SUPPLEMENTARY INFORMATION

Characteristics of *let-7* recognition sequences that correlate with transcript degradation

Using FIRE (Finding Informative Regulatory Elements) (1), we searched the promoter regions and 3'UTRs for motifs that are highly informative about the gene clustering partition. The most informative motif discovered by FIRE, HNUACCUC (with H= [A/C/U] and N= [A/C/G/U]), is perfectly complementary to the *let-7* seed. Thus, among all possible sequence elements in either the promoter or 3'UTR, the single most informative sequence element for predicting the response of a gene to *let-7* over-expression was the presence of a let-7 seed in the gene's 3'UTR. To identify potential *let-7* targets, we looked for genes consistently down-regulated in the presence of *let-7* (i.e., genes within cluster 2 or 4) that also contain at least one *let-7* recognition site within their 3' UTR, resulting in 838 genes (Supplementary Table 1).

We compared this list of experimentally-predicted *let-7b* targets with those predicted computationally. To generate this second, computationally predicted list, we downloaded all *let-7b* targets predicted by TargetScan (2) and found 242 target genes present in the clustering partition in Supplementary Figure 1. or The overlap between the two sets contains 125 genes, many more than would be expected by chance ($p < 10^{-78}$) (Supplementary Table 2). Genes predicted computationally but not observed experimentally could be regulated in other cellular contexts or could be modulated at low levels (not detectable with this method), while genes down-regulated by *let-7* but not predicted by computational approaches might be missed by the target prediction algorithm because they contain non-conserved binding sites, or they may be indirectly regulated by *let-7* overexpression.

We anticipated that evolutionarily conserved recognition sites would be more likely to result in transcript degradation than sites that are not conserved. To test this prediction, we analyzed the extent of evolutionary conservation for recognition sites with a whole-genome multiple alignment of 17 vertebrate species available through the UCSC Genome Browser. For each *let-7* seed match in a human 3' UTR, we determined the number of species in which the seed match (TACCTC) is exactly conserved. The fraction of species in which it was conserved divided by the fraction of species for which the sequence aligned was defined as the conservation score, which could range between 0 and 1. *let-7* recognition sites within regulated genes (clusters 2 and 4) were more likely to be conserved than recognition sites within non-targets (clusters 1, 3 and 5) (Figure S1A) (Wilcoxon test, $p = 4.33 \times 10^{-15}$). We also performed the same comparison between clusters 2 and 4 versus clusters 1, 3 and 5 but monitored the extent of conservation of the *miR-1* seed match (CATTCC) as a control (Figure S1B). In this case, there was no difference between *let-7* targets and non-targets ($p = 0.99$). Thus, our findings demonstrate that *let-7* seed matches are more conserved in target genes, and this does not reflect a generally higher level of conservation in the 3' UTR of the target genes, but rather specificity for the *let-7* seed.

The presence of multiple miRNA recognition sites has also been correlated with increased functionality (3). We found no significant correlation between the absolute number of *let-7* recognition sites within a gene's 3'UTR and whether or not the gene was a functional *let-7* target defined by its presence in cluster 2 or 4 vs. cluster 1, 3 or 5. However, we did observe a significant correlation between the density of *let-7* seed matches in the 3' UTR, where the density is defined as the number of *let-7* seed matches divided by the 3' UTR length in bp (maximum Kendall tau rank correlation coefficient = -0.119, corresponding $p = 2.9 \times 10^{-7}$) (Supplementary Table 2).

We also considered the contribution of non-seed binding between the miRNA and its target mRNAs (3). We reasoned that if the portion of the miRNA downstream of the seed contributes to target recognition, we would expect more extensive pairing between the *let-7b* miRNA and its targets than for non-targets. To test this prediction, we determined the minimum free energy (MFE) of pairing between miRNAs and 3'UTRs, anchored at the *let-7* seed. Since all co-structures are required to have an exact *let-7* seed-pairing, lower co-structure free energies indicate stronger pairing across the non-seed part of the miRNA. We first asked whether there is a significant difference in the distribution of MFE between genes within cluster 2 (weaker targets) and the genes within cluster 4 (stronger targets). We did not detect any significant differences in *let-7b* pairing strength between cluster 2 and cluster 4 (p = 0.50).

