nature biotechnology

Allelic imbalance sequencing reveals that single-nucleotide polymorphisms frequently alter microRNA-directed repression

Jinkuk Kim[1–3] & David P Bartel[1,2]

© 2009 Nature America, Inc. All rights reserved.

Genetic changes that help explain the differences between two individuals might create or disrupt sites complementary to microRNAs (miRNAs)[1,2], but the extent to which such polymorphic sites influence miRNA-mediated repression is unknown. Here, we describe a method to measure mRNA allelic imbalances associated with a regulatory site found in mRNA transcribed from one allele but not found in that transcribed from the other. Applying this method, called allelic imbalance sequencing, to sites for three miRNAs (miR-1, miR-133 and miR-122) provided quantitative measurements of repression *in vivo* without altering either the miRNAs or their targets. A substantial fraction of polymorphic sites mediated repression in tissues that expressed the cognate miRNA, and downregulation was correlated with site type and site context. Extrapolating these results to the other broadly conserved miRNAs suggests that when comparing two mouse strains (or two human individuals), polymorphic miRNA sites cause expression of many genes (often hundreds) to differ.

MicroRNAs are ~23-nucleotide endogenous RNAs that pair to mRNAs to direct their post-transcriptional repression[3]. To explore miRNA regulatory diversity within a single species, we considered miRNA complementary sites that are created or disrupted by single-nucleotide polymorphisms (SNPs) in mice. Our study centered on three types of complementary sites that previous computational and experimental results indicated can mediate miRNA recognition[3]. Each of these three sites includes perfect Watson-Crick pairing to the miRNA seed (miRNA positions 2–7; **Fig. 1a**). One is a 7-nucleotide site, referred to here as the 7mer-m8 site, for which seed pairing is supplemented by a Watson-Crick match to miRNA nucleotide 8 (refs. 4–6). Another is the 7mer-A1 site, for which seed pairing is supplemented by an adenine nucleotide across from miRNA nucleotide 1 (ref. 4). And the third is the 8-nucleotide or 8mer site, which has both the m8 match and the A across from position 1 (ref. 4). We focused on sites recognized by three miRNAs—miR-1 (which for our purposes is synonymous with its paralog, miR-206), miR-133 and miR-122—because these miRNAs show strong, tissue-specific expression in relatively homogenous and accessible tissues, muscle (miR-1 and miR-133) or liver (miR-122)[7]. In agreement with previous reports[1,2], searching SNP databases[8,9] for polymorphisms within mRNA 3′ untranslated regions (UTRs), which are the regions most likely to be targeted by miRNAs[10], revealed many SNPs that create or disrupt sites for one of the three miRNAs (with gain or loss considered relative to an outgroup sequence). Because miRNAs often destabilize their target mRNAs[11], we reasoned that if these sites were functional in the tissue expressing the cognate miRNA, then less RNA might accumulate from the allele with the site. Moreover, in mice heterozygous for the SNPs, destabilization of mRNA from the target allele, but not from the nontarget allele, would contribute to allelic imbalance in mRNA steady-state levels. Hence, we developed allelic imbalance sequencing (AI-Seq) to measure such imbalances, reasoning that any imbalances would identify and quantify miRNA regulatory diversity within a species, and provide a unique opportunity to examine the molecular consequences of miRNA-mediated repression *in vivo* without perturbing either the miRNA or its targets.

Because lab strains lack the heterozygosity found in natural populations, we performed five inter-strain crosses to generate mice heterozygous for the parental alleles. Approximately 300 annotated SNPs that create or disrupt target sites for one of the three miRNAs were heterozygous in F_1 progeny from at least one of the five crosses. We chose a subset of these, preferring those that create or disrupt 8mer sites, those in messages with evidence of expression in the tissues expressing the miRNAs and those that were not linked to many nearby polymorphisms. Allelic imbalance was measured for 67 target sites (28 for miR-122, 28 for miR-1 and 11 for miR-133) in the tissue expressing the cognate miRNA.

For AI-Seq, mRNA fragments containing the SNPs were first reverse transcribed and amplified (PCR), and then the amplicon was subjected to high-throughput pyrosequencing[12] (**Fig. 1b**). To economize on sequencing, we pooled amplicons derived from different primers. Because the primers flanking the SNPs used for RT-PCR were gene specific but not allele specific, both alleles of the same gene were amplified by the same reaction, and their relative abundance could be inferred from the number of sequencing reads representing each allele. We quantified these relative abundances using the allelic ratio, defined here as the $\log_2$ of the number of reads representing the target allele

[1]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts, USA. [2]Howard Hughes Medical Institute and Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [3]Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts, USA. Correspondence should be addressed to D.P.B. (dbartel@wi.mit.edu).

