www.sciencemag.org/cgi/content/full/1121158/DC1

Supporting Online Material for

The Widespread Impact of Mammalian MicroRNAs on mRNA Repression and Evolution

Kyle Kai-How Farh, Andrew Grimson, Calvin Jan, Benjamin P. Lewis, Wendy K. Johnston, Lee P. Lim, Christopher B. Burge, David P. Bartel*

*To whom correspondence should be addressed. E-mail: dbartel@wi.mit.edu

Published 24 November 2005 on *Science* Express
DOI: 10.1126/science.1121158

This PDF file includes:

Other Supporting Online Material for this manuscript includes the following:
(available at www.sciencemag.org/cgi/content/full/1121158/DC1)

Supporting Online Material

Materials and Methods

Expression and Sequence Data Mouse expression data were obtained from the Novartis Research Foundation (*S1*). The data comprised two replicates of 61 different tissue samples hybridized to the Affymetrix GNF1M mouse chip, and were normalized using Affymetrix Microarray Suite 5.0 (MAS5) software. Human and mouse annotated 3'UTR sequence data were obtained from Refseq. Orthologous human, mouse, rat, and dog 3'UTR data were derived from the UCSC genome browser (*S2*) multiZ multiple genome alignments (*S3*).

We selected 8,551 genes for our analysis, using the following criteria: (1) each gene had unambiguous mouse and human reciprocal orthologs, (2) each ortholog had both mouse and human UTRs annotated by Refseq, and (3) the gene was represented on the GNF1M chip. Affymetrix probe IDs ending with *_x_at*, which do not uniquely complement the target sequence and hence are likely to cross-hybridize, were excluded. When multiple probes mapped to a single gene entry, we used the arithmetic mean of the probe intensities.

For each gene in each tissue sample, we assigned an absolute expression rank and a relative expression rank, thereby creating two 8,551 X 61 gene–by-tissue sample matrices. To calculate the absolute expression rank for each gene, the geometric mean of the two replicates was sorted with respect to the geometric means of the other 8550 genes in that sample; values were placed in 61 equal-sized bins, of increasing absolute expression. To assign the relative expression rank for each gene in a sample, we ranked the gene in that sample with respect to expression in all other samples, again using the geometric mean of the two replicate values. We then sorted the 8,551 genes for each tissue into 61 equal-sized bins according to their relative rank. Bin 1 contained genes which had the lowest expression compared with expression of those genes in other tissues, while bin 61 contained genes which were highest in that tissue compared with their expression in other tissues.

MicroRNA Families and Site Identification Our analysis included 73 miRNA families [listed in fig. S7; miRNA sequences can be found at miRBase (*S4*)], which were defined as sets of miRNAs with identical nucleotides at positions 2-8. All families were required to have at least one member conserved across human, mouse, rat, dog, and zebrafish. Except for sites in Fig. 2A, potential miRNA regulatory sites were found by searching the 3'UTR sequences for 7-nt matches, which included a 6-nt match to the miRNA seed (nucleotides 2-7) and either a seventh Watson-Crick match to miRNA nucleotide 8 or an adenosine opposite nucleotide 1 (*S5*). Conserved sites were identified using UCSC MultiZ alignments of orthologous 3'UTR regions from human, mouse, rat and dog (*S3*). Genes were considered to contain conserved sites if their 3'UTR contained an aligned 7-nt match in all four genomes. Mouse and human annotated UTRs often differed slightly in length; for such cases, the alignments were based on the species with the shorter annotated UTR. Mouse and human genes with nonconserved sites (Fig 2B and 2C, respectively) were identified using the Refseq annotated 3'UTR set for that genome. To

simplify the description of our methods, we sometimes refer to genes with conserved sites as “conserved targets” and those with nonconserved sites as “nonconserved targets,” recognizing that these are not necessarily biological targets.

For the purposes of evaluating our method (e.g., Fig. 4C), we generated a cohort of control sequences in which the miRNA seed regions were shuffled while maintaining dinucleotide and mononucleotide frequencies approximating actual miRNA seed regions, which were further required to differ in sequence from any known conserved mammalian miRNA.

Background Calculation We calculated the likelihood of a 7-nt miRNA seed matching with a given 3'UTR as a function of trinucleotide frequencies and length. The second order Markov probability of a 7-nucleotide seed matching an arbitrary 7mer $x_1x_2x_3x_4x_5x_6x_7$ in the UTR was calculated as:

$$P(\text{matching\_a\_given\_7mer}) \sim P(x_1x_2x_3) \times P(x_2x_3x_4|x_2x_3) \times P(x_3x_4x_5|x_3x_4) \ldots \times P(x_5x_6x_7|x_5x_6)$$
$$= P(x_1x_2x_3) \times P(x_2x_3x_4) / P(x_2x_3) \times P(x_3x_4x_5) / P(x_3x_4) \ldots \times P(x_5x_6x_7) / P(x_5x_6)$$

where $P(x_m..x_n)$ is the probability of matching nucleotides *m* through *n* of the 7mer.

The probability of each trinucleotide and dinucleotide was estimated for each UTR as the observed frequency of the trinucleotide or dinucleotide in that UTR, without using pseudocounts. This approach assumes that the overall dinucleotide and trinucleotide frequencies in a UTR approximate those for smaller windows within the UTR, and that the sequences of the miRNA seeds and the UTR are sufficiently nondegenerate such that the probabilities of di- and trinucleotides and 7mers are approximately independent. From this calculation of the probability of matching any given 7mer, we calculated the probability of the 7-nt sequence occurring in the UTR, based on length.

$$P(\text{UTR\_contains\_7mer}) = 1 - P(\text{UTR\_does\_not\_contain\_7mer})$$
$$= 1 - (1-P(\text{7nt\_match}))^{(\text{number\_of\_7mers})}$$
$$= 1 - (1-P(\text{7nt\_match}))^{(\text{UTR\_length} - 6)}$$

To validate that this approach was accurately estimating the likelihood of a given miRNA seed to match to a given UTR, we calculated the number of targets (either conserved or nonconserved) for each miRNA in the annotated mouse and human UTR sets. The total expected number of UTRs targeted by all miRNA families was 95% of the actual observed number of targets in mouse and 93% of the actual observed number of targets in human, and the Pearson correlation between number of targets for each individual miRNA in the expected and observed sets was 0.93 for both mouse and human. When we repeated this analysis for our control set of shuffled miRNA seeds, the total expected number of UTRs targeted by all miRNA families was 99% and 100% of the actual observed number of targets in human and mouse, respectively, and the correlations were 0.92 for human and 0.93 for mouse. Across all human UTRs, 15% of miRNA target sites are conserved between human, mouse, rat, and dog, and of these conserved sites, about half that number are conserved above background expectation (*S5*). The discrepancy between total numbers of observed and expected targets for real miRNA families in one

genome was primarily accounted for by the ~7.5% of targets with sites conserved above background.

The choice of a trinucleotide model over mononucleotide and dinucleotide models was based on the empirical performance of each model in estimating the observed number of targets matching each miRNA in one genome. To illustrate the higher-order effects captured by the trinucleotide model over lower-order models, scatterplots displaying the estimated and observed numbers of targets for each miRNA family in mouse are shown in fig. S5C.

To calculate the expected probability of a given miRNA matching a conserved site in a UTR, we calculated the probability of the miRNA matching a single 7mer in an analogous manner (the nucleotide frequencies were obtained from the human or mouse annotated UTR used as the basis for the alignment), but used the actual number of conserved 7mers in the aligned UTR, instead of the UTR length:

```
P(UTR_contains_conserved_7mer)
= 1 - P(UTR_does_not_contain_conserved_7mer)
= 1 - (1-P(7nt_match))^(number_of_aligned_conserved_7mers)
```

Dinucleotide and trinucleotide probabilities were estimated from the dinucleotide and trinucleotide frequencies for the entire UTR sequence, as opposed to using just conserved regions, because human UTRs are, on average, ~1000 nucleotides in length, while the number of conserved 7mers averages only ~70, too short for accurately estimating trinucleotide frequencies.

In contrast to the results in one genome, where the expected and observed numbers of targets correlated well, there was a marked enrichment for conserved targets of real miRNAs over expected. When considering all 73 vertebrate families, the ratio of the observed number of conserved targets to the expected number was 2.0 : 1, whereas the ratio for the controls was 0.93 : 1. The signal-to-noise values for each individual miRNA obtained via this approach approximated those obtained from TargetscanS (*S5*).

Gene Density Maps (Figs. 1, 3, S2, S5, and S6) A stepwise construction of a gene density map from Fig. 1A is illustrated in fig. S2A.

For each miRNA, the Observed targets gene-density map reflected the actual distribution of genes with sites matching the miRNA, based on the target finding approaches discussed above, whereas the Expected targets (or Background) map reflected the expected distribution of genes with sites matching the miRNA, based on properties of their UTRs (length, conservation, and trinucleotide composition) that influence the likelihood of a match occurring by chance.

To calculate the Observed gene density map, the position of each of the genes targeted by the miRNA was assigned in accordance with its absolute expression (*x*-axis) and its relative expression (*y*-axis) as illustrated (fig. S2A). Maps were smoothed using a squared Euclidean kernel function, with each target gene contributing a density of $1/(r^2 + k)$ to each cell on the heatmap, where r^2 was the squared Euclidean distance between the coordinates of the cell and the coordinates of the target gene, and k was a constant smoothing factor. The relatively large values for the smoothing constant k (61×0.4 for the mouse atlas, 24×0.2 for the C2C12 time course) were necessary for effective

visualization, because a typical miRNA has more than an order of magnitude fewer conserved targets than the total number of cells on the density map. All density maps were normalized to a mean density of 1.0 (green), and the colors represent positive (red) and negative (blue) deviations from mean density.

The Expected gene density maps were calculated analogously, using all genes and scaling the contribution of each gene by its fractional expected probability of matching the miRNA by chance. These expected probabilities were calculated as described above, and take into account the influence of UTR length (for the analysis of nonconserved targets), number of conserved 7mers (for the analysis of conserved targets), and trinucleotide composition (for the analysis of both nonconserved and conserved targets.)

The differences between the density maps (displayed in Figs. 1 and 3 in the main text) were calculated by subtraction of the density of the Expected (or Background) map from the density of the Observed map at each of the 61 x 61 cells in the two density maps (fig. S2 and S5). Local differences in density (both in the original and subtracted density maps) indicate differences relative to the mean density. The subtracted density maps are not sensitive to the order in which the smoothing function and the density subtraction were applied, because the density in each cell is the result of summing the contributions of each gene, applied using the kernel function.

For the analysis of conserved targets (Fig. 1 and fig. S2), instead of normalizing both the Observed and Background gene density maps to have the same mean density, the Background maps were scaled to a reduced intensity, based on the signal-to-noise of the miRNA, with the intent of subtracting out the density contributed by spurious targets conserved due to chance. Signal-to-noise was calculated as the number of observed conserved targets divided by the number of expected conserved targets (described in the previous section on calculating background). The signal-to-noise values for each of the six miRNAs used in the analysis were as follows: miR-133 (3.0), miR-1 (2.8), miR-122 (0.8), miR-142 (2.1) miR-9 (4.4), miR-124 (4.2). Because the signal-to-noise was below 1 for miR-122, the Background map for miR-122 was normalized to have the same mean density as the Observed map, as if signal-to-noise was 1.0. The other five subtracted gene density maps shown in Figure 1A reflected the distribution of predicted targets above estimated noise.

For the analyses of nonconserved targets (Figs. 3, S5, S6), both the Observed and Expected gene density maps were normalized to the same mean density. Thus, the subtracted maps reflected the relative differences in distribution between observed and expected nonconserved targets.

We caution that the gene density maps are strictly a visualization tool, and we employ them for the purpose of displaying general trends. The construction of the gene density map in figure 2A shows, for instance, that although the conserved targets of miR-133 tend to cluster in the lower right corner, miR-133 targets can be present in any region of the gene density map. In particular, all quantitative results and tests of statistical significance are derived directly from the absolute and relative bin indices of genes, without using the gene density map.

Modified Kolmogorov-Smirnoff Test To characterize a signal for a particular miRNA in a particular tissue, we first assigned each gene to one of 61 equal-sized bins based on its relative expression in that tissue. This was the same procedure used to construct the

gene density maps, except that genes were only binned along the relative axis. Genes with low relative expression were placed in low-numbered bins, while genes with high relative expression were placed in high-numbered bins. Each bin contained the same number of genes from the entire gene set, while the actual number of targeted genes in each bin varied.

Because different tissues preferentially express genes with different UTR lengths and trinucleotide compositions, and both of these variables affect the likelihood of matching to a particular 7-nt sequence, it was necessary to correct for these effects by estimating the expected number of target genes in each bin. For each gene in the entire gene set, we estimated the probability of targeting by a miRNA as a function of the gene's UTR length and trinucleotide composition (see above). Summing the probabilities in each bin gave the expected number of targets in each bin. This expected distribution was then normalized, setting the total number of expected targets summed across all bins equal to the total number of observed targets. Because the numbers of observed and expected targets were usually approximately equal, normalization amounted to multiplying by a number near 1.0 in most cases.

We used a modified Kolmogorov-Smirnoff test to compare expected and observed distributions (*S6*). Figure 3A (top panels) shows the results of subtracting the number of expected targets in each bin from the number of observed targets in each bin. The one-sided discrete Kolmogorv-Smirnoff test statistic was calculated by taking the running tally of the difference in each bin, across the entire distribution (Fig. 3A, red line in bottom panel) and using the largest cumulative negative difference as the KS test statistic. The negative displacement from zero on the *y*-axis in Figure 3A indicated the number of genes that were overrepresented on the left side of the distribution. To move the KS statistic back to zero, a corresponding number of targeted genes would have to be shifted from the left side of the distribution to the right side of the distribution.