We then compared the MFE between co-structures from clusters 2 and 4 (*let-7* predicted targets), and co-structures in clusters 1, 3, and 5 (non-targets). We found that the MFEs were significantly lower in clusters 2 and 4 compared to clusters 1, 3 and 5 (Wilcoxon test, $p \sim 4.1 \times 10^{-5}$; Figure S1C), consistent with a model in which non-seed binding contributes to miRNA function. In fact, all 7 members of the *let-7* family of miRNAs have significantly increased pairing strength with members of clusters 2 and 4, as compared to clusters 1, 3 and 5 (all p-values $< 10^{-3}$), suggesting that *let-7* family members may target similar sets of genes. As a control, we shuffled the non-seed portion of the *let-7b* sequence and repeated the analysis (Figure S1D). When the nucleotides in the non-seed portion were randomized, we found no significant difference between the distribution of MFE's from clusters 2 and 4, compared with the distribution of MFE's from clusters 1, 3, and 5 (p ~ 0.23; Figure S2). We repeated this analysis without shuffling but using the *miR-1* seed instead of the *let-7b* seed, and, as expected, found no difference between clusters 2 and 4, and clusters 1, 3, and 5 (p ~ 0.95; Figures S1E and S1F).

The 3'UTR 7-mers or 8-mers matching positions 1 through 8 (extended seeds) of miRNAs are highly conserved (4) and are predictive of miRNA targeting (3). To test whether the increased pairing strength in clusters 2 and 4 is only due to pairing at position 1 and/or position 8, we removed the two nucleotides around the seed and the seed match, and recomputed MFEs. Again, we found that MFEs were significantly lower in clusters 2 and 4 compared to clusters 1, 3 and 5 (Wilcoxon test, p = 0.0023). We conclude from these analyses that the 3' extremity of the *let-7* miRNA (i.e. the non-seed part past position 8) contributes significantly to target specificity, as has been reported for other miRNAs (3).

Supplementary Experimental procedures

Human foreskin fibroblast isolation—Foreskin fibroblasts (HCFS06) were isolated from newborn human foreskin obtained from the National Disease Research Interchange under a protocol approved by the Princeton Institutional Review Panel. Foreskins were washed twice in antibiotic/antimycotic solution (100 U/ml penicillin G, 100 μg/ml streptomycin, 2.5 μg/ml amphotericin B, Invitrogen Corp. CA), incubated in 100% ethanol for 1 min on ice, and washed in antibiotic/antimycotic solution for an additional 10 min. The subcutaneous connective tissue was recovered by scraping the dermal side of the tissue, and the tissue was transferred to 0.5% dispase solution (BD Bioscience), quickly minced into small ~0.5 cm pieces, and incubated at 37°C for 3-4 hours. The epidermal sheet was removed. The dermal sample was minced into 2-3 mm pieces, placed in a 15 ml polypropylene tube with 3 ml of 1000 U/ml collagenase (Fisher Bioreagents) and incubated at 37°C for 1-2 hours with vigorous stirring. After the addition of 3 ml ice cold DMEM with 7.5% FBS, the supernatant was filtered through a 70-mm nylon mesh strainer (BD Bioscience). Cells were centrifuged for 10 min at 150 x *g* at 4°C, resuspended in complete growth medium (DMEM, 8% FBS, 1X penicillin and streptomycin, 1mM sodium pyruvate, 10 mM HEPES buffer, 2 mM L-glutamine, 0.1 mM nonessential amino acids) and cultured. Characteristic fibroblast morphology was determined visually by light microscopy.

