Human polymorphism at microRNAs and microRNA target sites

Matthew A. Saunders, Han Liang, and Wen-Hsiung Li†

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

MicroRNAs (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. 1*A*).

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 ($MIR^{\Delta seed}$), (*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. 1*B*). Only three miRNAs (*hsa-mir-125a*, *hsa-mir-627*, and *hsa-mir-662*) have SNPs within the seed region (Fig. 1*B*). 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.

†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.

© 2007 by The National Academy of Sciences of the USA

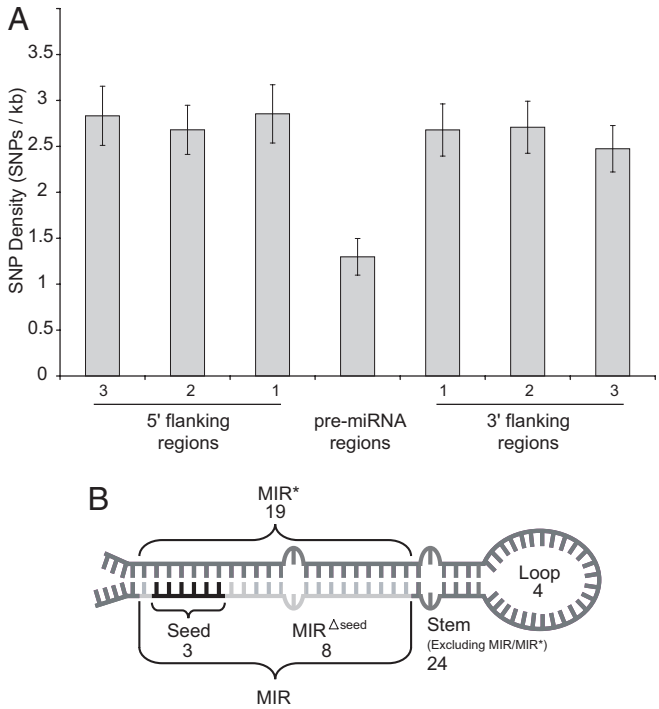

Fig. 1. SNPs in human pre-miRNAs. (*A*) SNP density in validated human pre-miRNAs and flanking regions. Flanking regions 1–3 represent successive, nonoverlapping windows of ≈100 nt (equal to the size of the given pre-miRNA) located immediately adjacent to the pre-miRNA. Error bars represent the standard deviation of the mean value. (*B*) Distribution of human SNPs across the secondary structure of pre-miRNAs. The pre-miRNA is apportioned in the following regions: seed (black), MIR$^{\Delta seed}$ (light shaded), MIR*, stem (neither MIR or MIR*), and loop. The number of SNPs identified in dbSNP for each region in human pre-miRNAs is denoted for each region.

hsa-mir-125b-1 and *hsa-mir-125b-2*, which likely creates some functional redundancy among these miRNAs.

Overall, ≈90% of human pre-miRNAs have no reported polymorphisms, and most observed polymorphisms are not within the seed region, demonstrating strong selective constraint on human pre-miRNAs.

Many Target Sites are Disrupted by High Frequency SNPs in Humans. We investigated the occurrence of human SNPs in miRNA target sites. Depending on the combinatorial effect of different miRNAs that regulate a given mRNA, a SNP that disrupts a true target site may have biological implications. To scan for SNPs in miRNA target sites, we compiled a list of target sites throughout the human genome from two sources (see *Methods*): (*i*) computational prediction of miRNA target sites (representing mRNA seed matches) in human 3′ UTRs that are conserved across four mammalian taxa (human, mouse, rat, and dog) and (*ii*) based on a list of targets with experimental evidence for control by miRNAs from the TarBase database (12).

We queried dbSNP at the coordinates of ≈29,000 computationally predicted sites and 706 sites with experimental supporting evidence (244 target sites are shared between both lists). From these data, we identified 383 SNPs (including indels) that disrupt computationally predicted sites [supporting information (SI) Table 2] and 17 SNPs (including indels) that disrupt target sites with experimental evidence (Table 1). These SNPs affect ≈350 protein-coding genes. The average SNP density in computationally predicted target sites is 1.9 SNPs per kb, whereas the density in flanking regions is ≈2.7 SNPs per kb (Fig. 2), suggesting overall the action of purifying selection on target sites. This finding is consistent with results of an independent comprehensive study using a similar methodology that has also demonstrated strong negative selection on miRNA target sites (14).