Received 20 January; accepted 6 April; published online 26 April 2009; doi:10.1038/nbt.1540

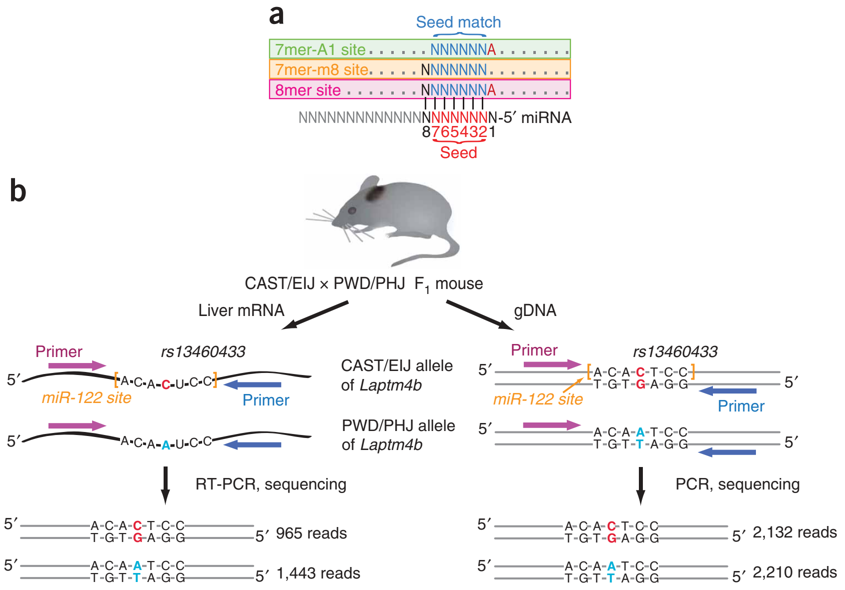
a
Seed match
7mer-A1 site . . . . . . NNNNNNA . . . .
7mer-m8 site. . . . . NNNNNNN. . . . .
8mer site . . . . NNNNNNNA . . . .
NNNNNNNNNNNNNNNNNNNNNN-5′ miRNA
87654321
Seed
b
CAST/EIJ × PWD/PHJ F1 mouse
Liver mRNA
gDNA
Primer
rs13460433
5′
A-C-A-C-U-C-C
miR-122 site
Primer
CAST/EIJ allele of Laptm4b
5′
A-C-A-A-U-C-C
PWD/PHJ allele of Laptm4b
RT-PCR, sequencing
5′ A-C-A-C-T-C-C T-G-T-G-A-G-G 5′ 965 reads
5′ A-C-A-A-T-C-C T-G-T-T-A-G-G 5′ 1,443 reads
Primer
rs13460433
5′ A-C-A-C-T-C-C T-G-T-G-A-G-G 5′
miR-122 site
Primer
5′ A-C-A-A-T-C-C T-G-T-T-A-G-G 5′
PCR, sequencing
5′ A-C-A-C-T-C-C T-G-T-G-A-G-G 5′ 2,132 reads
5′ A-C-A-A-T-C-C T-G-T-T-A-G-G 5′ 2,210 reads

Figure 1 Measurement of mRNA allelic imbalances associated with heterozygous miRNA target sites. (**a**) Canonical 7- to 8-nucleotide miRNA target sites. (**b**) AI-Seq, illustrated for SNP rs13460433, which generates a heterozygous miR-122 target site in *Laptm4b* of CAST/EIJ × PWD/PHJ F_1 mice. When using liver mRNA, the 965 and 1,443 reads obtained from the target and the nontarget allele, respectively, implied an allelic imbalance of 0.67 (965/1,443). Because allele-specific PCR bias might have influenced this ratio, amplification and sequencing was performed in parallel with the same primers but using gDNA instead of mRNA. The gDNA template produced a target/nontarget ratio of 0.96 (2,132/2,210), enabling the raw allelic imbalance to be corrected to 0.70 (0.67/0.96) or –0.51 in $\log_2$ scale. This allelic ratio implied that in liver, a tissue expressing miR-122, the mRNA abundance of the target allele is 70% of that of the nontarget allele.

divided by the number of reads representing the non-target allele, after normalizing to the ratio obtained using genomic DNA (**Fig. 1b**).

If none of the intact miRNA sites directed repression, the allelic ratios would be expected to center on zero, with individual ratios deviating from zero because of experimental noise. However, contrary to this null hypothesis, when liver tissues were assayed using AI-Seq, the distribution of the allelic ratios for the 28 miR-122 sites centered below zero (**Fig. 2a**), consistent with the hypothesis that mRNA from some of the alleles with target sites was destabilized. If miR-122 caused this destabilization, then the shift from zero should depend on the presence of this miRNA. To test this dependency, we measured ratios for 22 of the 28 miR-122 sites in muscle, which does not express miR-122. Ratios for the remaining six sites were not considered because four were in messages not expressed in muscle, and the other two were unusual in that the SNP disrupting the miR-122 site (CA<u>C</u>TCCA and ACA<u>C</u>TCC, SNP underlined) simultaneously created a site for miR-1 (CA<u>T</u>TCCA and ACA<u>T</u>TCC), which is expressed in muscle, thereby precluding the use of these as negative controls. As expected for a miRNA-mediated effect, the shift from zero disappeared in muscle (**Fig. 2a**). When analyzing the ratios for the 39 sites for miR-1 or miR-133, which are expressed in the muscle but not in the liver, the reciprocal pattern was observed—namely, the distribution of the ratios measured in liver centered on zero and that of the ratios measured in muscle was skewed toward lower values (**Fig. 2b**).