Microarray analysis— To monitor gene expression changes associated with *let-7* over-expression, dermal fibroblasts were transfected with pre-*let-7b* or negative control RNA. Twenty-four hours after transfection, cells were serum-starved (DMEM with 0.1% serum) for 36 hours and re-stimulated by adding medium with serum (DMEM with 10% serum). Samples were collected at 24 hours after transfection and at 0, 12, 24, and 36 hours after serum starvation. Total RNA was isolated using the mirVana miRNA Isolation kit. High quality RNA was confirmed using a Bioanalyzer 2100 (Agilent Technology) and the amount was determined with a NanoDrop spectrophotometer. Three hundred and twenty five nanograms of total RNA was amplified using the Low RNA Input Fluorescent Labeling Kit (Agilent Technologies) to incorporate Cyanine 3-CTP (Cy-3) or Cyanine 5-CTP (Cy-5) (Perkin Elmer). Cy-3-labeled negative cRNA was used as reference RNA to compare with corresponding Cy-5-labeled *let-7* over-expressing samples at each time point. The labeled cRNA was mixed and co-hybridized to whole Human Genome Oligo Microarray slides (Agilent Technologies) at 60°C for 17 hrs and subsequently washed with the Agilent Oligo Microarray Hybridization Kit. Slides were scanned with a dual laser scanner (Agilent Technologies). The Agilent feature extraction software, in conjunction with the Princeton University Microarray database, was used to compute the log ratio of the difference between the two samples for each gene after background subtraction and dye normalization. Of the ~44,000 probes on the microarray, 25,972 probes generated signal in four out of five arrays. Data for these probes was mapped to genes based on UniGene Clusters. If multiple probes mapped to a single gene, the values were averaged, resulting in 14,590 genes. From this set, a RefSeq identifier was located for 11,694 genes, and these were clustered with the *k*-means algorithm. The microarray intensity data will be made available through the PUMA Database (http://puma.princeton.edu) for archiving and analysis.

Motif discovery— We used FIRE (with default parameters) to search the promoter and 3'UTRs of human genes for motifs that were informative about the gene clustering partition obtained from *k*-means (Figure 2A). FIRE uses a two-step search algorithm to look for informative motifs (1). The first step consists of an exhaustive, *k*-mer-based exploration of motif space. For each possible *k*-mer, the mutual information between the *k*-mer's pattern of presence and absence across the 11,694 genes (in promoters or 3'UTRs) and the cluster indices of the clustering partition is calculated. The most highly informative *k*-mers are then selected, and optimized into more degenerate and more informative motif representations (using the degenerate code). The statistical significance and robustness of the optimized motifs was then determined using nonparametric tests based on randomizing the gene labels of the clustering partition.

miRNA:target folding— We extracted 18 bp 5' and 1 bp 3' of all seed matches in 3'UTRs to produce a 25 bp RNA fragment. We determined the MFE hybrid co-structure with the *let-7b* miRNA sequence and the corresponding free energy using RNAcofold (5).

RNA interference— For Wee1 small interfering RNA (siRNA): ribonucleotides (sense: 5' GGAGAUCAAUGGCAUGAAATT 3' and anti-sense 5'UUUCAUGCCAUUGAUCUCCTT 3') and (sense: 5' UUGUAAUGGUGGAAGUUUATT 3' and anti-sense 5'UAAACUUCCACCAUUACAATT 3') were synthesized and annealed by Sigma. siRNAs targeting the unrelated protein luciferase (sense: 5' CGUACGCGGAAUACUUCGATT 3' and anti-sense 5' UCGAAGUAUUCCGCGUACGTT 3') were used as a control. Cells were reverse transfected with 100 nM of each of the above siRNAs mixed with 100 nM of *pre-let-7b* or negative control pre-miR using Oligofectamine. After 4 hours, growth media plus serum for a final concentration of 10% was added, and cells were incubated in this medium for another 24 hours before being used for experiments.

Cell lysate preparation— Cells were lysed in a RIPA lysis buffer (0.1% Nonidet P-40, 50 mM HEPES pH 7.5, 250 mM NaCl, 20 mM EDTA) containing a protein inhibitor cocktail (Roche). Cell lysates were incubated at 4°C for 15 min, sonicated and clarified by centrifugation (14,000 x *g*) for 10 min at 4°C. The supernatants were analyzed for protein concentration with the DC protein assay kit (Bio-Rad Laboratories) according to the manufacturer's instructions.

Supplementary Table 1: Putative direct *let-7* targets based on gene expression and the presence of at least one *let-7* binding site in the 3' UTR. (*) Represents the 125 genes also predicted by TargetScan.

Supplementary Table 2: Correlation coefficients and p-values for the correlation between *let-7* binding site density and transcriptional response.