In principle, SNPs with a functional effect may be of negligible population genetics importance if they are at a low frequency. To gain insight into the frequencies of the alleles, we retrieved population frequency information for the given SNPs that disrupt target sites. Many SNPs in dbSNP lack any population frequency information; however, 196 SNPs in target sites have frequency information for various panels of human populations (representing panels from HapMap, Perlegen, and other genotyping panels). Many types of ascertainment bias affect the availability of population frequency information for SNPs in dbSNP. For this reason, it

Table 1. SNPs in experimentally verified human miRNA targets

Target gene	miRNA	Target sequence	SNP	Frequency category	HapMap panel frequency, CEU/HCB/JPT/YRI	Evidence	Conservation
LDLRAP1	mir-124	G[T/*G*]GCCTTT	rs11583293	n/a	—	a	M, R, D, O
SLC16A9	mir-1/-206	TCATTCC[G/*A*]	rs2893808	**0**	0.0/0.0/0.02/0.0	a	M, D, O
SDC4	mir-1/-206	TCATT[C/−]CT	rs11475489	n/a	—	a	M, R, D, O
CD24	mir-93/-302/-372/-373	AGCACT[T/*A*]A	rs4030414	n/a	—	a	None
ATP6V0E	mir-124	ATGCCT[T/*C*]A	rs11539178	n/a	—	a, b	None
POLR2K	mir-1/-206	ACATT[C/*T*]CA	rs11555067	n/a	—	a	None
DVL2	mir-124	CT[G/*A*]CCTTT	rs1054280	n/a	—	a	M, R
IQGAP1	mir-124	CTGCCTT[T/*A*]	rs1042538	**2**	0.10/0.41/0.42/0.09	a, c	D, O
MYH9	mir-124	GTGC[C/*T*]TTA	rs8226	n/a	—	a	M, R, D, O, C
MKRN1	mir-93/-302/-372/-373	G[G/*C*]CACTTT	rs17620927	**0**	0.0/0.0/0.0/0.0	a	M, R, D, O
EZH2	mir-101	GTACTG[T/*G*]A	rs8829	n/a	—	c	M, R, D, O, C, X
TTC7A	mir-124	GTGCCT[T/*C*]T	rs28635788	n/a	—	a	M, R, D
TTC7A	mir124	GTGCCTT[T/*C*]	rs10196117	n/a	—	a	M, R, D
CAV1	mir-124	AT[G/*A*]CCTTA	rs11553391	n/a	—	a	D
ACAA2	mir-124	TTG[C/*G*]CTTA	rs7233791	**2**	0.05/0.11/0.14/0.19	a, b	M, R, O
KLHDC5	mir-1/-206	A[C/*T*]ATTCCC	rs1050288	2	—	a	M
MTPN	let-7/mir-98	GTA[C/*T*]CTCA	rs17168525	**2**	0.01/0.19/—/0.0	c	M, R, D, O

The target sequence column shows the polymorphic sites in brackets with nonreference alleles in italics. Also shown are values for the population frequency category for minor allele as defined in *Methods*. Bold text indicates availability in HapMap. The types of experimental evidence are microarray (a), RT-PCR (b), Luciferase assay (c). For the conservation information shown, the abbreviations are as follows: M, mouse; R, rat; D, dog; O, opossum; C, chicken; X, *X. tropicalis*.

is difficult to glean meaningful inferences regarding the true population frequency spectrum of alleles (*sensu stricto*) from these data. However, interestingly, 57 SNPs in the predicted target sites and four of the SNPs in experimentally verified targets are at relatively high minor allele frequencies of $q \geq 0.10$ (category 2; see *Methods*) in at least one population (Fig. 3).