To increase sample size and thereby achieve statistical significance, we combined data sets such that the ratios measured in the presence of the cognate miRNA were analyzed together and compared to those measured in the absence of the miRNA (**Fig. 2c**). A significantly large fraction of the allelic ratios were <0 in the presence of the miRNA ($P < 0.01$, one-sided exact binomial test), but not in the absence of the miRNA ($P = 0.6$), and the difference between the two distributions also was significant ($P = 0.02$, one-sided Kolmogorov-Smirnov (KS) test). Thus, we concluded that at least a subset of the interrogated target sites mediated repression.

On average, the polymorphic sites were associated with mRNA downregulation of 12% (**Fig. 2c**; 95% confidence interval of 5–18%, bootstrapping). Actual downregulation was likely greater because the signal could have been diluted by both nuclear mRNA and mRNA from cells that do not express the cognate miRNA, such as those from blood or vasculature. Effects of functional sites also might have been diluted by inclusion of nonfunctional sites. Nonfunctional sites presumably were enriched among the set of sites interrogated in this study because natural selection selects against polymorphisms that either disrupt beneficial functional sites or generate functional sites in messages that should not be repressed[13–15].

We estimated the lower bound for the fraction of functional sites to be 16% by analyzing the maximal vertical displacement of the cumulative distribution curves (correcting for the bumpiness of the distributions[10]). This estimate is likely to be conservative because simulations showed that under certain assumptions our analysis may only identify about a third of all sites simulated to be functional (see Methods). These simulations incorporated the variability observed from the tissues lacking the miRNA and assumed that all sites mediated target repression by 20%. If, as in this simulation, only about a third of the active sites were detected, then our lower bound of 16% might be only a third of the actual fraction, in which case about half of the examined polymorphic sites mediated repression.

The variability observed in our experiments can be attributed to multiple sources. One source is stochastic sampling error inherent to counting sequencing reads, which can be modeled by the binomial distribution (**Fig. 2d**). A second source is PCR variability, which can be estimated as the variability of the allelic ratios measured using gDNA minus the stochastic error (**Fig. 2d**; difference between gDNA and binomial distributions). A third source is biological noise, which could include differences in the epigenetic states of the two alleles or allelic differences in linked *cis*-regulatory elements. To begin to estimate the biological noise, we examined the distribution of the allelic ratios of control mRNAs that were not predicted to be repressed in an allele-specific manner by the three miRNAs, such as mRNAs from tissues lacking the cognate miRNA. Allelic ratios of these control mRNAs were substantially more variable than those of gDNA, suggesting frequent allelic imbalance not attributable to the sites under investigation (**Fig. 2d**). However, we were unable to quantify the frequency or magnitude of this potentially widespread allelic imbalance because of the possibility that reverse transcription variability also contributed to the greater variability of the mRNA controls compared to that of the gDNA controls.

Our quantitative assay of site efficacy *in vivo* did not perturb either the miRNAs or their targets and thereby provided a fresh opportunity

© 2009 Nature America, Inc. All rights reserved.

© 2009 Nature America, Inc. All rights reserved.