Supplementary Figure S1:
(A) Functional *let-7* recognition sequences are more likely to be evolutionarily conserved. Whole-genome multiple alignments of 16 vertebrates were downloaded from the UCSC Genome Browser. The 3' UTRs were extracted from RefSeq. For each *let-7* seed match in a human 3' UTR, we determined the fraction of aligned species in which it was conserved (the conservation score). The distribution of conservation scores was significantly different between *let-7* recognition sequences in genes within clusters 2 and 4 versus genes within clusters 1, 3 and 5.
(B) The *miR-1* recognition sequence is not more conserved in the 3' UTR of *let-7* target genes compared with non-targets. The same analysis was performed as in (B), except conservation of the *miR-1* seed match (CATTCC) was determined in targets and non-targets. No difference was observed in the distribution of conservation scores.
(C) Pairing of *let-7* with target sequences outside the seed region contributes to target specificity. For each *let-7* seed present in the 3' UTR of genes within the clustering partition in Figure 4, the *let-7* miRNA and the extended *let-7b* seed were co-folded. The distribution of minimum free energies (MFEs) for the folding is plotted for the genes within clusters 2 and 4 (target seeds) in grey, and for the genes within clusters 1, 3 and 5 (non-target seeds) in black. Minimum free

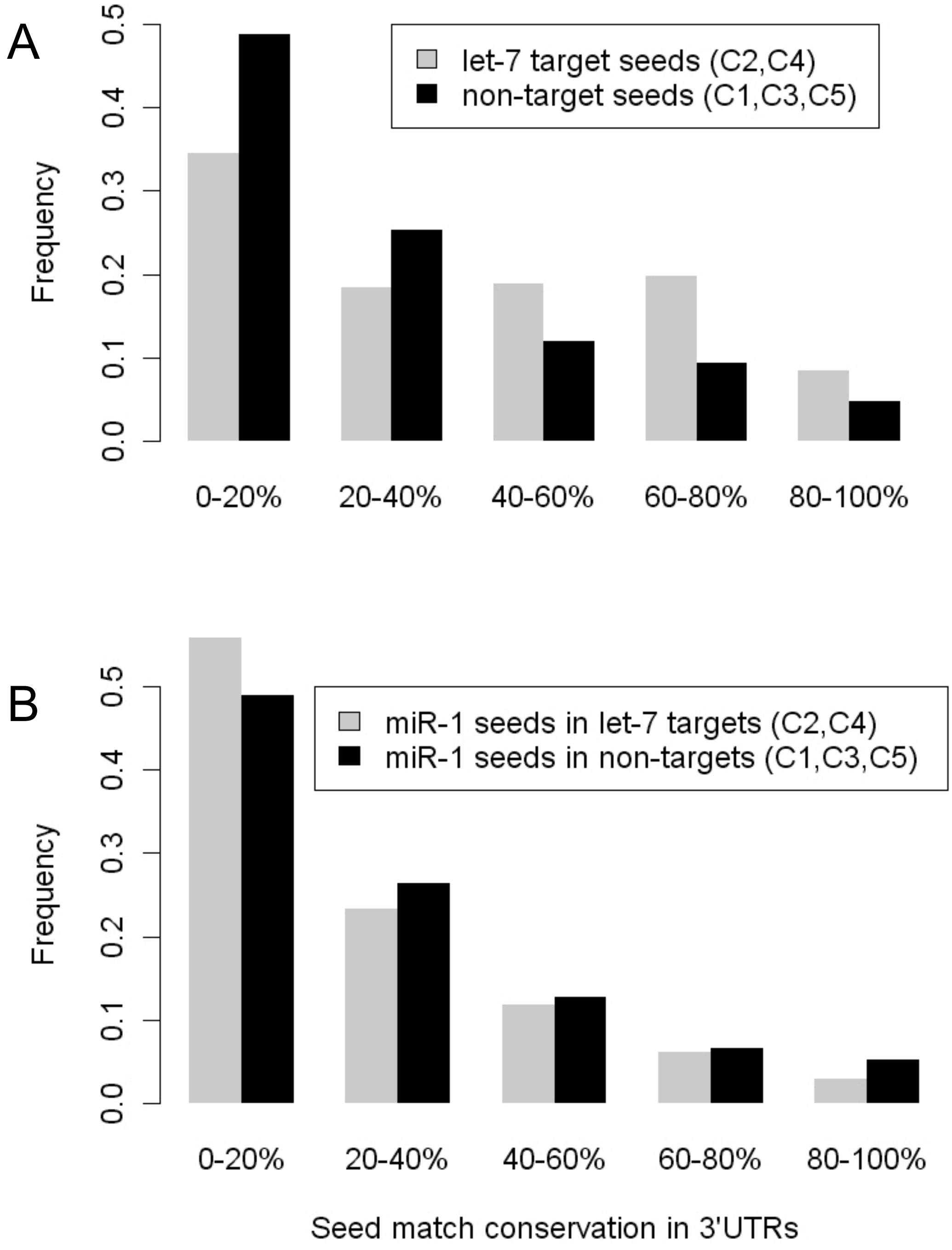
energy of *let-7* binding is lower for target genes than for non-targets.
(D) Scrambling of the sequence 3' to the seed on the miRNA eliminates the difference in MFE. The distribution of MFEs for co-folding *let-7* with *let-7* target sequences in clusters 2 and 4 was compared with the MFE distribution for co-folding of non-target seeds after scrambling non-seed nucleotides. In this case, the distributions of MFEs for target seeds (grey) versus non-target seeds (black) were not significantly different.
(E) Presence of a seed for the *miR-1* miRNA does not correlate with gene expression changes induced by *let-7* overexpression. *miR-1* target sequences in the 3' UTRs of genes within the clustering partition were co-folded with the *miR-1* miRNA. The distribution of MFEs for folding genes within clusters 2 and 4 (*let-7* targets, plotted in grey) is not different from the distribution for genes within clusters 1, 3 and 5 (*let-7* nontargets, plotted in black).
(F) Extended base pairing beyond the 8-mer extended seed contributes to miRNA-target pairing. The nucleotides in the 1 and 8 positions around the seed and seed match were removed and the distribution of MFE was computed for clusters 2 and 4 (grey) versus clusters 1, 3 and 5 (black). A significant difference was observed, consistent with the importance of base-pairing outside the 8-mer miRNA seed.