To assess the significance of KS test statistic values, we counted the number of genes targeted, and selected an equal number of genes from the entire gene set as a control cohort. Genes targeted by the miRNA were allowed to be selected for the control set. The likelihood of a gene being selected for the control set was proportional to its probability of matching the miRNA by chance, as a function of its UTR length and trinucleotide composition. For each miRNA, we generated 1,000 control cohorts and obtained discrete KS test statistics for these controls (for ease of visualization only 100 control cohorts are shown in Figure 4A.) We used the KS test statistic values of the 1,000 control cohorts to build an empirical background distribution, from which a p-value for the KS test statistic value of the actual miRNA was determined. For KS test statistics beyond the 98th percentile of the empirical distribution (i.e., those more significant than 980 of the 1000 controls), there were insufficient numbers of controls to accurately estimate *P* values. For each of these, we fit the asymptotic KS test statistic tail probability $Q = e^{-2nx\wedge 2}$ (*S7*), where *x* is the value of the KS test statistic, to the tail of our empirical distribution to derive an approximate *P* value for the miRNA-tissue pair. The parameter *n* was estimated by finding the value of *n* at which the 98th percentile of the theoretical tail probability matched that of the empirical distribution.

Using the set of 73 miRNA families conserved in mammals and zebrafish, and the 61 tissues of the mouse atlas, we performed 73 x 61 KS tests, i.e., each test involved comparing a particular miRNA-tissue pair to 1,000 control cohorts (fig. S7; a partial,

clustered version is displayed in Fig. 3B). In general, we observed significant KS-test values for miRNA-tissue pairs when the miRNA is expressed specifically and strongly in that tissue. We caution, however, that a significant signal for a particular miRNA-tissue pair does not always indicate that the miRNA is expressed in that particular tissue. For instance, miR-10 may be expressed in spinal cord neurons, which express many genes in common with cerebral cortex neurons, perhaps explaining the signal seen for a wide range of brain regions. Selection against acquiring a target site in these genes would be apparent beyond that specific cell type to cell types sharing similar expression profiles. Conversely, lack of a signal in a particular tissue may be due to heterogeneity in the tissue, because the genes most highly expressed in the tissue may come from cells in which the miRNA is absent. Heterogenous tissues such as lung, prostate, etc., gave weak signals for all miRNAs.

Estimation of Selective Avoidance The 73 miRNA families in our analysis had an average of 1,087 human and 1,050 mouse targets among our set of 8,551 genes, with an average overlap of 367 genes that were shared among the two species. To calculate the number of genes avoiding target sites during evolution, we considered the genes that were above median in terms of both relative and absolute expression in the tissue of miRNA expression (genes falling in the upper right quadrant of the gene density map, Fig. 3B.) We calculated both the observed and expected number of targets for this set of genes, in each case considering only the subset of genes from the total set which were not targeted by that miRNA in mouse. Although the sites were in human and not mouse UTRs, it was the mouse (not human) expression data that was used to determine which genes fell into the upper right quadrant, effectively preventing direct mRNA-destabilizing effects from contributing to the signal. The difference between the expected and observed numbers of targets represented our estimate for the number of genes affected by selective avoidance. The probability for a gene to contain a site matching the miRNA was calculated as described in the previous section (Background Estimation), and was dependent on UTR length and trinucleotide composition. Because the expected and observed numbers of genes for the total gene set differed slightly, we scaled the expected numbers of genes so that the total number of expected targets would equal the total number of observed targets when considering all 8,551 genes in the gene set. The signal for selective avoidance would therefore be constituted by the relative depletion of genes with sites within the upper right quadrant of the gene density map in Figure 3B (i.e. those genes which were expressed both strongly and specifically in this tissue.)

For example, to estimate the number of genes avoiding miR-133 sites in skeletal muscle, we considered the 2,661 genes of the mouse expression atlas that were above median in both relative and absolute expression. Of these genes, 207 were targets in mouse and thus were excluded from the analysis; such genes could potentially show confounding effects due to direct miRNA-mediated mRNA degradation. Of the remaining 2,454 genes, we observed that 156 had miR-133 target sites in their orthologous human UTRs, whereas we would have expected 188.4 genes to be targeted, based on the expected probabilities calculated for their UTR length and trinucleotide composition. Dividing these two numbers, this estimates that in the upper right quadrant, ~17.2% of genes are avoiding miR-133 target sites. Given that there are 2,454

genes in the quadrant, we conclude that ~420 genes expressed in muscle are under selection to avoid 7-nt miR-133 target sites.

Estimation of the Extent of Target Depletion To estimate the maximal extent of target depletion, we started with the mouse-only nonconserved analysis (Fig. 3A), and focused on the subset of genes that were most highly and specifically expressed in tissues with our six miRNAs (defined as genes that were in the top ten percent of genes in both relative expression rank and absolute expression rank.) We calculated the expected number of nonconserved targets among this subset, and compared this figure to the actual number of nonconserved targets in the subset, normalizing the total number of expected targets to reflect the total number of observed targets when summed across all genes. For miR-133 there were 403 genes that were both highly and specifically expressed in muscle, and of these, we observed 10 targets while expecting 24.3, giving us a depletion of 59%. The other microRNAs had depletion percentages as follows: miR-1, 43%; miR-122, 57%; miR-142, 54%; miR-9, 42%, and miR-124, 31%.

We chose only a small subset of genes that were highly expressed in both absolute and relative terms in order to account for the possibility of mRNA degradation effects. For messages in which miRNA targeting might cause mRNA degradation, both the absolute and relative ranks of the gene would be reduced. The result would be an effect in which targets were shifted incrementally to the lower left corner of the gene density map, and therefore looking in the middle of the map would be misleading, because target depletion would be partially obscured by genes of formerly higher rank cascading down. Thus, to quantify the maximal extent of target depletion, we looked at the genes that were most highly and specifically expressed, because there was no possibility of other genes falling into that region during such a cascade. The numbers we derived represent lower limits for the percentage of 7-mer sites responsive to highly expressed miRNAs because some targets may be translationally repressed with little or no changes in mRNA levels and insufficient time or selective pressure for site loss.

Combining Mouse and Human UTR Information Because of the noise in performing sequence analysis in one genome, we incorporated both human and mouse sequence information in our analysis for Figure 4, counting a gene to be a miRNA target if it contained a 7-nt target site in its UTR in either species. The expected probability of a gene being targeted by a miRNA was calculated as above, but trinucleotide frequencies were tallied only after the UTRs were filtered so that long runs of conserved sequence (stretches of 7 or more nucleotides conserved in four genomes) were counted only once.

The total expected number of UTRs targeted by all miRNA families was 98% of the actual observed number of targets in the mouse + human analysis. This compared to 93% in mouse and 95% in human. The Pearson correlation between the number of targets for each individual miRNA in the expected and observed sets was 0.93 for the human set, 0.93 for the mouse set, and 0.95 for the combined mouse and human set.

When performing the human-only and mouse-only nonconserved analyses in Figures 3, S5, and S6, we excluded genes from the analysis that were targeted in the other species. In cases where genes had highly conserved UTRs, this meant that if the gene lacked a site in one species, it had a substantially reduced likelihood of having that site in the other species. To account for this, we did not tally nucleotide counts from long runs

of conserved sequence (again, defined as stretches of 7 or more conserved nucleotides) for purposes of determining UTR length and nucleotide frequencies in both the human-only and mouse-only nonconserved analyses.

C2C12 Myotube Differentiation Timecourse

The expression data used in the analysis of C2C12 murine myoblast cell line differentiation (*S8*) consisted of 24 individual microarray experiments hybridized to the Affymetrix U74Av2 chip, reflecting eight time points assayed in triplicate. The first three time points (days -2, -1, 0) reflect gene expression prior to the onset of differentiation, and the latter five time points (days 2, 4, 6, 8, 10) reflect gene expression during the course of differentiation. The U74Cv2 chip data was not incorporated into our analysis, because it was missing the three experiments at the day 8 time point. The data were normalized using Affymetrix Microarray Suite 5.0 (MAS5) software, and we selected 4,965 genes for our analysis, using the same criteria as we followed in choosing genes for the main analysis.

Because of the smaller number of samples, we treated each of the 24 individual microarray experiments as its own separate sample for the gene density map analysis shown in Figure 1B (i.e, we did not merge the triplicate experiments). The gene density maps for miR-1, miR-133, and let-7 conserved targets were otherwise constructed in the same manner as in the main analysis, producing smaller maps of size 24 X 24 instead of 61 X 61. For each time point, the gene density maps from the three triplicates were averaged to produce the composite maps shown in Figure 1B. miR-1 and miR-133 are two muscle-specific miRNAs that accumulate beginning at day 0 and increase over the course of C2C12 differentiation (*S8*). The non-muscle-specific microRNA let-7 is shown alongside for comparison.

The mean change in the expression levels of the miRNA targets over the course of differentiation was calculated (fig. S3) as the geometric mean of the targets before differentiation (days -2, -1, 0) divided by the geometric mean of the targets after differentiation (days 8, 10); only genes that were expressed above median both before and after differentiation were included in the calculation. For each miRNA family with at least 100 conserved targets among the 4,965 genes, we calculated the mean change in expression levels of their targets, and found that miR-1 and miR-133 targets decreased by the greatest magnitude, with miR-1 targets decreasing an average of 23%, and miR-133 targets decreasing an average of 16% (fig. S3D.) In comparison, the typical decrease in expression for conserved targets of other miRNAs was ~5%, which we attribute to a propensity for differentiated myotubes to express genes with shorter, less well-conserved UTRs compared to undifferentiated myoblasts.

To evaluate the significance of the decrease in mean expression observed for miR-1 and miR-133 conserved targets, we repeated our analysis with 10,000 control cohorts for each miRNA, in a manner analogous to the modified Kolmogorov-Smirnoff test. We counted the number of genes targeted and selected an equal number of genes for each control cohort. Genes were randomly selected to populate the cohort based on their probability of having a conserved site to the miRNA by chance (see Background Estimation.) Only genes that were expressed above median both before and after differentiation were included in the analysis. For each miRNA, we derived an empirical background distribution describing the mean expression change due to chance, and used

it to estimate a P-value for the observed decrease in the mean expression of the miRNA's targets. The P-value for miR-1 was < 0.0001, indicating that the decrease in expression of miR-1 targets was more significant than all 10000 control cohorts, while the P-value for miR-133 was 0.0100. The conserved targets of the other miRNAs were not significantly downregulated; the next most significant were the targets of miR-24, with P-value of 0.2524.

We also extended our analysis to nonconserved targets, calculating the mean decrease in expression in the same manner as for the conserved targets, and the using control cohorts to evaluate significance. Nonconserved targets of miR-1 and miR-133 decreased an average of 7% and 8%, respectively, neither of which were significant, due to the general tendency of genes with longer UTRs to be expressed at lower levels in myotubes compared to myoblasts. Nonconserved targets of the other miRNAs also were not significantly downregulated.

While the decrease in expression of the conserved targets of miR-1 and miR-133 is highly significant, the significance comes from the consistency with which each target gene is downregulated, as opposed to large changes of two-fold or more. Hence, lists of genes with the largest foldchanges in expression have little overlap with the conserved targets of miR-1 and miR-133.

Secondary Structures Flanking Conserved and Nonconserved Sites 79 mammalian miRNA families (*S9*) were searched against a database of multiz alignments of 3'UTR sequences constructed by identifying the annotated 3'UTR regions for 22383 RefSeq mRNAs (*S10*) mapped by the UCSC genome browser [(*S2*); genome.cse.ucsc.edu]. 7- and 8-nt sites conserved in human/mouse/rat/dog/chicken containing Watson-Crick pairing to bases 2-7 of the miRNA supplemented by either or both a Watson-Crick match to base 8 or an adenosine across from position 1 of the miRNA were identified. Those 7- and 8-nt sites found in human 3'UTR sequence but not observed to be conserved in the 5-vertebrate alignments were collected and included as the nonconserved set. In addition, a control set consisting of 4 cohorts corresponding to each of the 79 miRNA sequences were searched against the alignments and conserved and nonconserved sets were collected. Sets of control sequences were constructed for the 79 miRNA families so as to preserve properties affecting the likelihood of finding a match and score in a single genome using the TargetScan algorithm (*S9*). Notably, these control sequences preserved the predicted free energies associated with pairing to the miRNA seed region.

Zhao et al. (*S11*) report that the predicted secondary structures of sequences immediately flanking authentic miRNA binding sites have significantly less predicted stability than do average 3'UTR fragments. To explore this claim that authentic targets might be associated with less stable predicted secondary structures in mRNAs, we evaluated the predicted folding energies of sequences surrounding sites found to be conserved in 3'UTR alignments of 5 vertebrates and the remaining sites found in human. To enable evaluation of predicted folding energies of sequences flanking the 7- or 8-nt matches, only those sites located >70 nt downstream of the 3'UTR start in human and those sequences located >70 nt upstream of the 3'UTR terminus in human were included in the analysis. These 70-nt fragments were folded using the RNAfold routine from the Vienna RNA package (*S12*) and the average folding free energy of the upstream and downstream fragments were calculated for both the 5-vertebrate conserved set and the

nonconserved set. Results obtained when examining regions flanking conserved sites corresponding to real miRNAs were indistinguishable from those for the conserved sites corresponding to control cohort sequences.