It has been suggested that the false-discovery rate for computationally predicted target sites is ≈50% (3, 11, 17), which implies that some of these SNPs likely represent true target sites that harbor relatively high-frequency allelic variants with potential functional significance. Of particular interest are SNPs found within experimentally verified target sites (Table 1). For example, rs17168525 resides within the target site for *let-7* within the 3′ UTR of the gene *MTPN* (myotrophin). The human consensus sequence for this target site has been experimentally verified via a luciferase reporter assay to effect *MTPN* protein levels (6). The given SNP is found at a frequency of 0.19 in the HapMap panel from East Asia (ASN). Another SNP, rs1042538, disrupts a target site for *hsa-mir-124* within the 3′ UTR of the gene *IQGAP1* (IQ motif-containing GTPase-activating protein 1). This SNP was found at frequencies of 0.40, 0.10, and 0.10 for the HapMap panels of Asia (ASN), Europe (CEU), and Africa (YRI), respectively. Experimental evidence based on microarray data and a luciferase reporter assay (4) suggests that *IQGAP1* transcript levels are controlled by *hsa-mir-124*.

Novel Biologically Relevant (NBR) Target Sites. Alternative alleles for a human SNP may create (or disrupt) a target site that is not evolutionarily conserved. For example, a SNP in the 3′ UTR may create a sequence match to the seed of a miRNA that previously was not associated with the given mRNA. However, the presence of such a target at the primary sequence level of a 3′ UTR is not biologically relevant unless the appropriate cognate miRNA is coexpressed (temporally and spatially) with the given mRNA. Therefore, we predicted NBR target sites in the human genome by considering naturally occurring human SNPs in 3′ UTRs, in context with coexpression information from microarray data for miRNAs and protein coding genes (see *Methods*).

Using stringent thresholds for high expression levels at both miRNAs and mRNAs we predicted 257 NBR target sites in 209 human genes (SI Table 3). A large proportion of the SNPs associated with NBR sites are at a high frequency (category 2; data not shown). Although this observation may be attributed to a high false-discovery rate in our computational prediction approach that does not rely on evolutionary conservation, some of these NBR sites may bear biological importance (see below).

Tests for Recent Positive Selection on miRNA Target Sites. A polymorphism that leads to a functional difference may be subject to selection. Recent positive selection that acts on an advantageous SNP may leave a signature at the population genomic level of exceptionally long range linkage disequilibrium with neighboring SNPs. To test for signatures of recent positive selection on SNPs in miRNA target sites, we considered candidate SNPs (see *Methods*) among our predicted conserved targets, experimentally verified targets and NBR targets in context with long-range haplotype (LRH) analyses.

The integrated haplotype score (iHS) has recently been developed as a genomically standardized measure for conservation of LRHs associated with a given SNP in a population (18). Exceptionally high values of $|\text{iHS}|$ suggest a signature of recent positive selection for the given haplotype, and this test is most powerful for selected alleles with frequencies of >0.5. We obtained iHS scores for all available candidate SNPs (see *Methods*). The overall iHS distribution for the candidate SNPs is not significantly different from the genomic empirical distribution, and the majority of $|\text{iHS}|$ values are <2.0, suggesting that these SNPs have not recently experienced strong positive selection (Fig. 4). However, two SNPs, rs7284767 (in YRI) and rs11755 (in ASN), exhibit significantly high iHS values of +3.23 and −2.65 (within approximately the top 1% of genome-wide outliers), respectively.

SNP rs7284767 (G/A) is located in a NBR target on the gene *TUG1* (taurine up-regulated gene 1), where the human reference allele (G) is at a frequency of 0.44 in YRI (and 0.37 in ASN). The alternative allele (A) creates a NBR target site for *hsa-mir-20* in humans. The positive iHS seen for this SNP indicates that the allele state seen in the chimpanzee reference genome (G) is associated with a conserved LRH in humans. However, it remains unclear which allele is in fact the human ancestral state.