Supplementary Figure S2:
A549 (**A** and **B**) and HepG2 (**C** and **D**) cells transfected with *pre-let-7b* (+) or negative control (NC) (-) were collected at the indicated times. Cell cycle profiles were monitored using flow cytometry (**A** and **C**). This experiment was performed in triplicate; the data shown reflects the averaged results. Excess *let-7b* correlated with a larger fraction of cells in G0/G1 phase. Lysates were collected at the indicted time points after transfection and analyzed by immunoblotting with antibodies to Cdc34 and Wee1 (**B** and **D**). GAPDH was used as a loading control. In both A549 and HepG2 cells, *pre-let-7b* transfected cells contained lower levels of Cdc34 compared to cells transfected with control pre-miR. In A549, but not HepG2 cells, decreased Cdc34 resulted in elevated levels of Wee1.

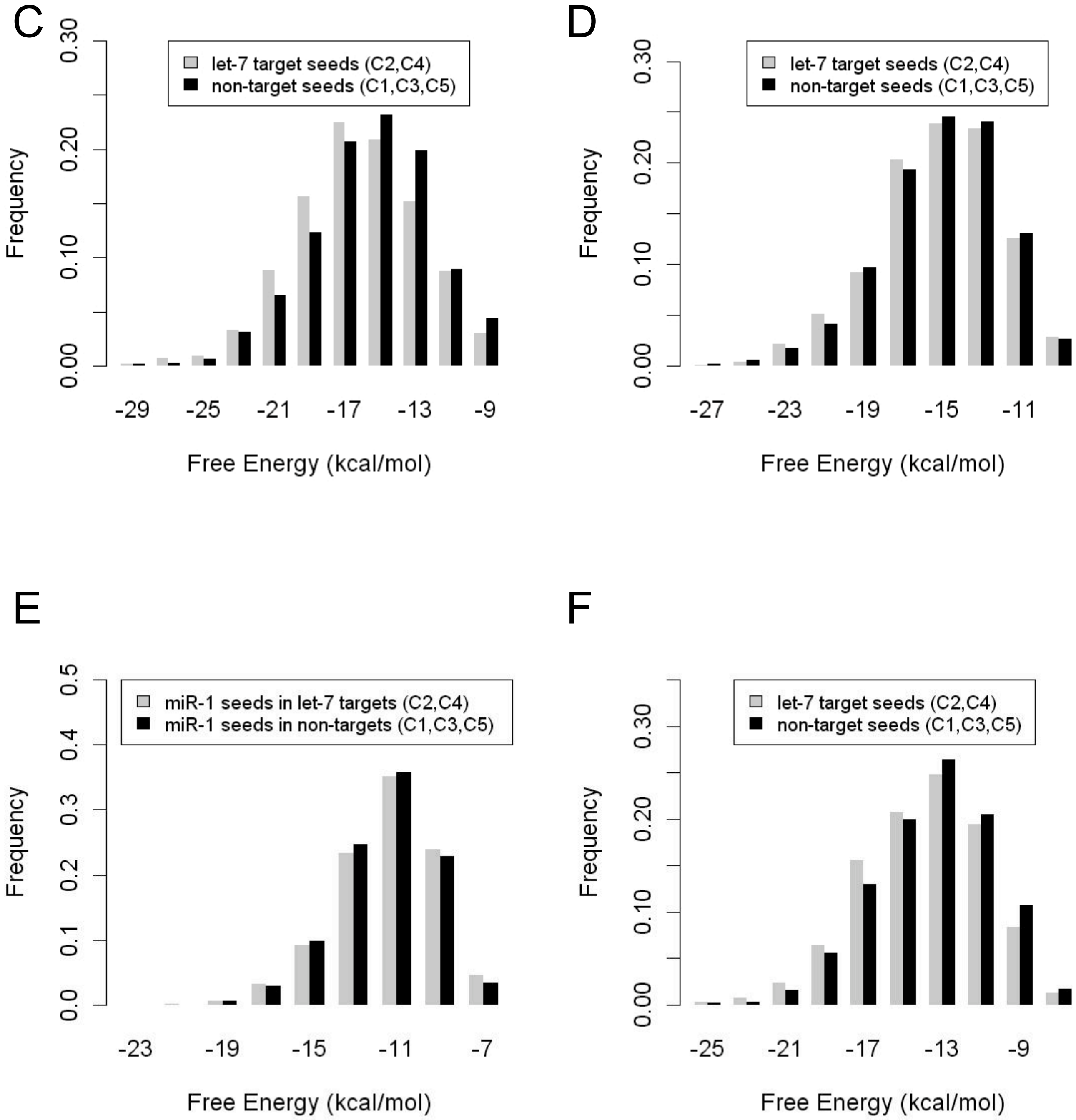
Reference

1. Elemento, O., Slonim, N., and Tavazoie, S. (2007) *Mol Cell* 28(2), 337-350
2. Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P., and Burge, C. B. (2003) *Cell* 115(7), 787-798
3. Grimson, A., Farh, K. K., Johnston, W. K., Garrett-Engele, P., Lim, L. P., and Bartel, D. P. (2007) *Mol Cell* 27(1), 91-105
4. Chan, C. S., Elemento, O., and Tavazoie, S. (2005) *PLoS Comput Biol* 1(7), e69
5. Bernhart, S. H., Tafer, H., Muckstein, U., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2006) *Algorithms Mol Biol* 1(1), 3

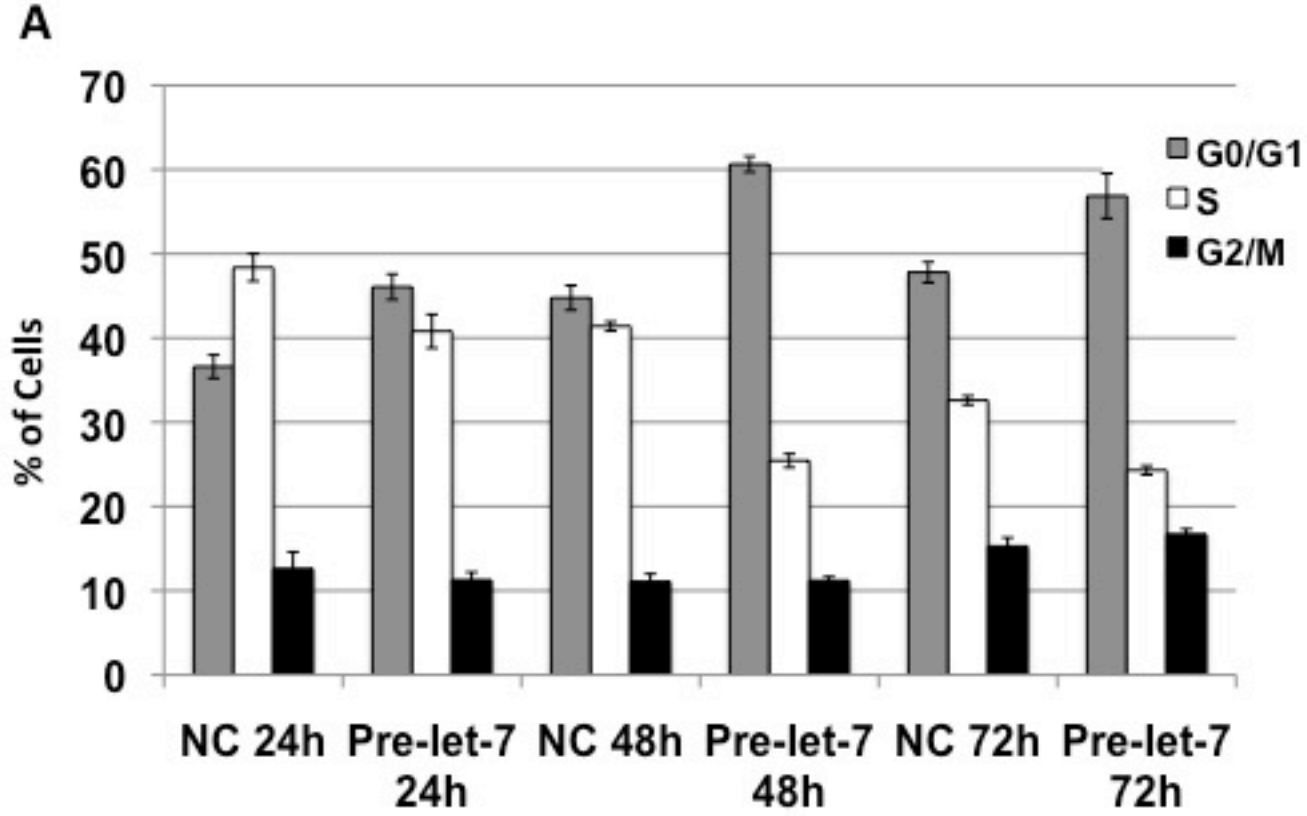
Supplementary Figure S1.

