.2010.42.4.032
12. Saito T, Saetrom P. MicroRNAs–targeting and target prediction. New Biotechnol 2010; 27:243-9; http://dx.doi.org/10.1016/j.nbt.2010.02.016
13. Thomas M, Lieberman J, Lal A. Desperately seeking microRNA targets. Nat Struct Mol Biol 2010; 17:1169-74; PMID:20924405; http://dx.doi.org/10.1038/nsmb.1921
14. Yue D, Liu H, Huang Y. Survey of Computational Algorithms for MicroRNA Target Prediction. Curr Genomics 2009; 10:478-92; PMID:20436875; http://dx.doi.org/10.2174/138920209789208219
15. Kertesz M, Iovino N, Unnerstall U, Gaul U, Segal E. The role of site accessibility in microRNA target recognition. Nat Genet 2007; 39:1278-84; PMID:17893677; http://dx.doi.org/10.1038/ng2135
16. Miranda KC, Huynh T, Tay Y, Ang YS, Tam WL, Thomson AM, et al. A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes. Cell 2006; 126:1203-17; PMID:16990141; http://dx.doi.org/10.1016/j.cell.2006.07.031
17. Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, Bouyioukos C, Mourelatos Z, et al. A combined computational-experimental approach predicts human microRNA targets. Genes Dev 2004; 18:1165-78; PMID:15131085; http://dx.doi.org/10.1101/gad.1184704
18. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human MicroRNA targets. PLoS Biol 2004; 2:e363; PMID:15502875; http://dx.doi.org/10.1371/journal.pbio.0020363
19. Lall S, Grun D, Krek A, Chen K, Wang YL, Dewey CN, et al. A genome-wide map of conserved microRNA targets in C. elegans. Curr Biol 2006; 16:460-71
20. Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, et al. Combinatorial microRNA target predictions. Nat Genet 2005; 37:495-500; PMID:15806104; http://dx.doi.org/10.1038/ng1536
21. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 2005; 120:15-20; PMID:15652477; http://dx.doi.org/10.1016/j.cell.2004.12.035
22. Smalheiser NR, Torvik VI. Mammalian microRNAs derived from genomic repeats. Trends Genet 2005; 21:322-6; PMID:15922829; http://dx.doi.org/10.1016/j.tig.2005.04.008
23. Borchert GM, Holton NW, Williams JD, Hernan WL, Bishop IP, Dembosky JA, et al. Comprehensive analysis of microRNA genomic loci identifies pervasive repetitive-element origins. Mob Genet Elements 2011; 1:8-17; PMID:22016841; http://dx.doi.org/10.4161/mge.1.1.15766
24. Borchert GM, Lanier W, Davidson BL. RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol 2006; 13:1097-101; PMID:17099701; http://dx.doi.org/10.1038/nsmb1167
25. Devor EJ, Peek AS, Lanier W, Samollow PB. Marsupial-specific microRNAs evolved from marsupial-specific transposable elements. Gene 2009; 448:187-91; PMID:19577616; http://dx.doi.org/10.1016/j.gene.2009.06.019
26. Piriyapongsa J, Jordan IK. A family of human microRNA genes from miniature inverted-repeat transposable elements. PLoS One 2007; 2:e203; PMID:17301878; http://dx.doi.org/10.1371/journal.pone.0000203
27. Yan Y, Zhang Y, Yang K, Sun Z, Fu Y, Chen X, et al. Small RNAs from MITE-derived stem-loop precursors regulate abscisic acid signaling and abiotic stress responses in rice. Plant J 2011; 65:820-8
28. Yao C, Zhao B, Li W, Li Y, Qin W, Huang B, et al. Cloning of novel repeat-associated small RNAs derived from hairpin precursors in Oryza sativa. Acta Biochim Biophys Sin (Shanghai) 2007; 39:829-34; PMID:17989873; http://dx.doi.org/10.1111/j.1745-7270.2007.00346.x
29. Girardot M, Pecquet C, Boukour S, Knoops L, Ferrant A, Vainchenker W, et al. miR-28 is a thrombopoietin receptor targeting microRNA detected in a fraction of myeloproliferative neoplasm patient platelets. Blood 2010; 116:437-45; PMID:20445018; http://dx.doi.org/10.1182/blood-2008-06-165985