These sets of sites also were evaluated for accessibility using a method resembling one used to predict miRNA target sites in *Drosophila* 3'UTRs (*S13*). A 100-nt region surrounding each site was folded using RNAfold from the Vienna RNA package (*S12*) and the predicted local structure at the 7- or 8-nt sites was searched for sequences of three consecutive unpaired bases. The set of conserved sites in 5 vertebrates was enriched for open structure relative to the set of nonconserved sites. However, the results obtained when identifying sites for sets of control cohort sequences that preserve the binding free energy of the seed region (in addition to properties affecting the likelihood of finding a target site in human 3'UTRs) were highly similar to those found for the real miRNAs. In summary, our analyses indicate that non-occlusive secondary structure, as measured by previously reported algorithms (*S11, S13*), does not influence miRNA-directed targeting in mammals.

Selection of predicted targets and nonconserved cohorts for reporter assays

Nonconserved TargetScan-like targets were selected randomly from human 3'UTR sequences that had human TargetScan scores within the range of those of miR-1 predicted targets with experimental support (*S9*), but were not TargetScan predictions because they did not score above the cutoffs in mouse or rat. Four of the six were not scored in mouse or rat because they lacked seed matches in the orthologous mouse or rat UTR. The other two (N1 and N3) were scored in both mouse and rat but had scores below the cutoffs.

TargetScanS predictions were randomly selected from a list of predicted human targets (*S5*) that had exactly two sites (7- or 8-nt matches to the seed region) in the 3'UTR, which were within ~1 kb of each other. Nonconserved TargetScanS-like targets were selected to resemble the TargetScanS predictions in human, in that they had two sites in the human 3'UTR, which fell within orthologous aligned regions of the human, mouse, rat and dog genomes. However, the aligned segments corresponding to both human sites were diverged to include mismatched nucleotides in mouse, rat or dog sequences. In all cases, both sites were concurrently disrupted in at least one of the mouse, rat or dog orthologous sequences. In two of 17 cases (N18 and N19), there was an additional non-aligned site within the 3'UTR of the ortholog lacking the two aligned sites. However, these two cases do not bias our interpretation, because only one of the two mediated repression.

To simplify the experimental analysis, we choose to examine UTR fragments with two matches to the same miRNA family, even though most UTRs with a conserved match to a miRNA family do not have a second conserved match to the same miRNA family (*S5*). This simplification was justified based on the observation that UTRs with a conserved match to one miRNA usually have a second conserved match to a second miRNA (*S5*), and in cells in which both miRNAs are expressed the repression presumably would be equivalent to that observed with two matches to the same miRNA (*S14*). The same would be true for nonconserved matches, in that more than 90% of UTRs have nonconserved matches to multiple miRNAs. One concern was that UTRs

with two nonconserved sites to the same miRNA might be more likely to be important species-specific targets. To address this possibility, we investigated whether such UTRs occur more or less frequently than would be expected by chance, comparing UTRs containing multiple matches to miRNAs with the number containing multiple matches to control sequences of similar overall abundance that do not match miRNAs. For both control sequences and miRNA matches, we plotted the number of UTRs with one match versus the number with more than one match, and found that the double-site UTRs occurred in the same relative proportion for the control sequences as for miR-1 and miR-124. Therefore, UTRs with multiple miRNA matches occur as frequently as expected by chance, indicating that there is no strong selection for or against multiple occurrences over selection acting on single occurrences.

Renilla reporter plasmids were constructed by insertion of PCR-amplified 3'UTR fragments into a p-RL-SV40-derived vector (Promega). Mutant derivatives were constructed by QuikChange site-directed mutagenesis (Stratagene). Insert sequences were confirmed by sequencing and are provided in Table S1.

Transfection and luciferase assays HeLa cells were transfected using Lipofectamine 2000 (Invitrogen) in 24-well plates (~0.5 x 10^5 cells / well) with 25 ng firefly luciferase control reporter (pIS0, (*S9*)), 10 ng *Renilla* luciferase reporter, 1.25 μg pUC19 and an appropriate amount of miRNA duplex. miR-1 duplex comprised oligonucleotides: 5'-UGGAAUGUAAAGAAGUAUGUA-3' and 5'-CAUACUUCUUUACAUUCAAUA-3'; miR-124 duplex comprised: 5'-UAAGGCACGCGGUGAAUGCCA-3' and 5'-GCAUUCACCGCGUGCCUUAAU-3'. Firefly and *Renilla* luciferase activities were measured 24 hours after transfection with the Dual-luciferase assay (Promega). *Renilla* activity was normalized to firefly activity to control for transfection efficiency. Values plotted in Fig. 1 are geometric means of replicate values.

Northern blotting For Fig. 4D, 9 μg total RNA was loaded per lane. All tissues were from rat. RNA from cortex, cerebellum, spinal cord, hippocampus, hypothalamus, pituitary, olfactory epithelium and dorsal root ganglia was purchased from Analytical Biological Services (Wilmington, DE); RNA from heart, liver and spleen was purchased from Ambion (Austin, TX). Northern blotting was performed as described previously [(*S15*); http://web.wi.mit.edu/bartel/pub/protocols/], with the following DNA oligo probes:

miR-7a: CAACAAAATCACTAGTCTTCCA
miR-7b: AACAAAATCACAAGTCTTCCA
miR-124: TGGCATTCACCGCGTGCCTTAA
U6 snRNA: TTGCGTGTCATCCTTGCGCAGG

miR-7 was probed with a combination of miR-7a and miR-7b probes. For comparison, the blot was stripped and reprobed for miR-124, then U6 snRNA.

Supporting Online Text

Specificity determents beyond 7- or 8-nt matches to miRNA seed regions

We present experimental and computational evidence indicating that determinants outside the matches to the seed region play only small roles in specifying targeting. Additional specificity determinants proposed previously include pairing to the 3' portion of the miRNA, mRNA secondary structure that could occlude miRNA pairing, and sites for RNA-binding proteins that might occlude or recruit the miRNA-programmed silencing complex. Although early computational analyses, including ours, assumed that pairing to the 3' portion of mammalian miRNAs would usually provide added specificity (*S9, S16*), more recent studies suggest that 3' pairing is primarily important for sites that do not have perfect seed matches, which appear to be much less prevalent than those that have perfect matches (*S5, S17, S18*). In contrast to the emerging consensus regarding the role of 3' pairing, two recent studies have suggested that non-occlusive mRNA secondary structure is a major determinant of targeting (*S11, S13*). To investigate this possibility, we implemented these two algorithms with the idea that if they successfully identified non-occlusive secondary structure important for targeting that the conserved sites would preferentially fall into these open areas when compared to the nonconserved sites (Materials and Methods). A slight preference for conserved sights was detected, but with further investigation this preference did not appear to be associated with miRNA targeting, in that it persisted when we replaced the miRNAs with non-miRNA control sequences. We conclude that, mRNA secondary structure, as measured by previously reported algorithms (*S11, S13*), does not perceptively influence miRNA-directed targeting in mammals, when considering results summed over all targets.

Although our data support the idea that overall specificity determinants outside the 7- or 8-nt match to the seed region overall play relatively minor roles in specifying targeting by highly expressed miRNAs, for any individual site under study they might still play important roles. Although messages preferentially co-expressed with miRNAs are generally ~50% depleted in miRNA matches, they are not totally depleted, and only 13/17 fragments with nonconserved TargetScanS-like sites mediated repression. We speculate that some of the non-responsive sites are in regions of the UTR made less accessible by secondary structure or protein binding. Furthermore, determinants outside the 7- or 8-nt match might explain some of the variability in repression observed in the reporter assay, ranging from <1.3-fold to 4-fold.

Number of genes avoiding miRNA targeting

The finding that ~170 to ~440 genes of the mouse expression atlas appear to be selectively avoiding targeting to the six miRNA families of Figure 3 raised the question of how many genes of the atlas are avoiding targeting to all miRNAs. This question cannot be addressed by simply summing the antitargets of each miRNA family because the antitargets of one miRNA might be the same as those of another. For example, the antitargets of miR-133 would be expected to largely overlap with those of miR-1 because these two miRNAs are expressed in the same tissues. However, the antitargets for miRNAs expressed in different tissues cannot all overlap; when estimating the number of genes selectively avoiding targeting by a miRNA, we use relative rank, and a gene cannot be relatively high in all tissues. When considering that a single miRNA can have more

than 400 antitargets and that there are numerous tissues and cell types that specifically express miRNAs, we estimate that the aggregate number of antitargets is on the order of thousands, not hundreds. Note that an antitarget of one miRNA family can be a conserved target of another family and a nonconserved target of a third family.

Signatures for sites in coding sequence

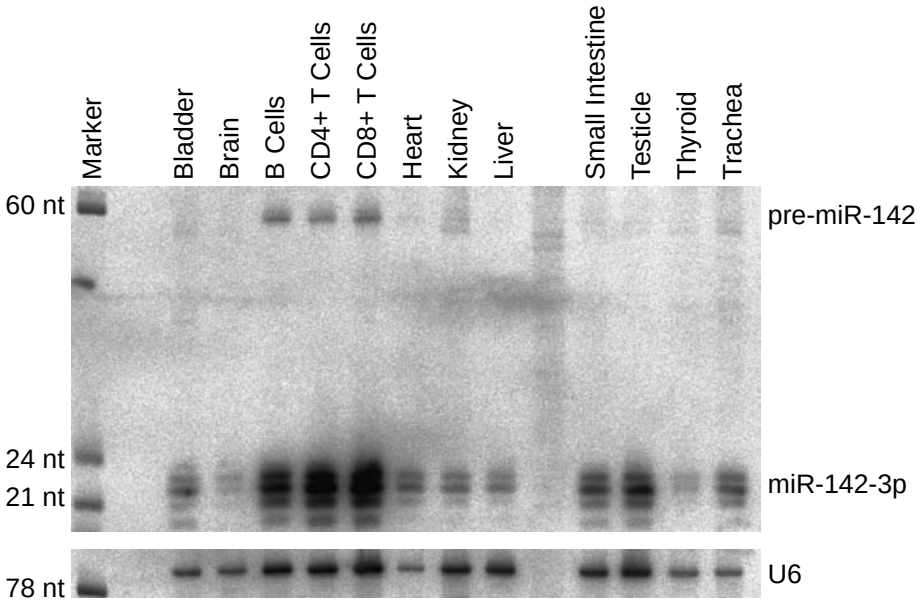
To examine targeting within the coding sequence, we repeated the analysis of Figure 4B using coding sequences rather than 3'UTRs. Some miRNA families, including let-7, miR-9, miR-29, miR-122, miR-124, miR-142, miR-133, miR-125 gave robust, accurate signatures, but the P-values were generally less significant than for the 3'UTR analysis, and other miRNAs gave spurious signatures. We conclude that substantial targeting involving perfect seed matches occurs in coding sequence but that 3'UTRs are more hospitable for targeting. This result agrees with previous target-prediction analysis, which showed that 8-nt sites within coding sequences were under selective pressure to preserve miRNA pairing but that this signal for conserved targeting, although present in coding sequence, is highest in 3'UTRs (*S5*).