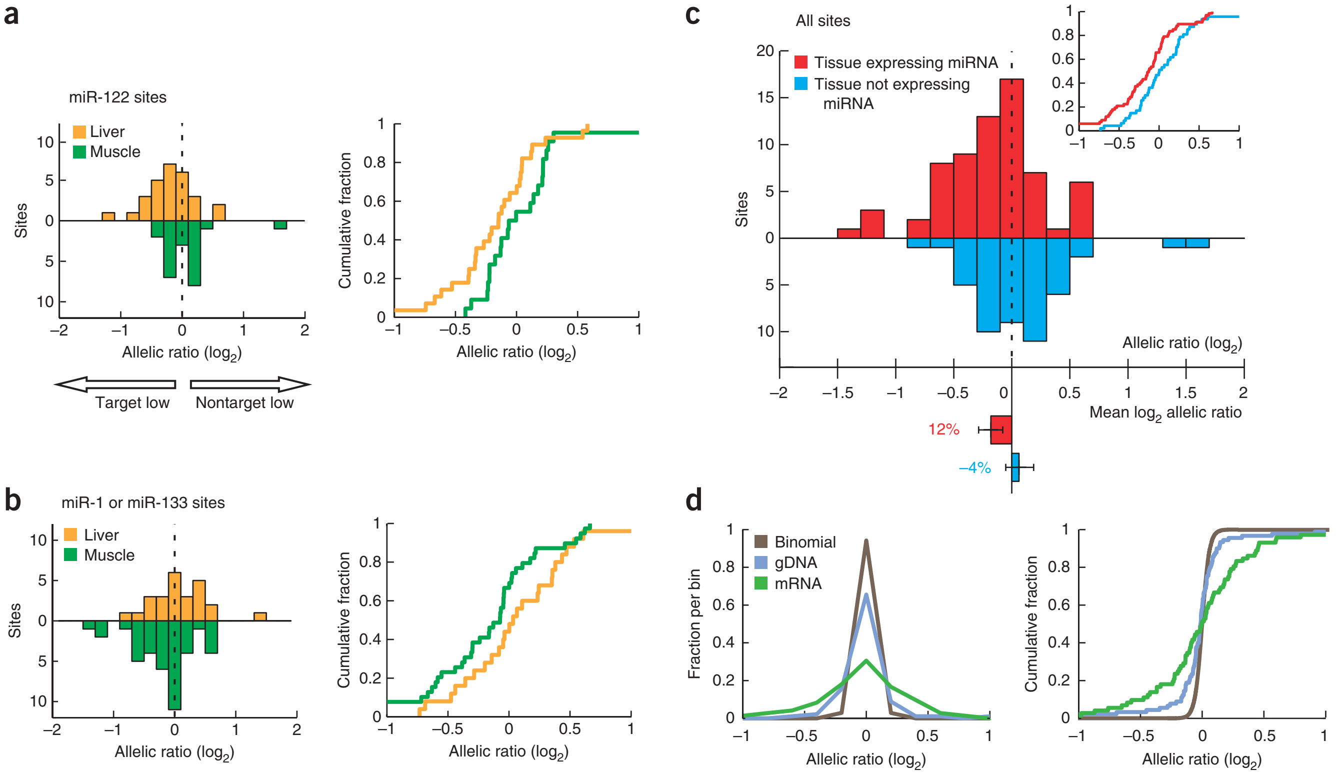

Figure 2 Impact of heterozygous target sites on mRNA allelic imbalance. (**a**) Distribution of allelic ratios, $\log_2$ (target/nontarget), measured for miR-122 polymorphic sites using mRNA from either liver ($n = 28$) or muscle ($n = 22$), plotted as a histogram (left, 0.2-unit bins) and cumulative distribution (right). (**b**) Distribution of allelic ratios measured for miR-1 and miR-133 polymorphic sites using mRNA from either liver ($n = 25$) or muscle ($n = 39$), plotted as in panel **a**. (**c**) Distribution of allelic ratios, pooling ratios from panels **a** and **b**, measured using either mRNA from the tissue expressing the cognate miRNA ($n = 67$), or mRNA from the tissue not expressing the cognate miRNA ($n = 47$). In the inset are the cumulative distributions, plotted as in panels **a** and **b**. Below the histogram is the mean offset from zero for the two distributions, with error bars indicating 95% confidence intervals (bootstrapping) for the mean, and the percentages indicating the average downregulation of target alleles compared to nontarget alleles. (**d**) Sources of variability in allelic ratios, depicted with standard (left, 0.2-unit bins) and cumulative (right) distributions. Total variability not attributable to the cognate miRNAs was measured using mRNA with heterozygous sites not predicted to be regulated by the cognate miRNAs ($n = 72$). Of the 72 ratios determined, 47 were from mRNAs of tissues lacking the cognate miRNA, and 25 were from mRNAs without predicted potential for allele-specific repression mediated by the three miRNAs. (These 72 ratios were not normalized to corresponding gDNA ratios.) PCR variability was measured using gDNA ($n = 90$). Stochastic counting error was simulated using the binomial model ($n = 9{,}000$, 100 simulations per gDNA measurement), with total counts for each simulated amplicon chosen to match those of the gDNA measurements. The differences between each of the three possible pairs of distributions were statistically significant ($P < 0.01$ for each comparison, two-sided KS test).

to examine the influence of site type and site context on miRNA activity. The 8mer sites performed significantly better than did 7mer-m8 or 7mer-A1 sites (**Fig. 3a**; $P = 0.005$ and $P = 0.001$ respectively, one-sided KS test), and 7mer-m8 sites tended to perform slightly better than did 7mer-A1 sites, although this difference was not statistically significant ($P = 0.1$, one-sided KS test). The overall rank order of the efficacy of the three types was consistent with previous observations from experiments that ectopically expressed or deleted miRNAs[10,16–18].

To consider the influence of site context, we calculated the 'context score' for each polymorphic site. Context scores quantitatively evaluate site type and three features of site context (that is, surrounding AU content, position within the 3′ UTR and pairing to the 3′ region of miRNA) to predict site efficacy[10]. Context scores significantly correlated with target downregulation in the presence of the cognate miRNA, but not in the absence of the miRNA ($P < 0.001$; **Fig. 3b**). Significant correlation was retained in the presence of the miRNA even after the contribution of site type had been factored out, thereby indicating that site context, as scored by this model, influences the efficacy of polymorphic sites ($P < 0.01$; **Fig. 3c**).

Our experiments focused on three of the 87 miRNA families conserved in chicken or more divergent vertebrates[19]. Expanding our SNP database search to the other 84 broadly conserved miRNA families and the ~8 million SNPs annotated in 15 mouse strains[8] showed that any two strains have on average 2,430 distinct polymorphic sites (bottom and top 2.5 percentile, 810–4,600; median, 1,470) and 1,510 genes with at least one polymorphic miRNA site (bottom and top 2.5 percentile, 520–2,790; median, 950). These numbers would increase if sites recognized by the hundreds of additional annotated miRNAs were also considered. However, because species-specific miRNAs and those conserved only within mammals tend to be expressed at lower levels, their 7- to 8-nucleotide sites are thought to be less frequently sufficient for mediating repression[3,19]. Therefore, to guard against overstating the impact of polymorphic sites, we did not consider these additional miRNAs.

