

# Human polymorphism at microRNAs and microRNA target sites

Matthew A. Saunders, Han Liang, and Wen-Hsiung Li<sup>†</sup>

Department of Ecology and Evolution, University of Chicago, 1101 East 57th Street, Chicago, IL 60637

Contributed by Wen-Hsiung Li, December 21, 2006 (sent for review December 4, 2006)

**MicroRNAs (miRNAs) function as endogenous translational repressors of protein-coding genes in animals by binding to target sites in the 3' UTRs of mRNAs. Because a single nucleotide change in the sequence of a target site can affect miRNA regulation, naturally occurring SNPs in target sites are candidates for functional variation that may be of interest for biomedical applications and evolutionary studies. However, little is known to date about variation among humans at miRNAs and their target sites. In this study, we analyzed publicly available SNP data in context with miRNAs and their target sites throughout the human genome, and we found a relatively low level of variation in functional regions of miRNAs, but an appreciable level of variation at target sites. Approximately 400 SNPs were found at experimentally verified target sites or predicted target sites that are otherwise evolutionarily conserved across mammals. Moreover, ~250 SNPs potentially create novel target sites for miRNAs in humans. If some variants have functional effects, they might confer phenotypic differences among humans. Although the majority of these SNPs appear to be evolving under neutrality, interestingly, some of these SNPs are found at relatively high population frequencies even in experimentally verified targets, and a few variants are associated with atypically long-range haplotypes that may have been subject to recent positive selection.**

human evolution | positive selection | single-nucleotide polymorphism

**M**icroRNAs (miRNAs) function as important posttranscriptional regulators of mRNA expression by binding to the 3' UTR and repressing translation (1, 2). Indeed, a large proportion of protein-coding genes appear to be regulated by miRNAs, suggesting that miRNAs have a critical role in affecting a variety of biological functions (3–6).

A miRNA gene is transcribed and processed initially into a precursor miRNA (pre-miRNA) that is ~100 bp in length and forms a stem-loop foldback structure (7–9). The pre-miRNA is further processed into a mature miRNA (MIR) that is ~22 bp long and binds to a specific target site on an mRNA to exert posttranscriptional repression. The critical region for MIR binding in animals is the “seed” region (nucleotides 2–7 from the 5' end of the MIR), which most often binds to a target site in the 3' UTR of the given mRNA by perfect Watson–Crick complementarity (3, 10). Recent studies have demonstrated that a single mutation in the match of the miRNA seed to its target site can abolish repression (5). Different computational approaches have been developed to predict miRNA target sites throughout the genome (reviewed in ref. 11), and a small portion of these predicted target sites have been experimentally validated to show relatively high accuracy for target site prediction (12).

Because the prevalence and importance of miRNAs in animals has been recognized only in recent years, few studies have described naturally occurring human polymorphisms associated with miRNAs and their target sites (13, 14). Due to the stringent recognition requirement between the miRNA seed region and its target, a naturally occurring SNP may have significant functional implications for MIR binding and posttranscriptional regulation. For example, a recent study has shown that a SNP may affect organismal phenotype by altering a miRNA target site, leading to

a significant alteration of protein expression (15). Given the wealth of data that is currently available in databases for human SNPs, we can begin to identify naturally occurring variation associated with miRNAs and their targets using an *in silico* approach. SNPs in critical components of the miRNA system may have important phenotypic consequences, with implications for both evolutionary studies and biomedical research.

In this study we conducted a bioinformatic genome-wide survey of human SNPs in miRNA target sites and in miRNAs themselves, and found an appreciable level of variation within predicted miRNA target sites as well as target sites that have been experimentally verified for posttranscriptional regulation of mRNAs. In addition, we searched for SNPs that would potentially affect novel target sites in humans. We speculate that some of these variations may have functional effects, and we show that some SNPs are associated with suggestive signatures of recent positive selection.