Additional considerations and implications

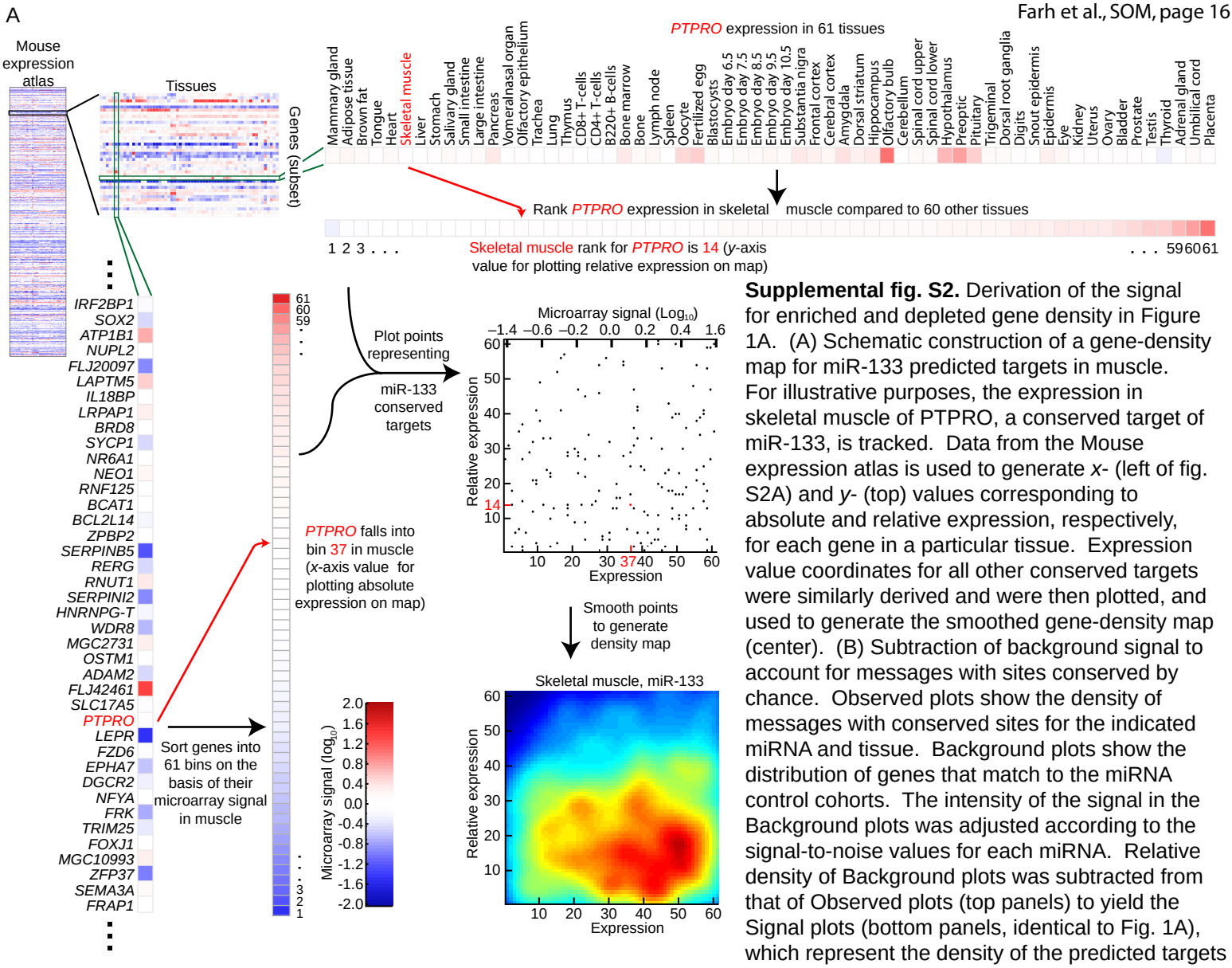
Our analyses of the expression of messages with conserved and nonconserved sites reports propensities and trends. It is important to bear in mind that some messages do not follow the dominant trends. For example, the miR-133 analysis in skeletal muscle showed a propensity of messages with conserved sites to be expressed in muscle, but expressed at lower levels in muscle compared to other tissues of the atlas (Fig. 1A). As mentioned in the text, this trend, together with the trend during myoblast differentiation, is concordant with the ideas that miRNAs often 1) dampen the output of preexisting messages to facilitate a more rapid and robust transition to a new expression program, 2) optimize protein output without eliminating it entirely, and 3) destabilize many target messages to further define tissue-specific transcript profiles. However, messages with conserved sites populate all regions of the gene density map (fig. S2A), with some having no detectable expression in muscle or myoblasts (as modeled by C2C12 cells). Because it is doubtful that all, or even most, of these sites are conserved by chance, we conclude that a sizable minority of conserved targets represent exceptions to the principles enumerated above. These include messages that are destabilized to imperceptible levels with the help of miRNA-mediated repression—a class of targets that might be particularly abundant among targets of the miR-302 family (Fig. 1C). They also include failsafe targets, which are messages that are nearly completely repressed at the transcriptional level but require miRNAs to assure fidelity of this repression (*S19*).

On the whole, analyzing the apparent imprint of miRNAs on mRNA expression and evolution revealed the spatial and temporal expression of miRNAs during mammalian development (Fig 4). This striking correspondence between the signatures and the expression patterns indicated that the signals we observed in Figures 1 and 3 have biological meaning and conversely that miRNA expression patterns have biological meaning, i.e., that miRNAs generally are active in the tissues where they are expressed and are not sequestered in an inactive form.

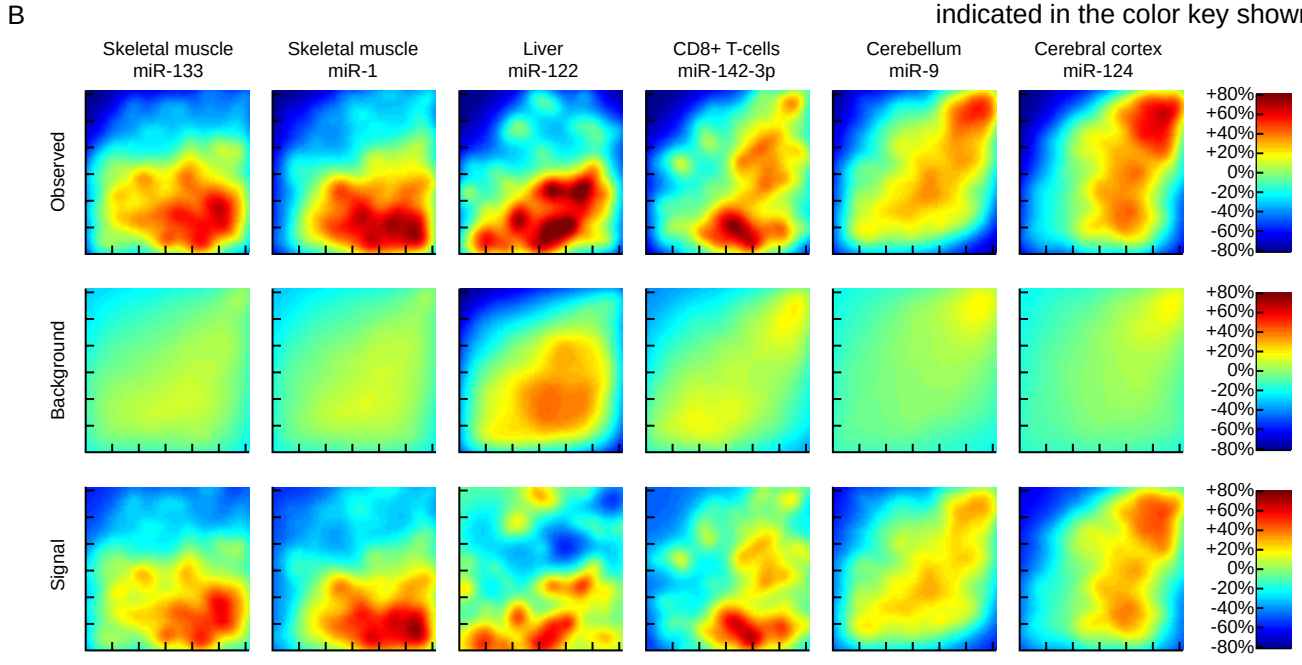
Our results have ramifications for interpreting results of reporter assays and TargetScanS—two of the main tools for studying and identifying mammalian miRNA targets. In our heterologous reporter assay, conserved sites mediate repression indistinguishable from that of nonconserved sites (Fig. 2 and fig. S4). Although not every message with conserved sites is an authentic target, and some that have nonconserved sites might be authentic species-specific targets, it is reasonable to propose that those with conserved sites are substantially enriched in biological targets compared to those with nonconserved sites. The observation that conserved and nonconserved sites mediate similar repression in the reporter system calls into question the utility of such a reporter system for distinguishing biological targets from messages with fortuitous pairing to the miRNA. Although nonconserved sites are less likely than conserved sites to be biological (in large part because they are less likely to be in mRNAs that are coexpressed with the miRNA), the abundance of nonconserved sites and our in vivo evidence for function of such sites (Fig. 3 and 4) suggest frequent nonconserved repression. Therefore, the TargetScanS predictions represent only a fraction (probably a minority) of targets repressed in the animal. A more complete list can be compiled by considering the messages coexpressed with the miRNA that have a conserved or nonconserved 7-nt TargetScanS-like site. Nonetheless, a focus on conserved predictions enriches for interactions that evolution has preserved and are therefore most likely to be consequential. Furthermore, the use of conservation as the criterion for distinguishing features of miRNA targeting from equally plausible fortuitous features has been informative for discovering general principles of miRNA targeting in the past [e.g., in defining Watson-Crick seed pairing (*S9*) and the A anchor (*S5*) as generally important for targeting] and will undoubtedly reveal additional insights in the future.

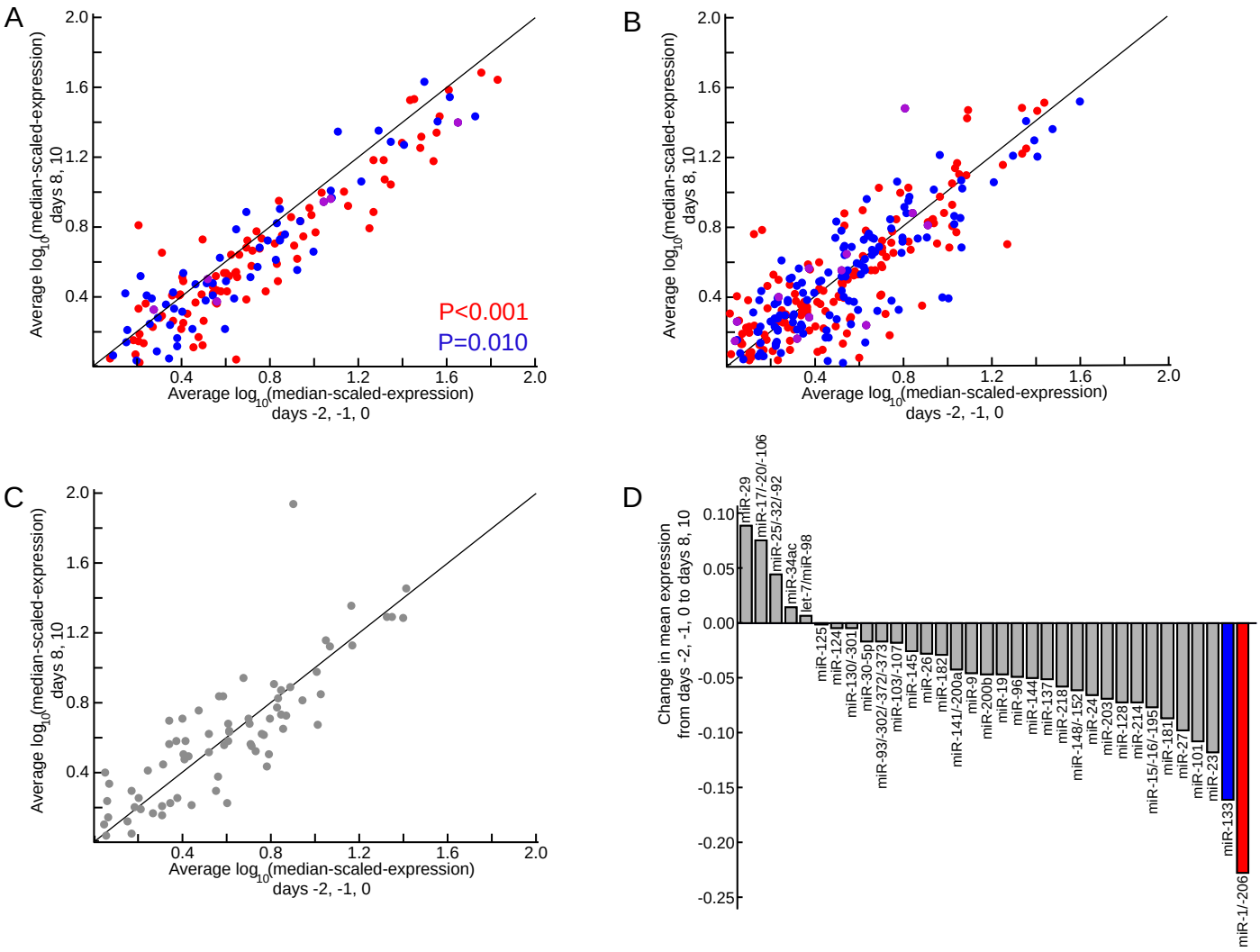


Supplemental fig. S1. miR-142-3p accumulates to high levels in human B-cells, CD4+ T cells and CD8+ T cells from peripheral blood. Detection of pre-miR-142 suggests active expression of miR-142-3p in purified, differentiated peripheral blood cells. Small RNA blotting was performed using 15 µg human total RNA per lane, the human miR-142-3p miRCURY LNA probe (Exiqon, Vedbaek, Denmark), and a U6 DNA oligo probe (TTGCGTGTCATCCTTGCGCAGG), as previously described [(S14); http://web.wi.mit.edu/bartel/pub/protocols/]. RNA from peripheral human B-cells, CD4+ T-cells, and CD8+ T-cells was purchased from AllCells (Berkeley, CA). All other samples were purchased from Ambion (Austin, TX).