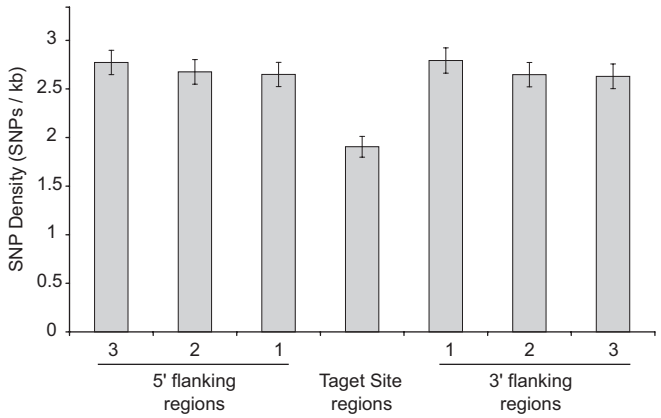

Fig. 2. SNP density within computationally predicted target sites and flanking regions. Flanking regions 1–3 represent successive, nonoverlapping windows of 7–8 nt (equal to the size of the given target) located immediately adjacent to the target. Error bars represent the standard deviation of the mean value.

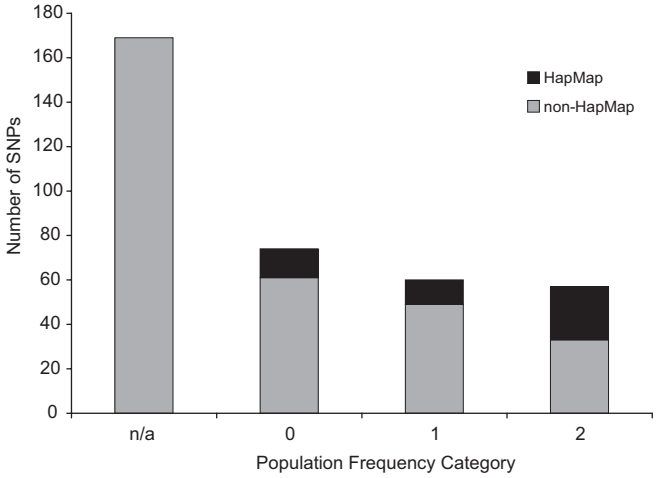

Fig. 3. Frequencies of minor alleles for SNPs in computationally predicted target sites. SNPs with population frequency estimates from any population in dbSNP were categorized into four categories based on minor allele frequency information: (*i*) SNPs without frequency information available for any population (category n/a), (*ii*) SNPs that are monomorphic in all populations genotyped (i.e., minor allele frequency $q = 0$) (category 0), (*iii*) SNPs that have been genotyped in at least one population and show a maximal frequency of $0 < q < 0.10$ in the population(s) surveyed (category 1), and (*iv*) SNPs that have been genotyped in at least one population and show a frequency of $q \geq 0.10$ in at least one population (category 2). Because a given SNP may have different frequencies in different populations, a SNP was categorized based on its highest frequency in any given population.

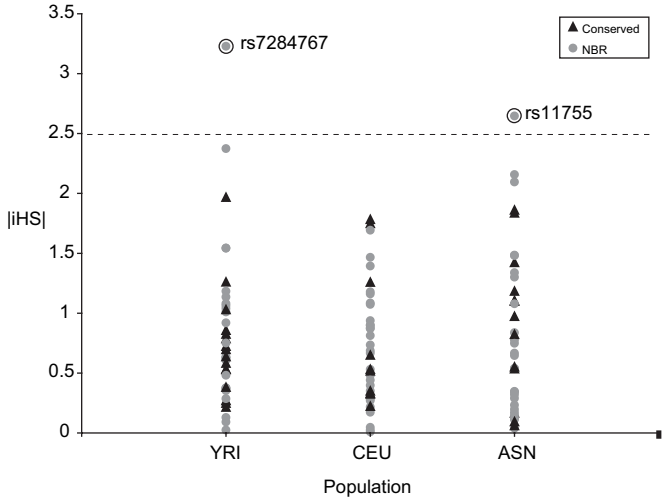

Fig. 4. Distribution of iHS estimates for SNPs in miRNA target sites. |iHS| values (18) were obtained for category 2 SNPs in computationally predicted sites (black triangles) and in NBR sites (gray circles) for each of the three HapMap populations (i.e., YRI, CEU, or ASN). Note that iHS values were not available for some SNPs, including those found in experimentally verified targets.