## Results

**Low Polymorphism in miRNA Genes.** We identified SNPs in human miRNA genes by querying the Single-Nucleotide Polymorphism Database (dbSNP) at the genomic coordinates of 474 pre-miRNAs. We identified 65 SNPs (including indel polymorphisms) in 49 pre-miRNAs, thus exhibiting a SNP density of ~1.3 SNPs per kb. For comparison, we also queried dbSNP for the flanking regions around the pre-miRNAs. As the regions flanking miRNAs are most often intergenic regions, likely with weak or no functional constraint, these regions exhibit a higher SNP density of ~3 SNPs per kb (Fig. 1A).

The pre-miRNA is composed of different domains with different functional significance. To gain insight into the potential functional importance of the identified polymorphisms, we mapped the SNPs to five different domains of the pre-miRNAs: (i) the seed region, (ii) the mature region excluding the seed region (MIR<sup>Δseed</sup>), (iii) the stem region complementary to the MIR (MIR\*), (iv) the stem region that is neither the MIR nor MIR\*, and (v) the loop region (Fig. 1B). Only three miRNAs (*hsa-mir-125a*, *hsa-mir-627*, and *hsa-mir-662*) have SNPs within the seed region (Fig. 1B). No population frequency information is currently available for these SNPs (rs1297533, rs2620381, and rs9745376), but it is likely that these are not common polymorphisms because most common polymorphisms have been sampled in HapMap (16). Thus, these variants would not be of significant population genetic importance. Furthermore, *hsa-mir-125a* shares an identical seed sequence with

Author contributions: M.A.S. and H.L. contributed equally to this work; M.A.S., H.L., and W.-H.L. designed research; M.A.S., H.L., and W.-H.L. performed research; M.A.S. and H.L. analyzed data; and M.A.S., H.L., and W.-H.L. wrote the paper.

The authors declare no conflict of interest.

Abbreviations: miRNA, microRNA; pre-miRNA, precursor miRNA; MIR, mature miRNA; NBR, novel biologically relevant; iHS, integrated haplotype score; LRH, long-range haplotype; EHH, extended haplotype homozygosity.

<sup>†</sup>To whom correspondence should be addressed. E-mail: whli@uchicago.edu.

This article contains supporting information online at [www.pnas.org/cgi/content/full/0611347104/DC1](http://www.pnas.org/cgi/content/full/0611347104/DC1).

© 2007 by The National Academy of Sciences of the USA









and some SNPs are at relatively high population frequencies. The discovery of SNPs in experimentally verified targets provides the most suggestive evidence for functionality of these variants, and it follows that miRNA target polymorphism might underlie human variation at the phenotypic level. This is an exciting prospect, because fine-tuning of expression levels can be mediated by miRNAs (6) and adaptation by regulatory effects is thought to be important for evolution in general (21). To date, most evolutionary consideration of regulatory regions has focused on cis-promoter elements and transcription factors.

Clearly, a majority of the SNPs in predicted targets are likely neutral or in a mutation–selection balance. However, it is plausible that some polymorphisms in target sites confer a selective advantage by changing the expression level of a given target protein. Here we studied polymorphic miRNA target sites as candidates for targets of recent positive selection in humans by using computational tests based on LRH analyses. These tests were performed in the context of a single population and thus are suitable for detecting recent selection that is population-specific or that is shared among human populations. In contrast, inferences based solely on measures of population subdivision (e.g.,  $F_{ST}$ ) are useful only for cases of strong population-specific selection. It is also noteworthy that statistical tests based on LRHs are suitable only for detecting selection that has occurred recently, within the past  $\approx 30,000$  years (22), whereas selection that has occurred earlier in evolutionary timescales will not be detected.

We identified three SNPs in miRNA targets (rs7284767 in *TUG1*, rs11755 in *ARPC5*, and rs1042538 in *IQGAP1*) that are associated with unusual patterns of LRH conservation in the human genome. Noteworthy is the SNP rs1042538 in *IQGAP1*, which disrupts a miRNA target site sequence. Although the pattern associated with this SNP shows only suggestive significance of an LRH test in the YRI panel, the pattern may be biologically important because this gene has been experimentally verified to be affected by *hsa-mir-124*, providing further support that this variant may in fact be functional. Future functional assays on the SNP and detailed population genetic analyses will be required to determine the biological significance of this SNP and other SNPs of interest in this study.