An estimate of the direct impact of polymorphic miRNA sites on gene-expression variation within a species can be extrapolated from our results as follows. First, for the 67 sites examined, we observed average downregulation of 12%, with at least 16% of the sites

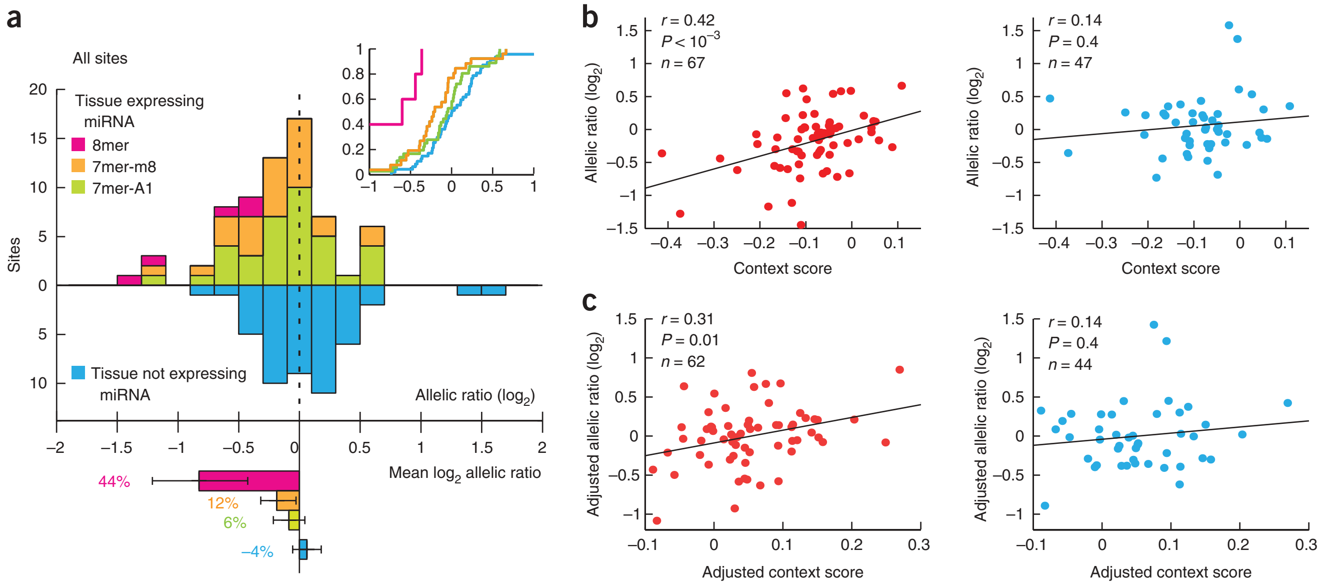

Figure 3 Dependence of target site efficacy on site type and site context. (**a**) Efficacy of target sites of different types (8mer, $n = 5$; 7mer-m8, $n = 26$; 7mer-A1, $n = 36$), plotted as in **Figure 2c**. (**b**) Relationship between allelic ratio and context score in the tissue expressing (left) and not expressing (right) the miRNA. Lines are the least-squares fit to the data (r, Pearson correlation coefficient, with P-value estimated by two-sided Pearson correlation test). When considering only the 47 sites in the left panel that were measured also in the absence of the cognate miRNA (right panel), the correlation remained significant ($r = 0.58$ and $P < 10^{-4}$). (**c**) Relationship between adjusted allelic ratio and adjusted context score for 7-nucleotide sites, plotted as in **b**. Context score and allelic ratio were each adjusted by subtracting the portion contributed by site type and the mean allelic ratio of sites of the same type, respectively. When considering only the 44 sites in the left panel that were measured also in the absence of the cognate miRNA (right panel), the correlation remained significant ($r = 0.53$ and $P < 10^{-3}$).

responsible for the observed downregulation. Correcting for our preference in choosing 8mer sites for analysis slightly lowered the average downregulation to 10% and the percentage of functional sites to 15% as our best estimates for all 7- to 8-nucleotide polymorphic sites in mRNAs expressed in cognate tissues. If we assume that 50% of the genes with these sites are coexpressed with the cognate miRNA in the same cell type, we can use our observed lower limit of 15% functional sites to estimate that at least 7.5% (0.5 × 0.15) of the genes with polymorphic sites will be differentially regulated between two strains. Thus, between many mouse strains, over a hundred messages are likely to be differentially regulated through polymorphic sites, with average mRNA downregulation for these messages >60% (correcting for dilutive effect of inactive sites by dividing the average downregulation of all sites, 10%, by 0.15). Using a less conservative estimate that 50% of the sites in cognate tissues are functional, proportionally more messages would be downregulated, with average downregulation of functional sites still ~20%. Because miRNAs also influence translation, effects on the proteome are presumed to be even greater. Overall, it is hard to escape the conclusion that polymorphic miRNA regulatory sites have a substantial impact on gene-expression variation within a species.

Our results in mice, considered with SNP frequencies in humans, indicated that any two unrelated humans probably have more than a hundred genes differentially regulated because of polymorphic miRNA targeting (**Supplementary Discussion** online). Assuming that some of these could explain differences in disease risk among individuals, our results suggest that, as more genome-wide association studies are conducted with improved coverage in 3′ UTRs, more miRNA target site polymorphisms will be associated with clinical conditions and individual traits[2].