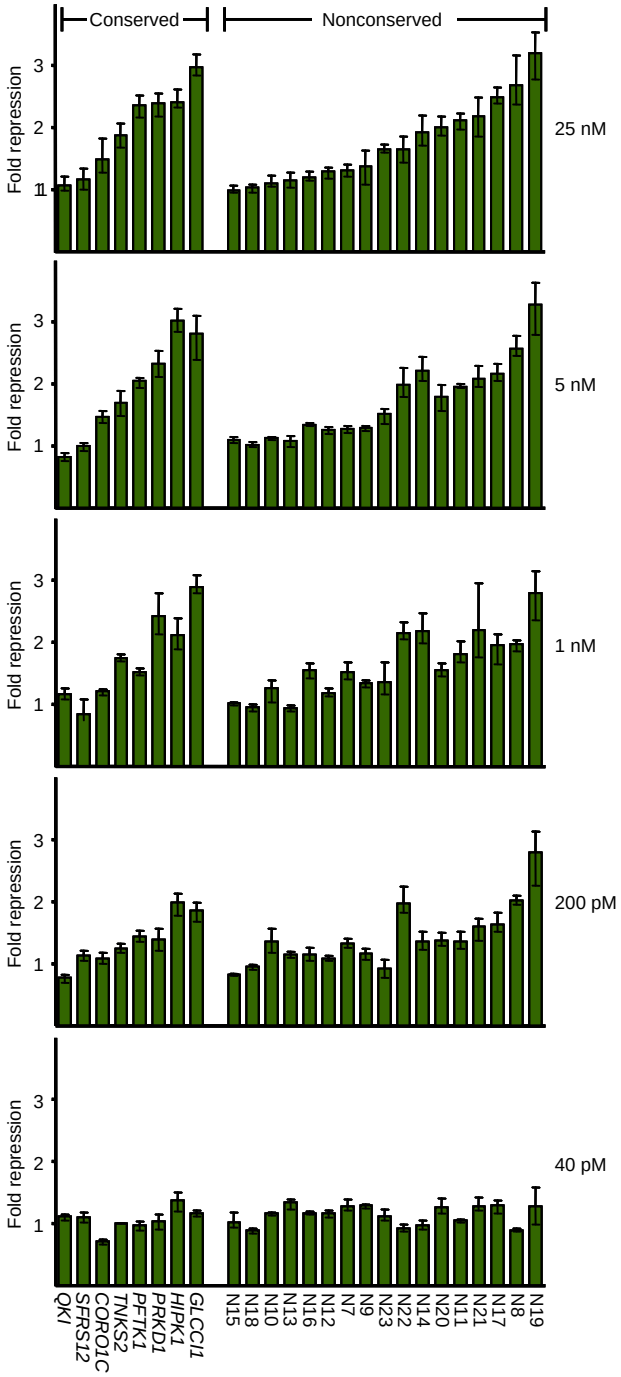


Supplemental fig. S2. Derivation of the signal for enriched and depleted gene density in Figure 1A. (A) Schematic construction of a gene-density map for miR-133 predicted targets in muscle. For illustrative purposes, the expression in skeletal muscle of PTPRO, a conserved target of miR-133, is tracked. Data from the Mouse expression atlas is used to generate *x*- (left of fig. S2A) and *y*- (top) values corresponding to absolute and relative expression, respectively, for each gene in a particular tissue. Expression value coordinates for all other conserved targets were similarly derived and were then plotted, and used to generate the smoothed gene-density map (center). (B) Subtraction of background signal to account for messages with sites conserved by chance. Observed plots show the density of messages with conserved sites for the indicated miRNA and tissue. Background plots show the distribution of genes that match to the miRNA control cohorts. The intensity of the signal in the Background plots was adjusted according to the signal-to-noise values for each miRNA. Relative density of Background plots was subtracted from that of Observed plots (top panels) to yield the Signal plots (bottom panels, identical to Fig. 1A), which represent the density of the predicted targets above noise. Red depicts local enrichment of miRNA targets, and blue depicts depletion, as indicated in the color key shown on the right.

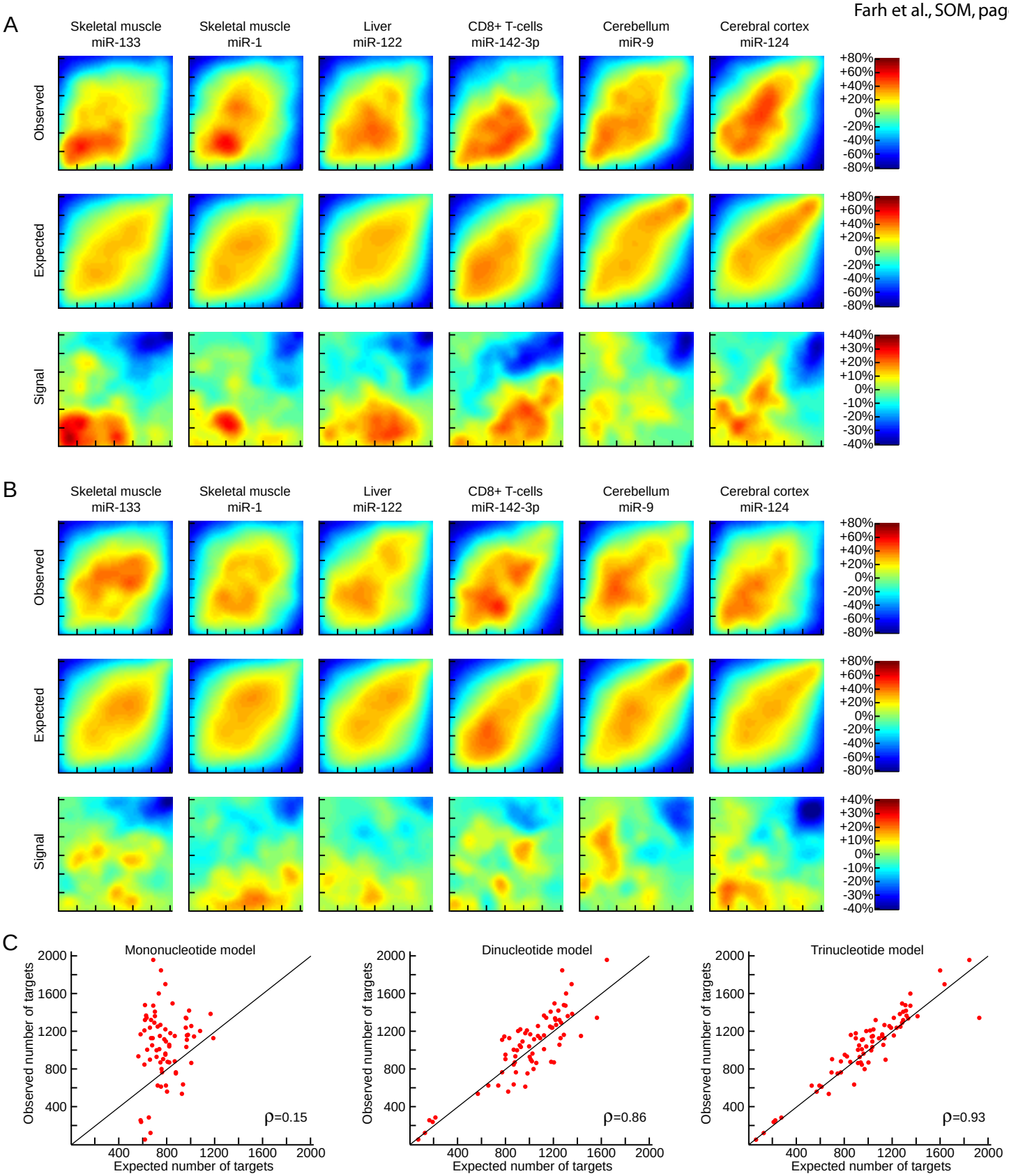




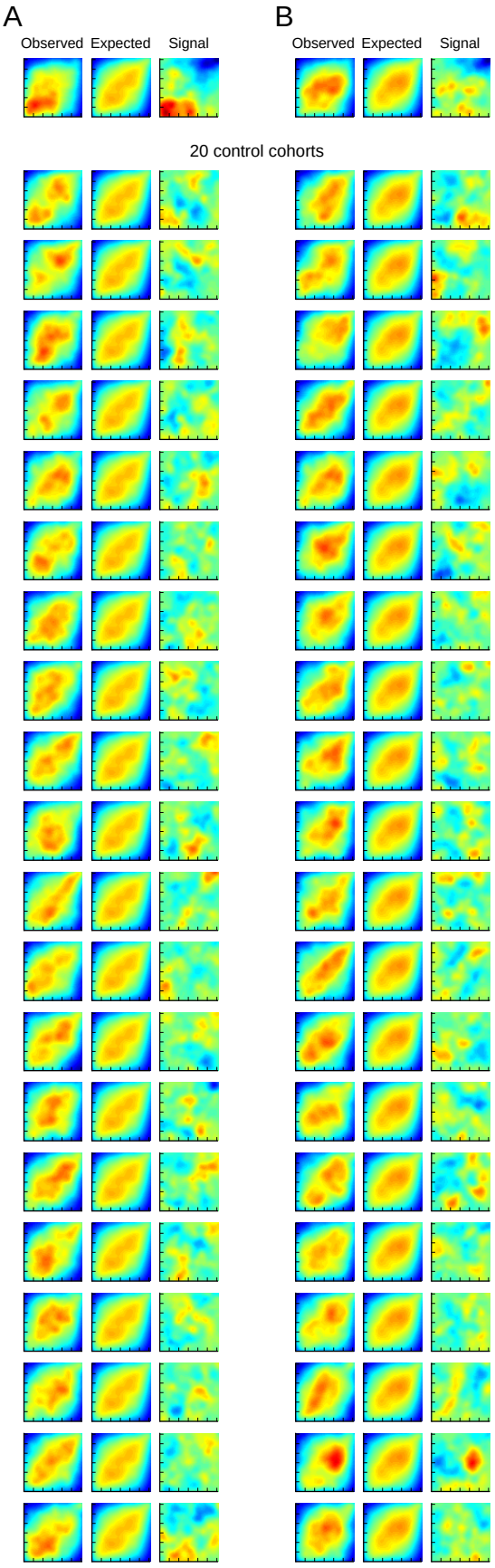
Supplemental fig. S3. Decreased expression of miR-1 and miR-133 predicted targets during myoblast differentiation. The scatter plots display genes that are expressed above median levels on the microarray both before and after differentiation. This set was enriched for conserved targets of miR-1 and miR-133 (60% were expressed both before and after) versus nonconserved targets of miR-1 and miR-133 (43% were expressed both before and after). (A) Conserved targets of miR-1 (red), miR-133 (blue) or both (purple), expressed above median both before and after differentiation. *x* axis: expression before differentiation. *y* axis: expression after differentiation. Conserved targets of miR-1 and miR-133 are consistently shifted from the midline, reflecting a decrease in their expression levels over the course of differentiation. The decrease in the expression of the 95 miR-1 targets averaged 23%, with a range of 7% to 36% at 25th and 75th percentiles. The decrease in the expression of the 62 miR-133 targets averaged 16%, with a range of -4% to 31% at 25th and 75th percentiles. Mean changes in the expression of miRNA targets were measured by the change in the geometric mean of their expression values. Because a substantial fraction of the myoblasts do not differentiate into myotubes, the decrease is expected to be greater in the subset of cells that differentiate. Bootstrap P values were derived using 10,000 control cohorts. (B) The 377 nonconserved targets of miR-1 and miR-133. Expression of these nonconserved targets decreased subtly (7% and 8%, respectively, P-values not significant) (C) Conserved let-7 targets did not markedly change. (D) Changes in mean expression of the conserved targets of 34 miRNA families. Values indicate the average change decreased over the course of differentiation. The 34 miRNA families shown are the subset of the 73 miRNA families in our analysis that had at least 100 conserved targets among the 4965 genes in the C2C12 analysis.

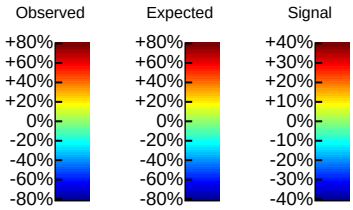


Supplemental fig. S4. MicroRNA titration of miRNA-mediated repression of reporter genes containing 3'UTR segments of target genes. All targets from Fig. 2B were re-assayed as in Fig. 2, except that miRNA concentrations were titrated to indicated levels. Luciferase values were processed as in Fig. 2, then values with cognate miRNA were normalized to those with non-cognate miRNA. Three replicates were performed (error bars represent largest and smallest values), except for the top panel, which was derived from 12 replicate values shown in Fig. 2B (error bars represent the third highest and lowest values).