Most often, even a functional polymorphism with a phenotypic effect will not have been subjected to positive selection, but it may still represent variation of anthropological or medical interest. Interestingly, our study shows that the SNP rs17168525 disrupts an experimentally verified target site in the *MTPN* 3' UTR. The derived allele frequency for this SNP is relatively high ( $q = 0.19$ ) in the ASN population, and, importantly, direct empirical evidence shows that this target site effects down-regulation of *MTPN* *in vivo* (6, 12). Although this SNP does not show evidence of recent selection, it may be of interest for future studies to examine potential effects related to *MTPN* expression levels.

In summary, we have shown that many miRNA target sequences are polymorphic and often these polymorphisms are at relatively high frequencies in different human populations. Experimental assays will determine the functional effects of these SNPs. Regardless of their fitness effect (i.e., deleterious, neutral, or advantageous), those variants that may indeed affect expression will shed light on mechanisms for phenotypic variability among humans. Future studies may also explore possible evolutionary implications of fixed differences between closely related species (e.g., human and chimpanzee) at miRNA target sites.

## Methods

**Identifying SNPs in miRNA Genes.** We obtained the genomic coordinates (hg 18; National Center for Biotechnology Information build 36) of all available human pre-miRNA genes ( $n = 474$ ) from the miRbase database (Release 9.0) (23) and identified SNPs (dbSNP build 126) within the miRNA genes by using the application programming interface tools in the ENSEMBL

database (24) (see below). SNPs in pre-miRNAs were further mapped to their locations within the miRNA secondary structure and were classified into five domains: seed, MIR<sup>Δseed</sup>, MIR\*, stem, and loop. For comparison, we also identified SNPs in regions flanking each miRNA gene spanning the windows with the same size of the given pre-miRNA, and we calculated the respective SNP density (excluding indels).

**Prediction of miRNA Target Sites.** We obtained 3' UTRs of the human genome from the Mammalian Genome Collection and RefSeq databases by using the University of California, Santa Cruz, genome browser. Only transcripts with a unique genomic location in the main assembly (National Center for Biotechnology Information build 36) were included in the analysis. Based on these well annotated human transcripts ( $\approx 45,500$ ), we defined nonredundant 3' UTR sequences for  $\approx 18,000$  human genes.

To predict miRNA target sites, we first constructed a four-species (human, mouse, rat, and dog) multiple alignment from the *MultiZ* 17 alignments (25) for the human 3' UTR sequences. Then we used the TargetScanS algorithm (3) to predict miRNA target sites for 73 conserved microRNA families (17). Briefly, the algorithm searches for conserved 7-mers (or 8-mers), including a 6-nt match to the miRNA seed (nucleotides 2–7) plus an additional anchor nucleotide. Throughout the paper, we refer to these 7- or 8-mers as (miRNA) target sites. We predicted  $\approx 29,000$  target sites in the human 3' UTRs. To calculate SNP density (to be distinguished from nucleotide diversity estimators of  $\theta = 4N_e\mu$ ) for target sites (excluding indels), we only considered a set of nonoverlapping target sites ( $\approx 25,000$ ). SNP density was also calculated for the flanking regions (both upstream and downstream). Analysis of each flanking region includes three successive, nonoverlapping windows that have the same length as the given miRNA target site.

**miRNA Target Sites with Experimental Evidence.** We downloaded all experimentally verified human target genes (including both translationally repressed mRNAs and down-regulated/cleaved mRNAs) from the TarBase database (12). When annotation for a miRNA target site was not available, we searched for the 6-nt seed-matches for the cognate miRNA in the 3' UTR and defined each target site as an 8-mer that includes a 6-nt seed match and two flanking nucleotides. Together, 706 target sites (representing 388 genes) with experimental evidence were included in our analysis.