Another approach for detecting effects of regulatory SNPs is provided by studies of expression quantitative trait loci (eQTL)[20]. In eQTL studies, correlation between the genotype of a polymorphic locus and expression of a gene is calculated for each locus:gene pair. In principle, these studies involving unrelated individuals should preferentially identify polymorphic targets as *cis*-regulated because the SNPs in functional target sites (and other linked SNPs) should be associated with expression of the targets. However, when we analyzed the results of a large-scale eQTL study that used >400 human liver samples[21], polymorphic miR-122 targets were not enriched among the genes identified as *cis*-regulated any more than were polymorphic miR-1 targets (data not shown). We attribute the greater sensitivity of AI-Seq to the internal reference provided by the nontarget allele, which normalizes for environmental differences, *trans*-acting genetic differences and other sources of sample variability, thereby more effectively isolating the influence of the site on expression. Also important for the success of our approach in detecting the relatively subtle effect of miRNAs was the precision achieved by high-throughput sequencing. Previous studies using heterozygous SNPs to detect allelic expression imbalances rely on allele-specific hybridization or primer extension[22–25], both of which, when compared at the gDNA level, were substantially noisier than our sequencing-based method (**Supplementary Fig. 1** online).

Despite detecting the effects of polymorphic miRNA sites in mouse tissues, miRNA effects were not detected in a HapMap[26] panel of lymphoblastoid cell lines when we used AI-Seq to measure the allelic imbalance of 56 heterozygous target sites for nine miRNA families most highly expressed in these cell lines (data not shown). In this case, the imbalances expected to result from polymorphic miRNA sites might have been overwhelmed by random monoallelic expression present in clonal subpopulations of these lines[27]. Moreover, the process of establishing lymphoblastoid cell lines, which involved Epstein-Barr-virus infection and subsequent transformation of B-cells, might have downregulated miRNA expression[28].

© 2009 Nature America, Inc. All rights reserved.

Experiments examining the influence of miRNA knockouts on the transcriptome and proteome have been informative for inferring the effects of conserved and nonconserved sites that are not polymorphic[17,29]. Our results complement these studies by revealing the influence of polymorphic sites without perturbing either the miRNAs or their targets. Following miRNA knockout, upregulation of targets can trigger feedback regulation that reduces the observed effect of losing the miRNA. Such a response is not likely to confound our AI-Seq results because feedback regulation is usually not allele specific and therefore is unlikely to change the relative expression of the target compared to the nontarget alleles. Our approach can be extended to characterize other *cis*-regulatory elements that might influence mRNA levels. As the capacity of high-throughput sequencing increases, we anticipate that RNA-Seq coverage will expand so that directed amplification of specific loci will no longer be required to accurately detect allelic imbalances. Then, our approach of correlating imbalances with predicted regulatory sites can be applied transcriptome-wide to reveal many of the polymorphic regulatory sites contributing to these imbalances.

METHODS

Mouse tissues and preparation of cDNA and gDNA. The study was approved by the MIT Committee on Animal Care. The Jackson laboratory performed five inter-strain crosses (CAST/EIJ × PWD/PHJ, FVB/NJ × PWD/PHJ, A/J × C57BL/6J, WSB/EiJ × MOLF/EiJ, A/J × DBA/2J) and dissected liver and skeletal muscle from two 4-week-old F_1 littermates of each cross. For each cross and tissue, ~0.6 g tissue (~0.3 g from each littermate) was homogenized for RNA extraction (RNeasy Maxi kit, Qiagen), and cDNA was synthesized from total RNA in reverse transcriptase reactions (Superscript III, Invitrogen) primed with random hexamers. For each cross, gDNA was isolated from ~50 mg of either liver or muscle from either littermate (DNeasy Blood and Tissue kit, Qiagen).

Computational identification of polymorphic sites. Genomic coordinates of known mouse SNPs on the July 2007 genome assembly (mm9) were obtained from NCBI dbSNP build 128 (ref. 9). Genotypes of mouse strains were obtained from dbSNP build 128, mm9 genomic sequence (for C57BL/6J strain) and the Perlegen data (http://mouse.perlegen.com/)[8]. Gene annotation on mm9 was obtained from UCSC genome browser. We identified SNPs that generate heterozygous sites for miR-122, miR-1 or miR-133 in at least one of the five crosses (123 SNPs for miR-122, 109 SNPs for miR-1 and 74 SNPs for miR-133; 7-nucleotide sites to 8-nucleotide sites ratio, 10.3), excluding those that modify the sites, for example, by converting a 7mer site to an 8mer site or vice verse.