30. Lee I, Ajay SS, Yook JI, Kim HS, Hong SH, Kim NH, et al. New class of microRNA targets containing simultaneous 5'-UTR and 3'-UTR interaction sites. Genome Res 2009; 19:1175-83; PMID:19336450; http://dx.doi.org/10.1101/gr.089367.108
31. Lytle JR, Yario TA, Steitz JA. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' UTR as in the 3' UTR. Proc Natl Acad Sci U S A 2007; 104:9667-72; PMID:17535905; http://dx.doi.org/10.1073/pnas.0703820104
32. Moretti F, Thermann R, Hentze MW. Mechanism of translational regulation by miR-2 from sites in the 5' untranslated region or the open reading frame. RNA 2010; 16:2493-502; PMID:20966199; http://dx.doi.org/10.1261/rna.2384610
33. Ørom UA, Nielsen FC, Lund AH. MicroRNA-10a binds the 5' UTR of ribosomal protein mRNAs and enhances their translation. Mol Cell 2008; 30:460-71; PMID:18498749; http://dx.doi.org/10.1016/j.molcel.2008.05.001
34. Schnall-Levin M, Zhao Y, Perrimon N, Berger B. Conserved microRNA targeting in Drosophila is as widespread in coding regions as in 3' UTRs. Proc Natl Acad Sci U S A 2010; 107:15751-6; PMID:20729470; http://dx.doi.org/10.1073/pnas.1006172107
35. Zhou X, Duan X, Qian J, Li F. Abundant conserved microRNA target sites in the 5'-untranslated region and coding sequence. Genetica 2009; 137:159-64; PMID:19578934; http://dx.doi.org/10.1007/s10709-009-9378-7
36. Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, et al. Ensembl 2006. Nucleic Acids Res 2006; 34(Database issue):D556-61; PMID:16381931; http://dx.doi.org/10.1093/nar/gkj133
37. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, et al. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 2005; 21:3439-40; PMID:16082012; http://dx.doi.org/10.1093/bioinformatics/bti525
38. Lehnert S, Van Loo P, Thilakarathne PJ, Marynen P, Verbeke G, Schuit FC. Evidence for co-evolution between human microRNAs and Alu-repeats. PLoS One 2009; 4:e4456; PMID:19209240; http://dx.doi.org/10.1371/journal.pone.0004456
39. Smalheiser NR, Torvik VI. Alu elements within human mRNAs are probable microRNA targets. Trends Genet 2006; 22:532-6; PMID:16914224; http://dx.doi.org/10.1016/j.tig.2006.08.007
40. Zhang R, Wang YQ, Su B. Molecular evolution of a primate-specific microRNA family. Mol Biol Evol 2008; 25:1493-502; PMID:18417486; http://dx.doi.org/10.1093/molbev/msn094
41. Smalheiser NR, Torvik VI. Complications in mammalian microRNA target prediction. Methods Mol Biol 2006; 342:115-27; PMID:16957371
42. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 2005; 110:462-7; PMID:16093699; http://dx.doi.org/10.1159/000084979
43. Kohany O, Gentles AJ, Hankus L, Jurka J. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 2006; 7:474; PMID:17064419; http://dx.doi.org/10.1186/1471-2105-7-474
44. de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet 2011; 7:e1002384; PMID:22144907; http://dx.doi.org/10.1371/journal.pgen.1002384
45. Smalheiser NR, Torvik VI. A population-based statistical approach identifies parameters characteristic of human microRNA-mRNA interactions. BMC Bioinformatics 2004; 5:139; PMID:15453917; http://dx.doi.org/10.1186/1471-2105-5-139
46. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. Nucleic Acids Res 2003; 31:439-41; PMID:12520045; http://dx.doi.org/10.1093/nar/gkg006
47. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res 2005; 33(Database issue):D121-4; PMID:15608160; http://dx.doi.org/10.1093/nar/gki081