**Prediction of NBR Target Sites Revealed by SNPs.** A SNP may create (or disrupt) a sequence in humans that corresponds to a miRNA target site. However, this newly affected target sequence is not biologically relevant unless the cognate miRNA is spatially and temporally coexpressed with the target mRNA. We refer to such a target site as a NBR target site. To identify NBR sites, the human SNP data onto 3' UTR regions of the human reference genome, and for each of the 73 miRNA families, we identified target site sequences that otherwise would not be revealed in the human reference genome. To further identify a set of target sites with coexpression of the cognate miRNA in the same spatial and temporal domain (i.e., NBR targets), we incorporated mRNA expression data (26) (downloaded from the National Center for Biotechnology Information GEO database) and miRNA expression data (27). mRNA expression data were then processed as described elsewhere (17) to calculate the expression level of target genes. We conservatively defined “high” expression in a given tissue with a threshold of 2,000 for a mRNA and a threshold of 5,000 for a miRNA. Both thresholds are 10 times higher than the background level in each data set. Together, there are five tissues shared between the two expression profiles (i.e., brain, liver, thymus, testes, and placenta), and we predicted 257 NBR target sites in 3' UTRs whose genes are coexpressed with the cognate miRNA in at least one of the five tissues.

Depending on the ancestral state of an allele, variations at these NBR target sites can either create or disrupt a target site in humans. To elucidate evolutionary changes at NBR sites, we retrieved the orthologous chimpanzee nucleotides from the *MultiZ* 17 alignment whenever available. We found that for the SNPs with an available orthologous chimpanzee position, >85% of the nucleotides in human reference genome have the same nucleotide in the chimpanzee genome, suggesting that in most cases, the alternative (nonreference) human allele did create a miRNA target site (SI Table 3).

**Retrieval of SNP information from dbSNP.** To access dbSNP information, we implemented perl application programming interfaces from ENSEMBL ("Variation" modules version 39.36a) by querying given human genome coordinates. For each SNP we retrieved the following information: allele states, population name/origin, population frequency, and availability in HapMap (version 21). For analyses, SNPs were binned into four categories based on population frequency: (i) SNPs that have no population frequency information available in dbSNP (category "n/a"), (ii) SNPs that have been genotyped in at least one population and are monomorphic ( $q = 0$ ) in the population(s) surveyed (category 0), (iii) SNPs that have been genotyped in at least one population and show a low polymorphic frequency ( $0 < q < 0.10$ ) in the population(s) surveyed (category 1), and (iv) SNPs that have been genotyped in at least one population and show an intermediate/high polymorphic frequency ( $q \geq 0.10$ ) in at least one population (category 2). For SNPs with different frequencies in different populations, categorization was mutually exclusive and each SNP was classified based on the maximal frequency of the minor allele in any available population.

**Tests of Positive Selection Based on LRHs.** For tests of selection, we considered SNPs in computationally predicted targets and in NBR target sites in frequency category 2 (see above) as "candidate SNPs" if they were available in HapMap (version 21). For these candidate SNPs, we performed analyses to test for signatures of recent selection by using phased LRH data. Analyses were performed on candidate SNPs only for the given population (i.e., YRI, Yorubans

from Ibadan, Nigeria; CEU, CEPH from Utah; ASN, combined Japanese from Tokyo, Japan, and Han Chinese from Beijing, China) showing a high frequency. We restricted our analyses to SNPs with minor allele frequencies of  $q \geq 0.10$  to provide the power to distinguish between extended haplotypes conserved due to young age vs. positive selection. For SNPs in experimentally verified targets we performed tests for all populations with any polymorphic frequency.

We obtained the iHS (18) for each candidate SNP via the program Haplotter (<http://hg-wen.uchicago.edu/selection/haplotter.htm>).