Site selection and DNA amplification. Polymorphic sites located <15 nucleotides from the stop codon were excluded[10]. Also excluded were those in the genes expressed, according to the mouse expression atlas[30], at a level lower than that of 90% of all genes in the tissue that expresses the cognate miRNA; sites in genes without an expression measurement were not excluded. Out of the remaining sites, all polymorphic 8mer sites were chosen. A subset of the 7-nucleotide polymorphic sites were chosen somewhat arbitrarily, preferring those with fewer additional SNPs in flanking regions, which could potentially interfere with primer annealing. A total of 138 SNPs were carried forward for primer design, performed with the aid of PRIMER3, and suitable primers were found for 124 of those, which corresponded to 136 polymorphic sites (7-nucleotide sites to 8-nucleotide sites ratio, 5.5). PCR amplification reactions of each of the SNPs were performed individually (Phusion Hot Start polymerase, New England Biolabs). All PCR reactions done with cDNA were accompanied by a matching no-reverse transcriptase control. Fragments that failed to be amplified at sufficient yield from either gDNA or cDNA were discarded. For each successful amplification, the procedure was repeated using cDNA from the noncognate tissue. As additional controls to examine variability of allelic imbalances that were not attributable to miRNA targeting, SNPs that do not generate polymorphic miRNA sites for the three miRNAs were identified in the open reading frames (ORFs) of 27 genes that had polymorphic sites in the 3′ UTRs, and these 27 SNPs were amplified using F_1 hybrids that were heterozygous for the ORF SNP but homozygous for the 3′ UTR SNP. In total, 240 amplicons (70 3′ UTR SNPs from cDNA of the tissue with miRNA expression, 49 3′ UTR SNPs from cDNA of the tissue without miRNA expression, 27 ORF SNPs from cDNA of either tissue and 94 SNPs from gDNA) were prepared for sequencing.

Mixing and purifying PCR products for pyrosequencing. Because gDNA-templated, liver-mRNA-templated and muscle-mRNA-templated amplicons all shared the same primers, they needed to be sequenced in separate pools, so that they could be distinguished from one another. Because one pyrosequencing plate can be divided into four segments without contamination between segments, the 240 amplicons were mixed into four pools. Each pool had ~60 amplicons, mixed in equimolar amounts after determination of each amplicon concentration (Bioanalyzer, Agilent Technologies). Each pool was deproteinated (phenol, chloroform with iso-amyl-alcohol), purified by native PAGE gel, taking precautions to avoid denaturing the double-stranded PCR products, and submitted to 454 Life Sciences for sequencing. Three sequencing runs were performed.

Analysis of sequencing reads. Of the ~1.194 million reads acquired, ~1.085 million (~91%) correctly mapped to the 3′-terminal 10-nucleotide fragments of the unique primer pairs. Of the 240 amplicons, 9 were excluded for at least one of the following reasons: (i) the number of reads obtained per amplicon was <300, (ii) a SNP was not detected, (iii) a severe allelic bias was observed with gDNA. The remaining 231 amplicons had a median of 3,978 reads (range, 541–18,714) and corresponded to 65 3′ UTR SNPs and 25 ORF SNPs. Information and results for each amplicon are provided (**Supplementary Tables 1** and **2** online). Although most of the sites were polymorphic for only one of the three miRNAs, five exceptional SNPs allowed one allele to have a miR-122 site and the other allele to have a miR-1 or miR-133 site. Three of the five were in mRNA expressed only in muscle, but the remaining two (NCBI dbSNP IDs: rs36333425, rs30114270) were expressed in both tissues, allowing the measurement taken from liver to report on the miR-122 site and that from muscle to report on miR-1 site. Other exceptions were the two polymorphic sites in the same mRNA (NCBI dbSNP IDs: rs32325030, rs32323893) that exist in the same cross and collectively allow one *Snap29* allele to have two miR-122 sites and the other allele to have none. For these, the allelic ratio was measured separately in liver for each site, but because each ratio was likely to reflect the effect of two target sites, each was analyzed after reduction by half the $\log_2$ value.

Evaluation of previous allelic imbalance measurement methods. To evaluate SNP arrays, we downloaded from the HapMap[26] website the raw signal-intensity data generated by hybridizing gDNA of a HapMap individual (NA19193) on the Affymetrix GeneChip 250K Nsp array. For evaluating the GoldenGate primer-extension assay, we downloaded from the Gene Expression Omnibus the raw signal-intensity data (GSM199494, GSM200074) generated from the Illumina GoldenGate assay with gDNA of a HapMap individual (NA10836)[24]. For both cases, SNPs annotated as heterozygous in the individual were identified from the HapMap genotype database, and the allele-specific probe intensity values for the SNPs were used to calculate $\log_2$ ratios of one random allele to the other.

Statistical analyses. MATLAB was used for all statistical analyses. To calculate the statistical significance for the observed number of allelic ratios with values < 0, we used the one-sided exact binomial test, in which the *P* value was the probability that a random variable following binomial distribution (parameters: $P = 0.5$, $n =$ [the total number of sites]) was equal to or larger than the observed number. To calculate the significance of difference between two distributions, we chose the KS test over the Wilcoxon rank sum test (Mann-Whitney U test) or *t*-test, because the KS test is based on fewer assumptions on the data and almost always provided the most conservative *P*-value compared to the other two tests. When estimating the lower bound for the number of active polymorphic sites, the contribution of experimental noise (illustrated by the bumpiness of the cumulative distributions) in increasing the maximal offset between the cognate and control distributions needed to be subtracted. To estimate the contribution of this noise, we merged the two distributions and

© 2009 Nature America, Inc. All rights reserved.