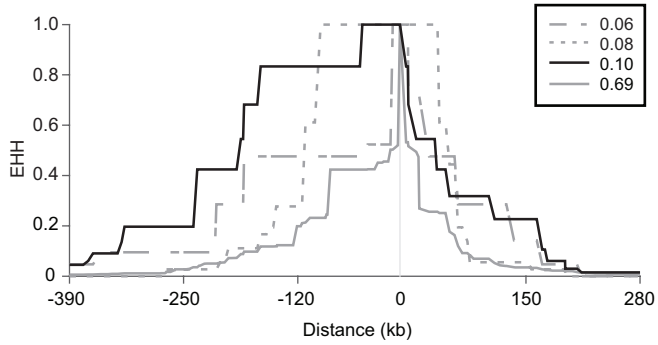

Fig. 5. EHH at various genetic distances for core haplotypes around rs1042538 in the YRI panel. The core haplotype bearing the derived allele at rs1042538 is marked in bold. All other haplotypes from the defined core are marked in gray and hatched lines. The population frequency of each core haplotype is denoted in box.

Although the allele in the human reference genome is the same as the orthologous nucleotide in the chimpanzee, mouse, and rat reference genomes, the human variant allele (A) is the same as the orthologous nucleotide in the macaque and other mammalian reference genomes. Thus, it is possible that the human ancestral state was in fact an A (with a miRNA target site, as seen in macaque), and the alternative allele G, which disrupts the target, has increased in frequency in humans and chimpanzees.

The other SNP found with a high iHS value, rs11755 (A/G), also is associated with a NBR target and is found in the gene *ARPC5* (actin related protein 2/3 complex, subunit 5) with the reference allele (A) at a frequency of 0.81 in ASN (and 0.43 and 0.17 in CEU and YRI, respectively). The alternative allele (G) creates a NBR target site for miRNA *hsa-mir-34ac*. This allele (G) is likely the ancestral state, because it is seen in all other mammalian reference genomes, and the negative value of the iHS indicates that the human derived allele (A) is associated with a conserved LRH. Therefore, it seems that this NBR target site is decreasing in frequency in some human populations. Interestingly, this SNP shows a high degree of population subdivision between ASN and YRI (F_{ST} = 0.40).

To perform a complementary test for recent positive selection, we used the LRH test (19). This test has been shown to be suitable to detect positive selection based on the extended haplotype homozygosity (EHH) parameter for selected alleles at relatively low frequencies (q = 0.1–0.2) (19, 20). We performed a LRH test for each candidate SNP in the study to determine whether it resides on a haplotype with unusually high EHH. One SNP (rs1042538 in YRI; q = 0.10) resides on a haplotype that shows suggestive evidence of recent positive selection. This SNP resides on a core haplotype that has a relatively high EHH value (0.81) at a distance of ≈0.05 cM, that is suggestive to be atypical among other haplotypes in the core (Fig. 5), and in an empirical genomic distribution (P = 0.08; data not shown). (The iHS value for this SNP was not available for comparison.) Interestingly, this SNP disrupts a target site found in the gene *IQGAP1*, which is experimentally validated to be affected by the given miRNA. As mentioned, the minor allele frequency of this SNP also is relatively high in other populations (q = 0.10 and 0.41 for CEU and ASN, respectively).

Discussion

In recent years, miRNAs have been recognized as important posttranscriptional regulators of gene expression in mammals. To date, >470 miRNAs have been identified in humans, and miRNAs are proposed to influence gene expression of >30% of protein-coding genes (3, 6). As miRNAs have a wide-spread effect on mRNA transcripts, a description of human natural variation associated with the miRNA system is warranted for understanding its functional and evolutionary significance. Recent development of methods to identify miRNAs and their binding sites (i.e., targets) along with the availability of comprehensive genomic databases of SNPs provides an unprecedented opportunity to explore human evolution at miRNAs and their targets. In this study we describe genome-wide patterns of human polymorphism in miRNAs and miRNA target sites found in 3′ UTRs of mRNAs.