To study the LRH data in a complementary fashion, we performed the "LRH test" (19) by using the software SWEEP according to standard documentation (December 2005 version). For a given population we retrieved the phased haplotypes of the respective HapMap panel spanning 600 kb centered on each candidate SNP. By using the phased haplotypes, nonoverlapping cores of haplotypes were defined as restricted to a maximum of 10 contiguous SNPs. We determined whether the core haplotype with the SNP of interest displayed an unusual level of EHH relative to other haplotypes in the core and relative to a population-specific empirical distribution of core EHH values given similar population frequencies and genetic distance from the SNP of interest. Statistical significance for the LRH test results was determined (by following standard software documentation) relative to an empirical distribution of EHH values of core haplotypes divided into 20 frequency bins by using data from 10 windows (spanning 600 kb each) of anonymous locations across the genome. Analyses were performed after adjusting for similar SNP densities across all regions. Given the well known weak power of detecting positive selection in humans, raw  $P$  values were reported without multiple comparison correction.

We thank L.-C. Hsieh and L. L. Chen for technical assistance and discussions and H. Kaessmann, M. W. Nachman, and R. Adkins for valuable comments. This work was supported in part by a United Negro College Fund–Merck Science Initiative postdoctoral fellowship (to M.A.S.) and by grants from the National Institutes of Health (to W.H.L.).

- Ambros V (2004) *Nature* 431:350–355.
- Bartel DP (2004) *Cell* 116:281–297.
- Lewis BP, Burge CB, Bartel DP (2005) *Cell* 120:15–20.
- Lim LP, Lau NC, Garrett-Engle P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM (2005) *Nature* 433:769–773.
- Brennecke J, Stark A, Russell RB, Cohen SM (2005) *PLoS Biol* 3:404–418.
- Krek A, Grun D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, Rajewsky N (2005) *Nat Genet* 37:495–500.
- Cai XZ, Hagedorn CH, Cullen BR (2004) *RNA* 10:1957–1966.
- Smalheiser NR (2003) *Genome Biol* 4:403.
- Lee Y, Ahn C, Han JJ, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S, Kim VN (2003) *Nature* 425:415–419.
- Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB (2003) *Cell* 115:787–798.
- Rajewsky N (2006) *Nat Genet* 38:S8–S13.
- Sethupathy P, Corda B, Hatzigeorgiou AG (2006) *RNA* 12:192–197.
- Iwai N, Naraba H (2005) *Biochem Biophys Res Commun* 331:1439–1444.
- Chen K, Rajewsky N (2006) *Nat Genet* 38:1452–1456.
- Clop A, Marq C, Takeda H, Pirottin D, Tordo X, Bibe B, Bouix J, Caiment F, Elsen JM, Eycheffe F, et al. (2006) *Nat Genet* 38:813–818.
- International HapMap Consortium (2005) *Nature* 437:1299–1320.
- Farh KKH, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, Burge CB, Bartel DP (2005) *Science* 310:1817–1821.
- Voight BF, Kudaravalli S, Wen XQ, Pritchard JK (2006) *PLoS Biol* 4:446–458.
- Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al. (2002) *Nature* 419:832–837.
- Saunders MA, Good JM, Lawrence EC, Ferrell RE, Li W-H (2006) *Am J Hum Genet* 79:1089–1097.
- King MC, Wilson AC (1975) *Science* 188:107–116.
- Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Vailly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES (2006) *Science* 312:1614–1620.
- Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ (2006) *Nucleic Acids Res* 34:D140–D144.
- Birney E, Andrews D, Caccamo M, Chen Y, Clarke L, Coates G, Cox T, Cunningham F, Curwen V, Cutts T, et al. (2006) *Nucleic Acids Res* 34:D556–D561.
- Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, et al. (2004) *Genome Res* 14:708–715.
- Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al. (2004) *Proc Natl Acad Sci USA* 101:6062–6067.
- Barad O, Meiri E, Avnion A, Aharonov R, Barzilay A, Bentwich I, Einav U, Glad S, Hurban P, Karov Y, et al. (2004) *Genome Res* 14:2486–2494.