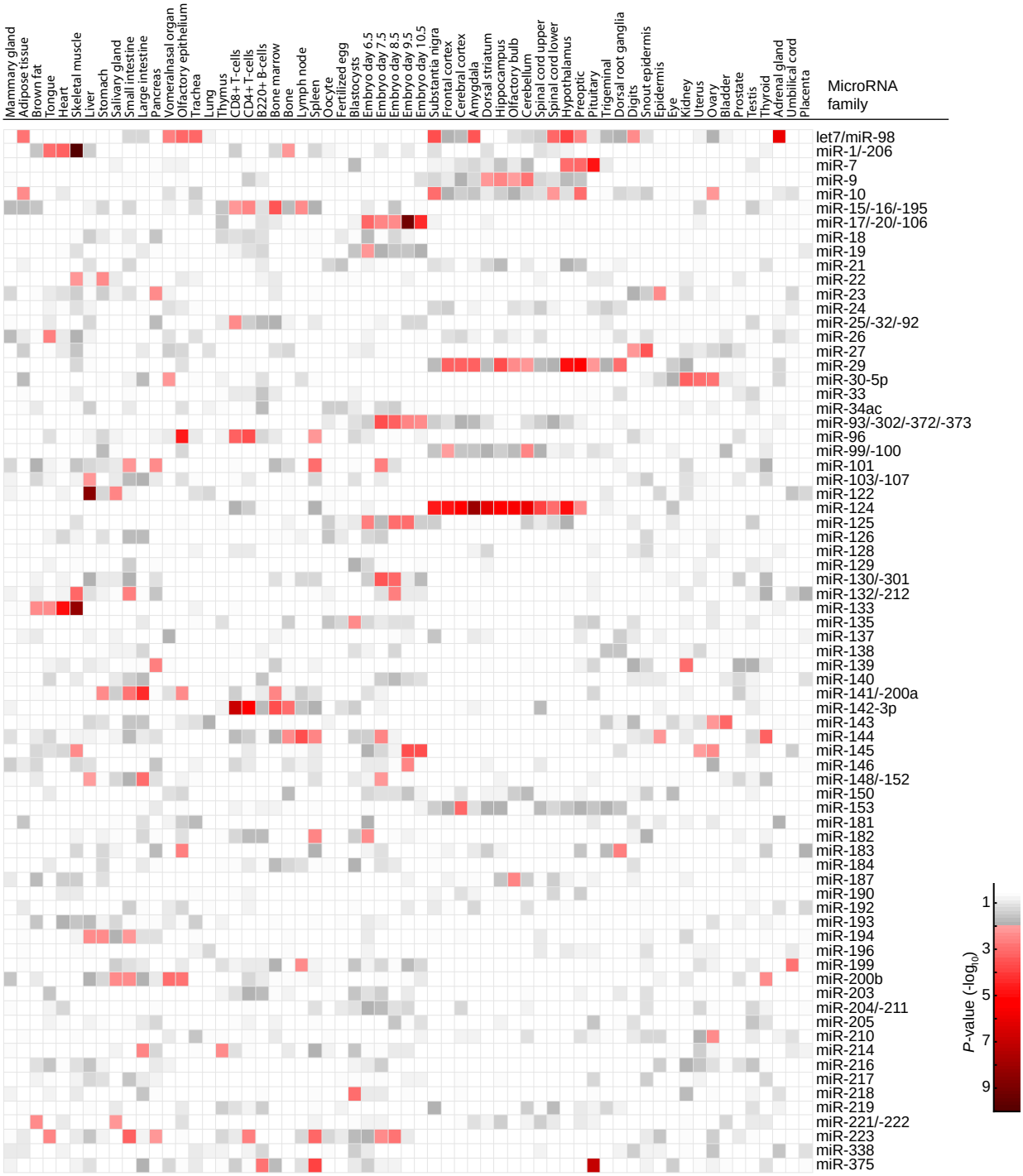


Supplemental fig. S5. Derivation of the signal for enriched and depleted gene density in Figure 3. (A) Derivation for Fig. 3A. (B) Derivation for Fig. 3B. Observed plots show the expression of messages with nonconserved sites for the indicated miRNA and tissue, generated as illustrated in figure S2A. Expected plots show the background distribution of genes likely to have a nonconserved match to the miRNA by chance. Relative density of each Expected plot was subtracted from that of the corresponding Observed plot (top panels) to yield the Signal plot (bottom panels, identical to Fig. 3), which represents the density of the nonconserved targets relative to expectation. Red depicts local enrichment of miRNA targets, and blue depicts depletion, as indicated in the color key shown on the right. (C) Comparison of the performance of mono-, di- and tri-nucleotide models for predicting the observed numbers of targets in mouse. Individual points correspond to the 73 miRNA families used for the analysis. ρ, Pearson correlation between observed and expected number of targets.





Supplemental fig. S6. Analysis of messages with nonconserved miR-133 sites in muscle (fig. S5), repeated with 20 control cohorts to illustrate the variability of the signal expected by chance. Each control cohort was populated with the same number of genes as for the Observed plot in figure S5. Genes for the control cohorts were selected according to their probability of containing the miRNA target site, on the basis of UTR length and nucleotide composition.



Supplemental fig. S7. Complete map showing KS-test P-values for each tissue-miRNA pair (61 tissues, 73 miRNAs) like that shown in Figure 4B. Darker areas denote increasingly significant values (Table S2), as indicated in color key.

Table S1. Sequences of UTR fragments inserted into luciferase reporters assayed in Figure 2 Listed is the gene identifier of the 3' UTR, its GenBank accession number, restriction sites used in cloning (5' site – 3' site), and the reporter plasmid name (in brackets). Occurrences of 7- or 8-nt matches to miRNAs are underlined. In the mutant versions of each reporter, two 7-nt matches were disrupted. For those that had a third site within the insert (G6PD, BDNF and N1), the middle site was not disrupted. To disrupt each miR-1 site, the seed match CATTCC was changed to gAaTgC (G6PD, BDNF, N1 – N6, PFTK1 and GLCCI1) or CtgaCC (CORO1C, TNKS2, N7 – N16). To disrupt each miR-124 site, the seed match TGCCTT was changed to TcggTT. Reporter plasmids were constructed in pIS2, which was derived from pRL-SV40 (Promega), and encodes Renilla luciferase. pIS2 cloning sites are shown below the insert sequences.

>G6PD NM_000402 SacI-XbaI [pAG68]
ccccgagtcgggaggactccgggaccattgacctcagctgcacattcctggccccgggc
tctggccaccctggcccgcccctcgctgctgctactacccgagcccagctacattcctc
agctgccaagcactcgagaccatcctggcccctccagaccctgcctgagcccaggagct
gagtcacctcctccactcactccagcccaacagaaggaaggaggagggcgcccattcgt
ctgtcccagagcttattggccactgggtctcactcctgagtggggccagggtgggaggg
agggacaaggggaggaaaggggcgagcacccacgtgagagaatctgcctgtggccttg
cccgccagcctcagtgccacttgacattccttgtca

>BDNF NM_170735 SacI-XbaI [pAG70]
cagtcatttgcgcacaacttaaaaagtctgcattacattccttgataatgttgtggttt
gttgccgttgccaagaactgaaaacataaaaagttaaaaaaaataataaattgcatgct
gctttaattgtgaattgataataaactgtcctctttcagaaaacagaaaaaaaacacac
acacacacaacaaaaatttgaaccaaaacattccgtttacattttagacagtaagtatc
ttcgttcttgttagtactatatctgttttactgcttttaacttctgatagcgttggaat
taaaacaatgtcaaggtgctgttgtcattgctttactggcttaggggatgggggatggg
gggtatattttgtttgttttgtgttttttttcgtttgtttgttttgttttttagttc
ccacagggagtagagatggggaaagaattcctacaatatatattctggctgataaaaga
tacatttgtatgttgtgaagatgtttgcaatatcgatcagatgactagaaagtgaataa
aaattaaggcaactgaacaaaaaaatgctcacactccacatcccgtgatgcacctccca
ggccccgctcattctttgggcgttggtcagagtaagctgcttttgacggaaggacctat
gtttgctcagaacacattctttccccccctcccccctctggtctcctctttgttttgttt
taaggaagaaaaatcagttgcgcgttctgaaatattttaccactgctgtgaacaagtga
acacattgtgtcacatcatgacactcgtataagcatggagaacagtgatttttttttag
aacagaaaacaacaaaaaataaccccaaaatgaagattatttttatgaggagtgaaca
tttgggtaaatcatggctaagcttaaaaaaaactcatggtgaggcttaacaatgtcttg
taagcaaaaggtagagccctgtatcaacccagaaacacctagatcagaacaggaatcca
cattgccagtgacatgagactgaacagccaaatggaggctatgtggagttggcattgca
tttaccggcagtgcgggaggaatttctgagtggccatcccaaggtctaggtggaggtgg
ggcatggtatttgagacattccaaaacga

>N1 NM_015318 SacI-XbaI [pAG74]
cgctgaatacagtgaacagggacattcccgccactcggggacagatgggcacaagggag
gggaaactccatcaggaagtgctcccctgggcagaggcgcccactgggtgctgtgggct
caggaggggcggggcaggagctggtgccaaccgggaaccagagccccacagccataca

gcccattggtgacaaggtcctgagaacacagtggccaggtgtccccaggctcctggccc
ctccgacgacctcaactctgcccagcccggtccctggccatcagcgacgctgtccgccc
cccgtcagatcccatgtgtgccatgtttatcatcagtgttttgtatttttgtactgagt
atcggagcactttacagaagctgactgtacattcctgttctgttgtgaagagaacattc
ccagaccc

>N2 CR601361 SacI-XbaI [pAG78]
gtacaagccattgaataagcctcttccttttttttgctcaaacattccacatccttgtg
gattcccctgcattgtttgttttatataacatttgatatttgttgtagcttgtatatga
acataattttctttagaggtagtcactgttctctccagtatgacccaggtttcttgact
ctgagtaatgcaccttctataactatctaaatttctattgaagcttttggattatgag
tatgctgacttttcacgattggctggtgcatgtttagacttaaatgtcatatccttcat
gtctcaaagccaaaatagtaacatctcatctcagaacagagctgtgaccacatgccaat
atatgtgtcacaaagtctacatatgttacattccttggaag

>N3 L40403 SacI-XbaI [pAG80]
aatcaattacatatgaaagggcattcttacattccagcagatatgtggcacagcctacc
caacatatgcataattgtctcttctttgctacaagatctccaaatttgagtttctgtca
acagaggcagaagaacttattgaatggggaggcagagggttgccactggctgctagatt
tgttcttctgcattttatcctaagtgacggggaaaaattaaatgactgtgtaggcaatg
gattttgctcaggctgattagaggcttacatagtgtgagtttaaagtttctgtgtacct
aatgtaaaatgctttcgatcacacctctgtagagatgtttgcctttccctgtcattcaa
gaatatgcgacctatagagttacgttgacttttcagtgtttactgaagtacagagatta
gcaatgccttggaatagtgttgctttttaaatccagaatattactcggtttatttaata
cttgagcccttgctaaagttgctcctcatattcttttgaaaatgtacaaaatgagaact
ctgattcctaccccctagagtatacagaatctccagggagtggtcccatcgaacccgtg
tgtgatatagaggactcttgataaaaactgtgtttatataaagatggtgtcaacctcta
cattccaagatttc

>N4 BC020379 SacI-XbaI [pAG82]
cagagacctcaggctctcagcacattccacaggctcctgagtccccgaggcctgggcca
gcttgggcaagccaagatcagatgtctctgtgttcgggaaggtctccgtgtgggaaagc
ccttgggggatcccgggtgaggagtgttgccccatccagagaatgaatgagttccttta
agtgccaccgccagcaagcccagaggcacacagtccgagtgcacccgcttagcctttac
attcctctccac

>N5 AP001157 SacI-XbaI [pAG84]
agtttcctacgtgtgcgacattcctaaggtcttcaagtttcctattctgttttgtgctg
tgtgaaagttctgatatgacccctgatgataaagtccataactgtagttaactggtcat
cactgctcaggtcccagttactaagaagactaccaaaatggcatcatcttaatcttgaa
gcgtatttctgacattcctgaaccg

>N6 AC116903 SacI-XbaI [pAG86]
ttttatattcatgcttgacattccaagaggcgttcttttcagcctgtggaggagacact
cttagatgtgggtttgtgtgtatgggtgcaacacttccggggataggaaagaagaacag
ttccctaacacaagctgttaagaagtctgtagcatctctgataacgaatagacccacaa
gctcctggaagctgtcgtgtgtccaatgcgtgaaaatgtctgtgacattttcagtgtta

gtcttgaactaaggcccttgtctgtgtgttttgaaggtgggtgaacctcagagaatgaa
cctgttatgatttggggtaagaatcacgttggaaaacctctcgcaaccagcacactagg
ccttctacggccatttctgtaaggggacagagcttaggtaaaggcacacactcatacac
acacaacaaagggccctgctaacttactcctcatgctgcttcagtgacctttgttcatg
tcacattcctggcgct

> PFTK1 NM_012395 SacI-XbaI [pAG72]
aaaaaatacatttaaaaaaacattccaagccaattggaagacatcattgggttcttac
tttaagacatctcctggaataactgttcaaatgcaggttttagaaacaatgcaggaatc
ttgctttaaagatgaaaaagggaatgggccagcttcccttactcaaggagttgagggac
cttggaggatgaaggcgagtatgtgacactggagaaaagtggaccaggcatgtctttg
ctttgatctggagggagggctgcctgatgcaggccggctcccagtggggcaggcctcgc
tgcagaatgcccagtagtactgcggccaaggggacagttaggagacttcatctaaagca
tgaaacctagctcctctacacacaaattcctatggaaatacctttgtgtacagtgtctt
acattttcctattagtcagaaagaaggagagaatgagtgagtgcttgaaatgtgtcata
ctgttttaggatcaagactaggaattaggagccaggttgacaaggacttttctgagag
ttgggtgagggtaaagcttttctataatcaagctcaatacaccaaggaaactggatcca
gaattcctaaactttaaaatggtactgtctgcggagtggagtatggatggttatgtcaa
agtcatagttcatcctatccagatgtagcattcatggtaaacttttaagtgctaagcaa
ggaattatttactgattggttttaaagagagcagaaaacacccaagtgtagaatgtcta
ctgtttgctacctagaaatcttttccattcctctttcatacattccaacccac

>GLCCI1 AK126731 SacI-XbaI [pAG76]
tattttatgcatttccctttcctcattacattccacattcttagaataagaagtgcatt
caatcctaggagaatgataatcctggacatgggtgaacatgaggagaaccagcaaaatc
tgtggtgtttgacatcactttgtcatgtggttacaagtaaaacaactgttgcattcact
gtttcaacatgtgtacatgtggctttttaaaagttcaggtgttgctcagtaaaggact
gtgacaatgttgcaaataaagtgttcagtactggactgtacataaacattccacattgt
g

>CORO1C NM_014325 SacI-SpeI [pAG193]
gagctggttattggtgtggtcctagggagggcggaaagggaggcactgccatttggaga
cattccatttcagatttgtcaaccagcgataggccacattccagtaagaactcaatttg
tctcccaaatttgcagaaacaaaacgtgatttaaaagctgagcttttatcagaaagct
tttttgatgttttaagtgttatgtgacttgttgaacttttaaaaagtgctacttttaa
aatcccagatactctgaattttagaaaacaaactaattctgattgtgtcgtgcccaagt
acccttttttttaatgaatagggaccaatgccacattgcttttatatttctttcttt
tttaatgttgccaaaaccaaaagtagctttg

