## 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. These small microRNAs (miRs) are only ~20 nts in length (Fig. 1A) and are capable of coordinating the expressions of networks of mRNAs through complementary basepairing. Strikingly, over 1,900 unique human miRs have been cloned since the first were discovered in 2001. 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
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 target and the extent 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\textsuperscript{15} and rna22\textsuperscript{16}), the principal, most widely accepted target prediction algorithms (DIANA-microT,\textsuperscript{17} miRanda,\textsuperscript{18} PiTar,\textsuperscript{19,20} and TargetScan\textsuperscript{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,\textsuperscript{22} the molecular events responsible for the genomic formation of many miR loci from transposable element (TE) sequences have now been described\textsuperscript{22-28} (Fig. 1C). Having recently performed a series of detailed genomic analyses describing the TE origins of ~2,400 distinct miRs,\textsuperscript{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 (Origin-based Identiﬁcation 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.\textsuperscript{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\textsuperscript{17-21} and existing experimental evidence.\textsuperscript{23} 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.\textsuperscript{23} In stark contrast to the principal miR target algorithms currently utilized (which typically predict several hundred putative mRNA targets for individual miRs\textsuperscript{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).

Targets sites are generally not preferentially located in 3’UTRs. miRs have now been conclusively shown to regulate target mRNAs through interactions with 5’ untranslated
miR-301a-5p which are predicted to preferentially (> 90%) target ORF sequences (Table 2, Fig. S1).

First identified in 2003 as arising from L2B LINE elements, miR-28 and miR-151 have long regions (UTR) and open reading frame (ORF) sequences similar to 3’ UTR interactions. As such, it is somewhat surprising that most target prediction algorithms predominately screen mRNA 3’ UTRs for miR regulatory sites. In this analysis, we assessed all publicly available human mRNA sequence regardless of functional annotation. 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).

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

---

**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.

---

**Table 1.** OrbId summary

|                         | Value       |
|-------------------------|-------------|
| 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 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. “Human miRs analyzed” correspond to the full set of human miR mature sequences identified by Borchert et al. as originating from TEs and were obtained from the miR Registry miRBase.
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 | PTPNB     | 5' UTR    |        |
|           | ENSG00000108309 | RNU6C3A   | 3' UTR    |        |
|           | ENSG00000106608 | UGCP      | 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 | TCEAN2C   | 3' UTR    |        |
|           | ENSG00000004468 | CD38      | 3' UTR    |        |
|           | ENSG00000226264 | HLA-DMB   | 3' UTR    |        |
|           | ENSG00000183908 | LRRCS5    | 3' UTR    |        |
|           | ENSG00000184040 | FAM23B.1  | 3' UTR    |        |
|           | ENSG00000148483 | TMEM236   | 3' UTR    |        |
|           | ENSG00000132623 | ANKRDS    | 3' UTR    |        |
|           | ENSG00000144455 | SUMF1     | 3' UTR    |        |
|           | ENSG00000215020 | ALS91684.1| 3' UTR    |        |
|           | ENSG00000215033 | AL603965.1| 3' UTR    |        |
| hsa-mir-1254-1 | ENSG00000081760 | AACS      | 5' UTR    |        |
|           | ENSG00000167077 | ME1       | 5' UTR    |        |
|           | ENSG00000238035 | AC13803S.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.
been recognized as being related, and their numerous representative sequences across mammalia are collectively referred to as the miR-28 family. 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-1513 and baring significant pre-miR homology to both miR-28 and miR-1512 (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 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.
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 algorithms16-18,20 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 repeats23,24, 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 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.
OrBId analysis to identifying the targets of miRs whose TE origins have been clearly defined using RepBase annotations, 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 miRNAs generally predicted by the principal algorithms, OrBId averaged ~8 putative mRNA targets per unique miR. While the average number of miRNAs 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. 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. 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 make strict thermodynamic modeling largely incapable...
of honing in on actual mRNA targets. Likely a closer estimation of true mRNA regulations, Orbld 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.\textsuperscript{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.\textsuperscript{41} We suggest that since TEs are present in multiple copies across the genome,\textsuperscript{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\textsuperscript{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), Orbld 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.\textsuperscript{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).\textsuperscript{2} Flanked genomic sequences were obtained for human miRs corresponding to genomes currently available in Ensembl (± 250 base pair flanks).\textsuperscript{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.\textsuperscript{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,\textsuperscript{43} 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-AFR12011-67021-30114 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