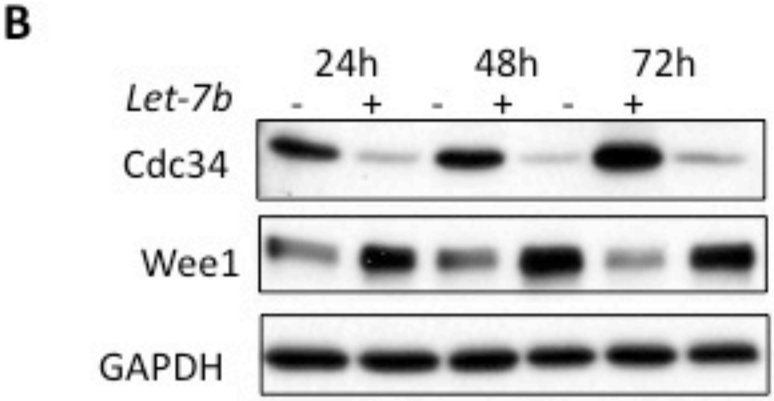


Supplementary Figure S1 cont.

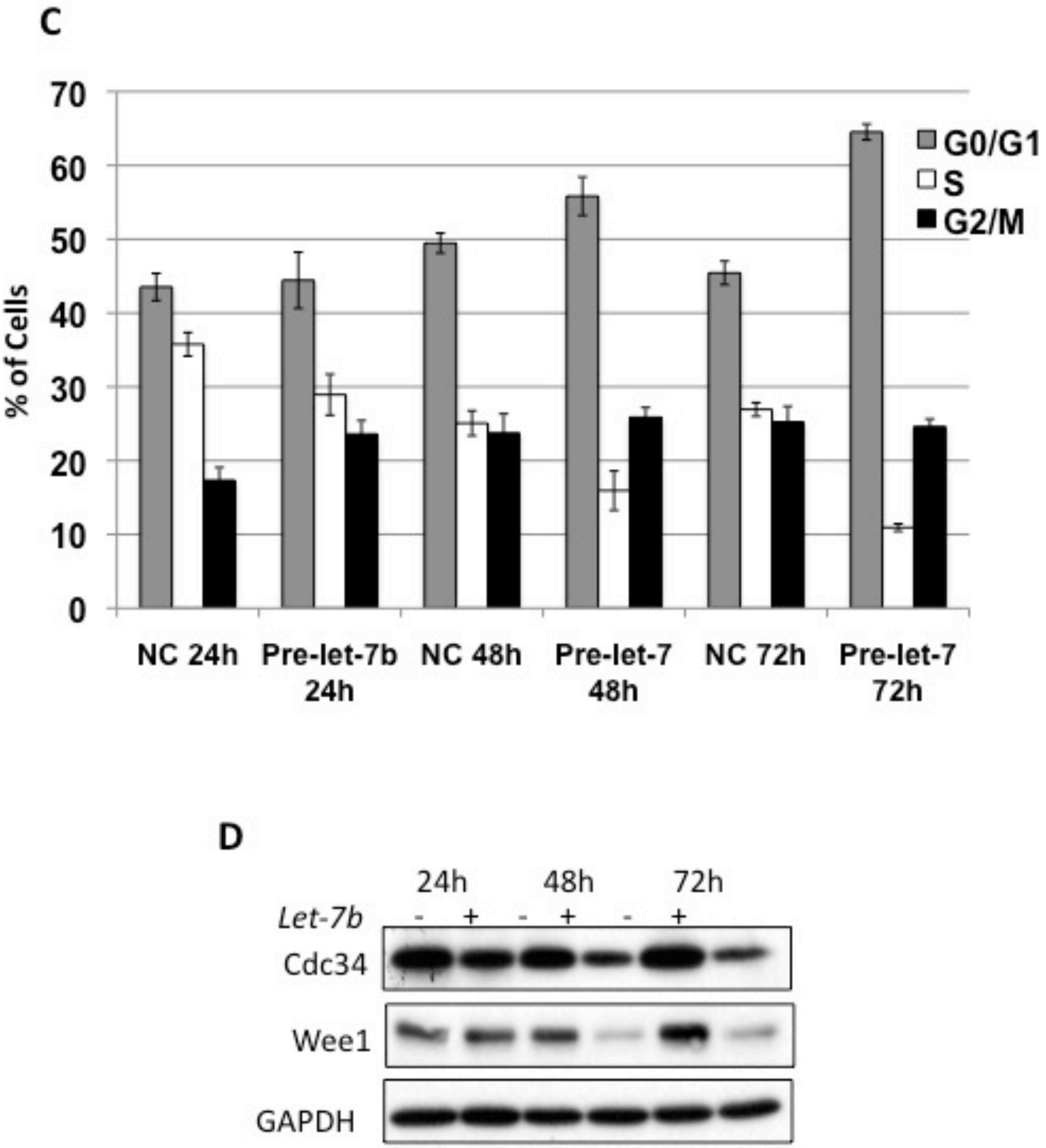


Supplementary Figure S2.





Supplementary Figure S2 cont.



Supplementary Table 2: Correlation coefficients and p-values for the correlation between let-7 binding site density and transcriptional response.

Sample	**Pearson correlation (p-value)**	**Kendall correlation (p-value)**
24 hr	-0.014 (0.68610)	-0.036 (0.1166)
24/36	-0.084 (0.01550)	-0.075 (0.001278)
24/36/12	-0.077 (0.02599)	-0.071 (0.002031)
24/36/24	-0.123 (0.0003452)	-0.119 (2.863e-0.7)
24/36/36	-0.102 (0.003113)	-0.081(0.000482)

Supplementary Table 3. Primers used for PCR amplification in this study.

Primer name	Primer Sequence (5'-3')
CDC34-F	gccactagtcaccaccagaataaacttgccgagtttacctcactagggccggacccgtggctccttagacgacagactacctcacggag
CDC34-m1&m2-F	gccactagtcaccaccagaataaacttgccgagttta**aag**cactagggccggacccgtggctccttagacgacaga**a**ta**aag**cacggag
CDC34-m1-F	gccactagtcaccaccagaataaacttgccgagttta**aag**cactagggccggacccgtggctccttagacgacagactacctcacggag
CDC34-m2-F	gccactagtcaccaccagaataaacttgccgagtttacctcactagggccggacccgtggctccttagacgacaga**a**ta**aag**cacggag
CDC34-R	gccaagcttcataaagtagttttatttagatttcaaacaaaccaaagcagggaaaagagtcggccacggtgaatccgt
LCS-WT-F	ctagtaaccacacaacctactacctcagctcagca
LCS-WT-R	agcttgctgagctgaggtagtaggttgtgtggtta
LCS-m-F	ctagtaaccacacaacctaataaagcagctcagca
LCS-m-R	agcttgctgagctgaggtattattctgtgtggtta

The restriction sites are underlined. Mutations are bold letters.