>TNKS2 NM_025235 AgeI-SpeI [pAG195]
gcaaaaggataaaaatgtgaacgaagtttaacattctgacttgataaagctttaataat
gtacagtgttttctaaatatttcctgtttttcagcactttaacagatgccattccagg
ttaaactgggttgtctgtactaaattataaacagagttaacttgaacctttatatgtt
atgcattgattctaacaaactgtaatgccctcaacagaactaattttactaatacaata
ctgtgttctttaaaacacagcatttacactgaatacaatttcatttgtaaaactgtaaa
taagagcttttgtactagcccagtatttatttacattgctttgtaatataaatctgttt

tagaactgcagcggtttacaaaatttttcatatgtattgttcatctatacttcatctt
acatcgtcatgattgagtgatctttacatttgattccagaggctatgttcagttgttag
ttgggaaagattgagttatcagatttaatttgccgatgggagcctttatctgtcattag
aaatctttctcatttaagaacttatgaatatgctgaagatttaatttgtgatacctttg
tatgtatgagacacattccaaagagctctaactatgataggtcctgattactaaagaag
cttctttactggcctcaatttctagctttcatgttggaaaattttctgcagtccttctg
tgaaaattagagcaaagtgctcctgtttttagagaaactaaatcttgctgttgaacaa
ttattgtgttcttttcatggaacataagtaggatgttaacatttccagggtgggaaggg
taatcctaaat

> N7 NM_022748 SacI-SpeI [pAG142]
atgtgtgagcagaagggaggatgaggaaaaagagaagaaaccccggtactgacaagctg
tttttgagtgccactgtttgccatcatctaagccactgaatcaagtgtatttcaggctt
atttcaacattccaatgccctggttttcctgcttgaatctgttcgtggtcaaaggtttg
ggggaatttgtgaccctggaacatccccagagtgaaagatggagctgggccacatcaga
ataaggccttggccccatcctctcacagcctaggtgctctgcaggcatgctgactgtcc
tgattgcgatccagcccgaaattccctcctctgctttcaaaagtcaaatcccccattct
taggccacactggtgtcacaagctcctgtcagggagctggggtttgggaatgtgctttg
tgaactctgctttaaagtgaggggccgaggaaaacttagaaacaggcagagttggaagc
agccaaatcacagtgggtgttgtgtgtgtgcgtgtgtgcatgcgtgcgtgtatgcgt
gtgtgaaagcaggtggaccattccacttttagctcctattgatgcaccaaaccaagtg
cctcatttctgtgccaaatgtttgccttggtcgttgtggacctccttctctaacttgcg
gtggc

>N8 AK024929 SacI-SpeI [pAG143]
cgatgccctctttttctcaggatagcagataacctgctttgaaagagggcttaattctg
tgggtcctaaattttctcctttctctctctctttctgtgtgtgtgtgttgggaaaatgg
caagtttccaataccagctttggaggaacgattacgttttccctccaatttcaagtccg
aaagaccagagccctcattccaaagccccccacccagatggatttttcgtttcatttg
tcatccgtcccatgggagggccccatgtctcctcagaacccatcctggaggcagcaggt
cgggtagagtgagtttggcctgctcatgacctccacccctgagattgtgaacaaggatg
tctggggcgatgctgagaatgttttgaagctgctcccagatgacgctgatgatcacac
cagattgagtgctgcgatcgccttgagtccaacctctgcataaacgaggttctcataaa
caagttcactctaccctaagctaagtctatgtgagcaaacccacttcatcctttgtacc
tggagacctggttacactaacctgatactgacctgttcatgtagctggaatggtgtgtt
tcatgcagtgtggaccaagcaatggcatggggtgtgtgtgtgtgtgtgtgtgtctgt
gtgtgtgtgtttgtgtatgcgttcacacttgtgtgtgtatatgtgcatgtagatgctgc
ataaatgattttgatgtcaaagacaaacacattccattgttttaaatattctattatg
taaacaatacgcagagggaccatatctactcttgtcatattatttgtgatggtaaaaca
tgcatttgcaataaattaagctttctgggaaggcaagcagtattggagccaaacgactg
tctcggaaca

>N9 NM_005116 SacI-SpeI [pAG144]
agttttccagggagggaatgatataattttcatgacaatgtcatgtgatgatagaattt
ctccattgatgaatctctagtatgtgtgtatactttatttacccatacgtatattttat
gaccatgagctaagttgttatttctttcattcatatccctgctgaaatattggtactag
caacttaaagttggcaagtctcatcttcagtaatacggatgccatgggagtggccaggc

cagattgaaggtgaatatggagcagtagcagaagcctcattccagaacccatctggcca
gagaagcagcagcatcctgggggatggccgtgcatggggtgtacactcgctataggcat
aggcccggcatggctgtcgctggacgccagctgtgcacacccagccacacctgctgcac
gccgcgttagtgtgcggctccgggcctgagcattcgcaaagctcgcttctccagggagc
ctcctcttggctttggaaaagagcccagggacttaacgtgctgcctgttgtactctgtc
ccctcatgacttttagacacagggactaggacatcgaccaccctctgcccctcctgca
gtcagaacccgagtagtgcaagaacggccacccggttttgccagcgttcagcgtctgcc
catggacccataaggcacattagcgttgggggtctcttttcagcagcctcacagacatt
cccgtcctttctggttggtgttgactgtgtttcagcatcactctaccctcccatgtcct
tagcattgct

>N10 NM_181358 SacI-SpeI [pAG145]
tttccattttggattctcggctgagttctcacagaagcattttccccatgtggctctct
cactgtgcgttgctaccttgcttctgtgagaattcaggaagcaggtgagaggagtcaag
ccaatattaaatatgcattcttttaaagtatgtgcaatcacttttagaatgaatttttt
tttccttttcccatgtggcagtccttcctgcacatagttgacattcctagtaaaatatt
tgcttgttgaaaaaaacatgttaacagatgtgtttataccaaagagcctgttgtattgc
ttaccatgtccccatactatgaggagaagttttgtggtgccgctggtgacaaggaactc
acagaaaggtttcttagctggtgaagaatatagagaaggaaccaaagcctgttgagtca
ttgaggcttttgaggtttctttttaacagcttgtatagtcttggggcccttcaagctg
tgaaattgtccttgtactctcagctcctgcatggatctgggtcaagtagaaggtactgg
ggatggggacattcctgcccataaaggatttggggaaagaagattaatcctaaaataca
ggtgtgttccatctgaattgaaaatgatatatttgagatataattttaggactggttct
gtgtagatagagatggtgtcaaggaggtgcaggat

>N11 NM_002924 SacI-SpeI [pAG146]
gttgtgacctggagcagaggacattagaacaagatgttgcatgagcaaaggacctaaat
tgttatttttgtgtgtacattccatctccaatggactcttccgtctcaatgcctccatt
ccaaactgttgtctgctttctttctccttctactatgctggatctgtgtctcttcctt

>N12 BC072452 SacI-NheI [pAG147]
tcctattcgccaaatgaagggcagtgccccgcacgtaagttggaatgatggacctgtgt
tcagagacttaacagaaccaacaagcaaaacaagtgagaacaggaaaaaggaagaggac
actggaatcaattcttgagagttgcactacttggtttttcttccattccaagtttcgtg
ggacccagagccttttttctttaaaagctaaaaaacaagtgtttaattcctcttttg
ttatctgttagataattgagatcacctagaaatgcgtttaatctgttcactcactgtaa
attttgaggacccagaattgtcttgtttaatttatactttcacccctgttgcagttaac
accagagaaggaacgtgaatgtcgagcacagccactacccttgttggcacttaatttag
aaatagggtgagaagtttaaaagcccatcttgattttatttcattccttttggttctc
tgtgtaataatagcaggctacatagtgacattccagttccaagaaggtacatcctgtcc
attcattaattgctttgattactaggagggtttctgttcagttttgtttttaaatgtct
tgctgatctagttctttcagatggaataaccttccagtcccttagagagtggaactagt
ccatataacccagcttcagtagcaaaagtagaagccgccacatcttttcatttctccaa
gaggagagtggggaaggttc

>N13 NM_000337 SacI-SpeI [pAG148]
cattaaagttttccacttcaccctcccatagtctagagggatgcccattccatggtcct
ccagagaatagttttgacttaacatgtctgtttagcccacatcacgtcagttatcaaca
ccgccactgtgcttactgttcctacagccacaccaggcttgaagagttagtgagaccaa
caaataattggaagtattggaaaaagcaaaatacatggggacaaaaaaaatacagtgaa
attctttttatcaaactgatgctgtgagaaaccagatgaatgccagtttggctttattt
ctaagaatctgggtcttcattctctggtgtagaaggaatgcaaaaaactataacaacaa
caacaaaaacatattttgaaaagacatattctgacatctctgcttgtgtgtggtaaggc
aggttcctatcagacatttatccctttggtcaagatccctttttgctcatccagggtttc
atactcaatatcgcttaaaaaaaaaaaagtatcagctagggatgactctggaagtatga
gtatcatggtggggaggaaggaatttttttttaaatgtaaatgacccccatttaccagac
cctaatcaaagtcacttaagggaatccctcagccttttatttggaaacagttgaaataa
actggcagcagctagatcagagtatcttgctttattttataaaggccaaaggtagtatg
aagtttggaccaaaaaggtaaatagatccattccagcacctgatactgatttttcaagg
ctctatgaaaggtcaaaaatttcattaaacaagaccagttctccctcttccccctgtcc
caagaaatctt

>N14 AB018693 SacI-SpeI [pAG149]
ttgccttagcacttctttcctctccaattgtaaaaaaaaaaaaaaaaaaaaaaaaaaaa
aaaaaattgcacaatttgagcaattcatttcactttaaagtctttccgtctccctaaaa
taaaaaccagaatcataatttcaagagaagaaaaaattaagagatacattccctatca
aaacatatcaattcaacacattacttgcacaagcttgtatatacatattataaataaat
gccaacatacccttctttaaatcaaaagctgcttgactatcacatacaatttgcactgt
tacttttagtcttttactcctttgcattccatgattttacagagaatctgaagctatt
gatgtttccagaaaatataaatgcatgattttatacatagtcacaaaaatggtggtttg
tca

>N15 AY585209 SacI-SpeI [pAG150]
gtgtctgctgactccagtgcatccatgaactctggggttcttctggttcggccatcacg
gctctcctccagtgggactcccatgctagcaggggtctctgagtatgagcttcccgaag
accctcgctgggagctgcctcgggacaggtaataatactttgactcccaacttttgaa
ggcattccttgcttgctattctagtggtggaaggtgtgtatactaacaccttctgggcg
gtgaaggttcatgaaatgcctgcagaaacattcattccactttctgctcgcagttgggt
gaaatacagagaggtggagatggggtgcctctcaaggtttgaacttcaccagccccaac
ttatgccactctctgtttcccccgaaagactggtcttaggcaaacccctgggagagggc
tgctttgggcaggtggtgttggcagaggctatcgggctggacaaggacaaacccaaccg
tgtgaccaaagtggctgtgaagatgttgaagtgtaagtgatgcttctgtccttgcaaag
aaaatcttgtccattccaagcaaacagcaggcctcgggcccagcagaacagcttctct
ggtgcctttgctgcccacacctgctccgctgcctgcttggggttacctgggagtggagg
gggtgttctgcaccttgagattgttctatgacgtcactctgttccttggcgt

>N16 NM_182691 SacI-SpeI [pAG151]
tcttctggcaaagggacatcattcctgtattatagtgtatgtaaatgcaccctgtaaat
gttactttccattaaatatgggaggggactcaaatttcagaaaagctaccaagtcttg
agtgctttgtagcctatgttgcatgtagcggactttaactgctccaaggagttgtgcaa
actttcattccataacagtcttttcacattggattttaaacaaagtggctctgggtta
taagatgtcattctctatatggcactttaaaggaagaaaagatatgtttctcattctaa

aatatgcattataatttagcagtcccatttgtgattttgcatatttttaaaagtacttt
taaagaagagcaatttccctttaaaaatgtgatggctcagtaccatgtcatgttgcctc
ctctgggcgctgtaagttaagctctacatagattaaattggagaaacgtgttaattgtg
tggaatgaaaaaatacatatattttggaaaagcatgatcatgcttgtctagaacacaa
ggtatggtatatacaatttgcagtgcagtgggcagaatacttctcacagctcaaagata
acagtgatcacattcattccataggtagctttacgtgtggctacaacaaattttactag
ctttttcattgtctttccatgaaacgaagttgagaaaatgattttccctttgcaggttg
cacacagttttgtttatgcatttccttaaaattaattgtagactccaggatacaaacca
tagtaggc

>PRKD1 NM_002742 SacI-SpeI [pAG198]
tgccaagacaaatcaacagaagcatttgtattttgtgtgaccaactgtgttgtattaac
aaaagttccctgaaacacgaaacttgttattgtgaatgattcatgttatatttaatgca
ttaaacctgtctccactgtgcctttgcaaatcagtgtttttcttactggagcttcattt
tggtaagagacagaatgtatctgtgaagtagttctgtttggtgtgtcccattggtgttg
tcattgtaaacaaactcttgaagagtcgattatttccagtgttctatgaacaactccaa
aacccatgtgggaaaaaaatgaatgaggagggtagggaataaaatcctaagacacaaat
gcatgaacaagtttaatgtatagttttgaatcctttgcctgcctggtgtgcctcagta
tatttaaactcaagacaatgcacctagctgtgcaagacctagtgctcttaagcctaaat
gccttagaaatgtaaactgccatatataacagatacatttccctctttcttataatact
ctgttgtactatggaaaatcagct