generated 1,000 pairs of distributions by random sampling, with replacement, maintaining the sizes of the original distributions. Then we calculated the maximum difference in cumulative fraction for each pair of simulated distributions and subtracted the median of the 1,000 values from the observed maximum offset. When simulating under the assumption that all sites mediate 20% downregulation, we started with the control distribution of 47 allelic ratios measured using the tissue lacking the miRNA and randomly drew (with replacement) 67 samples, which matched the size of the cognate distribution. To simulate 20% downregulation mediated by all sites, the sampled allelic ratios were each adjusted by offsetting them by –0.32 or $\log_2(0.8)$. We generated 1,000 such simulated distributions, each of which was compared to the control distribution to determine the maximum offset in cumulative fraction. The median of the resulting 1,000 values was considered as the representative estimate for the maximum cumulative difference between the simulated and control distributions. This difference was corrected for the bumpiness of the distributions, as explained above, to yield the detectable fraction of functional sites. The observed average downregulation of all examined polymorphic sites was corrected for our preferential choice of 8mer sites for analysis by recalculating the mean allelic ratio of all sites after reducing the contribution from 8mer sites by 1.87 fold (10.3/5.5), which was the enrichment of 8mer sites among the polymorphic sites analyzed. The observed lower bound for the fraction of functional sites was similarly adjusted.

Accession number. GEO, GSE15675.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

We thank Tim Harkins and the 454 Sequencing Facility for high-throughput sequencing, members of the laboratory for helpful comments on this manuscript, and T. DiCesare for illustration. Supported by a grant from the National Institutes of Health. D.P.B. is an investigator of the Howard Hughes Medical Institute.

AUTHOR CONTRIBUTIONS

J.K. performed all experiments and analyses. Both authors designed the experiments and wrote the manuscript.

Published online at http://www.nature.com/naturebiotechnology/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

© 2009 Nature America, Inc. All rights reserved.

1. Clop, A. *et al.* A mutation creating a potential illegitimate microRNA target site in the myostatin gene affects muscularity in sheep. *Nat. Genet.* **38**, 813–818 (2006).
2. Sethupathy, P. & Collins, F.S. MicroRNA target site polymorphisms and human disease. *Trends Genet.* **24**, 489–497 (2008).
3. Bartel, D.P. MicroRNAs: target recognition and regulatory functions. *Cell* **136**, 215–233 (2009).
4. Lewis, B.P., Burge, C.B. & Bartel, D.P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20 (2005).
5. Brennecke, J., Stark, A., Russell, R.B. & Cohen, S.M. Principles of microRNA-target recognition. *PLoS Biol.* **3**, e85 (2005).
6. Krek, A. *et al.* Combinatorial microRNA target predictions. *Nat. Genet.* **37**, 495–500 (2005).
7. Lagos-Quintana, M. *et al.* Identification of tissue-specific microRNAs from mouse. *Curr. Biol.* **12**, 735–739 (2002).
8. Frazer, K.A. *et al.* A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* **448**, 1050–1053 (2007).
9. Sherry, S.T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–311 (2001).
10. Grimson, A. *et al.* MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell* **27**, 91–105 (2007).
11. Lim, L.P. *et al.* Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**, 769–773 (2005).
12. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
13. Chen, K. & Rajewsky, N. Natural selection on human microRNA binding sites inferred from SNP data. *Nat. Genet.* **38**, 1452–1456 (2006).
14. Farh, K.K. *et al.* The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* **310**, 1817–1821 (2005).
15. Stark, A., Brennecke, J., Bushati, N., Russell, R.B. & Cohen, S.M. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. *Cell* **123**, 1133–1146 (2005).
16. Nielsen, C.B. *et al.* Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA* **13**, 1894–1910 (2007).
17. Baek, D. *et al.* The impact of microRNAs on protein output. *Nature* **455**, 64–71 (2008).
18. Selbach, M. *et al.* Widespread changes in protein synthesis induced by microRNAs. *Nature* **455**, 58–63 (2008).
19. Friedman, R.C., Farh, K.K., Burge, C.B. & Bartel, D.P. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.* **19**, 92–105 (2009).
20. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* **10**, 184–194 (2009).
21. Schadt, E.E. *et al.* Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.* **6**, e107 (2008).
22. Lo, H.S. *et al.* Allelic variation in gene expression is common in the human genome. *Genome Res.* **13**, 1855–1862 (2003).
23. Gimelbrant, A., Hutchinson, J.N., Thompson, B.R. & Chess, A. Widespread monoallelic expression on human autosomes. *Science* **318**, 1136–1140 (2007).
24. Tan, A.C. *et al.* Allele-specific expression in the germline of patients with familial pancreatic cancer: an unbiased approach to cancer gene discovery. *Cancer Biol. Ther.* **7**, 135–144 (2008).
25. Serre, D. *et al.* Differential allelic expression in the human genome: a robust approach to identify genetic and epigenetic cis-acting mechanisms regulating gene expression. *PLoS Genet.* **4**, e1000006 (2008).
26. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
27. Plagnol, V. *et al.* Extreme clonality in lymphoblastoid cell lines with implications for allele specific expression analyses. *PLoS One* **3**, e2966 (2008).
28. Lu, J. *et al.* MicroRNA expression profiles classify human cancers. *Nature* **435**, 834–838 (2005).
29. Rodriguez, A. *et al.* Requirement of bic/microRNA-155 for normal immune function. *Science* **316**, 608–611 (2007).
30. Su, A.I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* **101**, 6062–6067 (2004).