Our survey suggests that the occurrence of SNPs in pre-miRNA sequences is relatively rare. Only ≈10% of human pre-miRNAs have documented SNPs, and <1% (n = 3) of miRNAs have SNPs in the functional seed region. The population frequencies of the SNPs in seed regions are not yet available in dbSNP; however, it is likely that they are at low frequency and are thus of negligible population genetic importance. Changes that occur in other parts of the pre-miRNA, such as the stem and loop regions, may be under less stringent functional constraint, consistent with the observation that such SNPs are found at higher population frequencies (data not shown). Any given MIR may regulate hundreds of different targets at different spatial-temporal settings (4, 17). Therefore, functional mutations in miRNAs, as "top-level" regulators, are unlikely to be tolerated and would be removed by purifying selection. A recent resequencing study surveyed 173 human pre-miRNAs (a subset of those studied here) in a panel of 96 individuals and uncovered a total of only 10 SNPs in the pre-miRNAs (6%), none of which were in the seed regions (13). These results are consistent with our bioinformatic results.

Many human mRNA transcripts have target sequences in the 3′ UTR to which miRNAs bind and exert posttranscriptional repression. The inferred accuracy of the computational methods used here for target site prediction suggests that at least 50% of the target sites considered here are real; however, without functional validation it is difficult to determine what proportion of our discovered SNPs lie in true targets. Nonetheless, it is likely that some of these targets are real, and given the small size of the target sequence a single base change would likely have functional implications. In fact, experimental evidence is available for >700 targets, and we show that 2.4% of these targets are polymorphic,

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 ≈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, $MIR^{\Delta seed}$, 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 (≈45,500), we defined nonredundant 3′ UTR sequences for ≈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 ≈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 (≈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/cleavaged 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 a 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.).

1. Ambros V (2004) *Nature* 431:350–355.
2. Bartel DP (2004) *Cell* 116:281–297.
3. Lewis BP, Burge CB, Bartel DP (2005) *Cell* 120:15–20.
4. Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM (2005) *Nature* 433:769–773.
5. Brennecke J, Stark A, Russell RB, Cohen SM (2005) *PLoS Biol* 3:404–418.
6. 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.
7. Cai XZ, Hagedorn CH, Cullen BR (2004) *RNA* 10:1957–1966.
8. Smalheiser NR (2003) *Genome Biol* 4:403.
9. 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.
10. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB (2003) *Cell* 115:787–798.
11. Rajewsky N (2006) *Nat Genet* 38:S8–S13.
12. Sethupathy P, Corda B, Hatzigeorgiou AG (2006) *RNA* 12:192–197.
13. Iwai N, Naraba H (2005) *Biochem Biophys Res Commun* 331:1439–1444.
14. Chen K, Rajewsky N (2006) *Nat Genet* 38:1452–1456.
15. Clop A, Marcq F, Takeda H, Pirottin D, Tordoir X, Bibe B, Bouix J, Caiment F, Elsen JM, Eychenne F, *et al.* (2006) *Nat Genet* 38:813–818.
16. International HapMap Consortium (2005) *Nature* 437:1299–1320.
17. Farh KKH, Grimson A, Jan C, Lewis BP, Johnston WK, Lim LP, Burge CB, Bartel DP (2005) *Science* 310:1817–1821.
18. Voight BF, Kudaravalli S, Wen XQ, Pritchard JK (2006) *PLoS Biol* 4:446–458.
19. 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.
20. Saunders MA, Good JM, Lawrence EC, Ferrell RE, Li W-H (2006) *Am J Hum Genet* 79:1089–1097.
21. King MC, Wilson AC (1975) *Science* 188:107–116.
22. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, Palma A, Mikkelsen TS, Altshuler D, Lander ES (2006) *Science* 312:1614–1620.
23. Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ (2006) *Nucleic Acids Res* 34:D140–D144.
24. 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.
25. 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.
26. 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.
27. Barad O, Meiri E, Avniel A, Aharonov R, Barzilai A, Bentwich I, Einav U, Glad S, Hurban P, Karov Y, *et al.* (2004) *Genome Res* 14:2486–2494.