>QKI NM_206855 SacI-SpeI [pAG199]
aattcagggaagatttggtcacattgaagatacagtattttgtagtatttataaactg
ttctaaatgatagactatagaaaacatttttgtcatatgaaggtaaatcagtccattat
tttggatcatttaaactgaacattacaccttctgggctttgatttatgaagtggcacag
aatctaactttgcactgaatatcttttcattcacatctcctctgccttagtcacaatgg
caacagtacagaaaattctatcaaacctcagaatatagtagaaataattaagctgttga
atgagtcttaaaaattatactactgttaagtggaccaagtttggtgaagcagaatgtga
caaaggttgattaaggaaggaacaactcaaggacattgggaatgataacttttccactt
gagaactactttatgttttactgtaatttttaaagttttttgtccttttttgttatttt
gcaaaagaaaatagtatttacaggtggcttcttttaaaatataaaaatataaagcagga
atgtatatgaaatgtcagattttattgtatttgcagagtattagctttgaaattgaata
aagaaagctgtttgtagttttaaaatgccttattggtatcaattagaaatttcttctat
tttttgatgtacttagagcttttgagtgtagaattttaaatggcaggattttacagtg
tttacatgcaagtgcattttataagtgttctatatgtgtaaaatagtattttcaactgg
aaagtgttggctagtgcaaaa

>HIPK1 NM_181358 SacI-SpeI [pAG201]
tgtgatcctccagtgttatcccggagatggattgatgtctccattgtatttaaaccaaa
atgaactgatacttgttggaatgtatgtgaactaattgcaattatattagagcatatta
ctgtagtgctgaatgagcaggggcattgcctgcaaggagaggagacccttggaattgtt
ttgcacaggtgtgtctggtgaggagttttcagtgtgtgtctcttccttccctttcttc
ctccttcccttattgtagtgccttatatgataatgtagtggttaatagagtttacagtg
agcttgccttaggatggaccagcaagcccccgtggaccctaagttgttcaccgggattt
atcagaacaggattagtagctgtattgtgtaatgcattgttctcagtttccctgccaac

attgaaaaataaaaacagcagcttttctcctttaccaccacctctacccctttccattt
tggatt

>SFRS12 NM_139168 SacI-SpeI [pAG202]
tgggactacccacaaagtgagcatttctttttaaattttcttgacatttccaagcttat
tatgaataatattgcagtgtgtcttgtcagctgtaggtggcaaaggtgcccttataaaa
aaggaaactggcttttcaaaatgggctatgggagcacaagctgaagctttagtgccttc
tacaatgtggtatactgttttctagaattttatatgtgctagtcattctcaattcatat
ggaatctagatggatatttcatgcatacccatagagaagtgtgtaagtgatatgtcaga
agagcttcttactgatttcacctaaaatgagaaggaagtcctgttttcaagaatgacat
tagagtcatgcagctttgggaccatcagttttatactgtgataattgaaaatgaaacat
gttcttattttccttaaattgaagaaaacccttttagttgtctacattggatggccttat
tacctctcaatcatcttttcataaatgatgtgcagaaattgtacttaaggacttaggag
tatatgggaggttattggttttatgtttaaggatacgtttacttgagtttaagatacag
gtcatccatcattcttaggctcacttttttacagaaagtatgcaaatagtaaagtgacag
cactgctaatgtttttccccagtactataacttgtggtttctgaactcattattgttgt
atttccaaaaaagtaatacctttttaattagtgtattaaaagttaagtataattatttta
atgcaatctaatacaatcagattactcagttgccttacctcatgggaagagttactttt
ttagatctaaaaagctgaatagcatgttagttacttggtttcaacttgagttttctttt
aatgttaataagattgaaactttagtatttagtggggaatggaaagagttgcccttgtt
gcaagtaatgaagc

>N17 NM_014397 SacI-SpeI [pAG184]
gcgtggatgcaccgtgccttatcaaagccagcaccactttgccttacttgagtcgtctt
ctcttcgagtggccacctggtagcctagaacagctaagaccacagggttcagcaggttc
cccaaaaggctgcccagccttacagcagatgctgaaggcagagcagctgagggaggggc
gctggccacatgtcactgatggtcagattccaaagtcctttctttatactgttgtggac
aatctcagctgggtcaataagggcaggtggttcagcgagccacggcagccccctgtatc
tggattgtaatgtgaatctttagggtaattcctccagtgacctgtcaaggcttatgcta
acaggagacttgcaggag

>N18 NM_006457 SacI-SpeI [pAG185]
aaaaaccaatttcctgatggactattaaattcatcttagaataaattagtgaagaattt
aattttagaataaataatccaatctgaaataattataccttctttccttgttaggtagt
tatgagtaaatctgcaaaaggcaatgaaaatgccttaaattttatcaataacagaatta
ttgtatttaaaaaaaaactaatacttatctttaaaatagtaaataggattttaaacaga
gaattttatcagtaataggtgtcagtttttaaaaaattgcttgtaggctgagcgcggtg
gctcacgcctgtaatcccagcactttgggaggccaaggtgggtggaccacatgaggtca
ggagtttgagatcagcctggccaacatggtgaaaccccatctctactaaaaatacaaaa
attagccggacgcagtggcacgcgcctgtaatcccagctactcaagaggctgaggcacg
agaatcacttgaacccgggagggagaggttgcagtgagccaagatcgtaccactgcact
ccagcctgggtgacagagtgagactctgtctccaaaaaaaaactttgcttgtatattat
ttttgccttacagtggatcattctagtaggaaaggacaataagatttttatcaaaatg
tgtcatgccagtaaga

>N19 NM_002508 AgeI-SpeI [pAG186]
cgttgccctgacaacaccttgggagttgactgtatcgaacggaaatgaagacaagagtg
ccttatttcctttccaagtatttcacagcaacactctacttgaagcaacttggtccaga
ttgaaaagtgtcctctggctgagtggccactaggcccagacccagcccagcctgagccc
caacaacaacttttccctcactgttccccaaaacatgcaccctggacttctctaataga
aaagtctccacccctacacaaggacagaaccctccacccctacccccaaccctcagaca
gacttatacacccctgagtgaggattacatgcccatcccagtgtcctaggacctttcc
caatactagccccccagtggtgaacagaacctcccaaatttgagttgcacccttccctg
tggccttatgagctcagcctcgctttgaggtacccaccgtcctgtcagctccttgacct
atgagctggggcctgactaggaaaagttgggagttaaggaggaaattagcattccttaa
tgttttgttttggtgctctgaatttcttctttattatagtcctatagttttactcctca
gttcctcaccatcatcatcttgtctaagacccccattataatattcatgcgctgctttt
tcatcaaaacctaccctgtcctagagatctatgggcatttggtggatgataatgagcag
cccctcccagatagaatgtcaatatttgagcagtaggatattggcatttgttagttaaa
ggcttaaatcaaaagaatgtccaatggtaggaatttcaaggtgtaggtcagatatttga
gaataggggatttttttgatgtgccttaaattataccaaagattactaattattcctct
ttgcccaaaatacttgcatccaaggttctagtctctgttgctgtgctggtctttagccc
cactgctggcactgatgtccctccttttcacg

>N20 NM_020639 SacI-SpeI [pAG187]
tttatgcaacaaggaacgaatggtagcagccagctttgcggggcgtatgtgtggccagc
tcttaaccattccagtctattacttgggtgagtccttgtggacaaccacacacacgtgc
ccacatggtactagctgccgttcgtttctcgttgcctaagatgttttggcaactctaga
gccacaggcctaagtcattaaaaaattctccctttgtaacctcagtgctggggactgag
gcgagccccctcaggtcgctggagtgcaccagtcttggggaagaggtgcaggagaagct
gtgtttttatctccacacgcagtatgaagataaaattacatagtattacctagacata
gacagtattacctaggtagatgcactgctcacctgcgcccttcccagctctcatttttg
ttaggtgatttgggatagggatagtgttttggggtatgggggggagtgtttctgacctgc
tttgcagacgtgcctccgcacctcagcagtttggggtgtggccccagggcggttcttgg
atgtaaaagatgtggccatctagcctcgtaacttcactgtcacctgtgtcccatagggt
gccttctgaatactgttattagaataagtttgttgcagaacgtgaccctgcgtgcaaac
atgtaccgtggcctggtatatgatagagattgatattaatgtaccatgtatgttaatgt
gaatctgtgggcaggatactttccatggcaggaaatatccaagctgttgaaactggct
atgttttaatatgcctcattgtgcctttactgttgtgtggactgcgtgagggacaagaa

>N21 NM_178816 SacI-SpeI [pAG189]
tttcttctagcatcaaaggttgcttaaaataacagtggaacatttaaaagcaagagaat
tctaatttaccatcattcaaattatgttaaatccaactcaactttataagctaaggaac
atgtgaaggataaataaaaatttgatagttaagaagataaatatatgcagagaaattat
gtgatgagtcaaattgattattacaggacaatatactgtattaataaaacagactaaat
ttaaatgtattgtgcaatggcaaacctagtgccttcttggaaaaacatgccttaacaag
ctcatgtggttgcaaggttgataatatcttgataataacttatgttgttcttattccta
gtaattgctggcctccttatgaccttccgcctagtttcaggcagtagctaggagagcta
gactaactgctccagagaagtgaggccaaggcatgcaacagagtcacttctccgctttt
ccatttcaaagtag

>N22 NM_199072 SacI-SpeI [pAG191]
tgtgtggactctcttcctcagactgtggggacagatacaattccactcctgtccacagg
aacatgagatttagcagactaaggagatctgtaaagaatgaaccataccacaaggcata
ctgaagtgaggattataagagaaataaactcaaaatgctgttggaatatgcagagaatt
gctaccagaatattcagtaaggtttcagggagaatgtggcatttgaggactctcttaga
atgagtgattcacctgctatttaaatgaattatttagattttgacaaagatttaggtg
gacaccctaaactgtgtgtgcctttaaccagttaaaagaacagtgccttcagcatactt
ttttattagttgtaggaatacagcttttgaaaa

>N23 NM_002649 SacI-SpeI [pAG192]
atgatgcttccaaacatctccttagtgtctgcaggtgttagtggtgtgctaaaagcaag
gaaagcgagttagtcttttcagtgtcttttgcaattcaattcttttgtcatgtataact
gagacacacaaacacagcaggagaaatctaaaccgttgtgccttgaccttcctctgctg
gtcttgttccagggttatgaatatgaaaaaatagagatgagacttttgtgtcaactct
gtccacaagagtgagttatctagtatgattagtatagctttctccagcatggcagcagg
aagtaactacagggcctcttttatgcctgacatttcttcccttccttttttccctgcctc
ccttttttcatcaattgcaatgctcccacaactctttacagacttgtgaaatcttcaaga
acacctttactctataactcaaaaattagttgaaaaataattacttctcaaggattatt
agaatcttaggtacttatttgtaaagatgtttagtgacttttttttcaagtatcttatt
aaaggaggcattctagaaaatatgaattagtttccaaatgccttaattttaaactttgg
cctgaacagttttt

>pIS2 cloning site, pRL-SV40 sequence is in italics, restriction sites are underlined
*gaacaataattctag*gagctctataccggtctcgatatcactactagtgt*tctagagcg*
gccgct

Supporting References

1. A. I. Su *et al.*, *Proc. Natl. Acad. Sci. USA* **101**, 6062 (2004).
2. D. Karolchik *et al.*, *Nucleic Acids Res* **31**, 51 (2003).
3. M. Blanchette *et al.*, *Genome Res* **14**, 708 (2004).
4. S. Griffiths-Jones, *Nucleic Acids Res* **32**, D109 (2004).
5. B. P. Lewis, C. B. Burge, D. P. Bartel, *Cell* **120**, 15 (2005).
6. V. K. Mootha *et al.*, *Nat Genet* **34**, 267 (2003).
7. B. L. van der Waerden, *Mathmatical Statistics* (Springer-Verlag, Berlin, 1969), page 74.
8. P. K. Rao, M. Farkhondeh, S. Baskerville, H. F. Lodish, (data not shown).
9. B. P. Lewis, I. H. Shih, M. W. Jones-Rhoades, D. P. Bartel, C. B. Burge, *Cell* **115**, 787 (2003).
10. K. D. Pruitt, T. Tatusova, D. R. Maglott, *Nucleic Acids Res* **33**, D501 (2005).
11. Y. Zhao, E. Samal, D. Srivastava, *Nature* **436**, 214 (2005).
12. I. Hofacker, Fontana, W, Stadler, PF, Bonhoeffer, S, Tacker, M, Shuster P, *Monatshefte fur Chemie*, 167 (1994).
13. H. Robins, Y. Li, R. W. Padgett, *Proc. Natl. Acad. Sci. USA* **102**, 4006 (2005).
14. J. G. Doench, P. A. Sharp, *Genes Dev.* **18**, 504 (2004).
15. N. C. Lau, L. P. Lim, E. G. Weinstein, D. P. Bartel, *Science* **294**, 858 (2001).
16. B. John *et al.*, *PLoS Biol.* **2**, e363 (2004).
17. A. Krek *et al.*, *Nat. Genet.* **37**, 495 (2005).
18. J. Brennecke, A. Stark, R. B. Russell, S. M. Cohen, *PLoS Biol.* **3**, e85 (2005).
19. E. Hornstein *et al.*, *Nature*, in press (2005).