Supplemental Discussion

Combining signal and background distributions for the 10 UTR bins

Controlling for the local conservation rate is of vital importance in this type of study. Local conservation rate can affect the variability in the background estimate, thereby decreasing sensitivity and statistical power of the analysis. Moreover, the local conservation rate can dramatically distort estimates of preferential conservation (Lewis et al., 2005). To illustrate this distortion, we estimated the conservation signal and background without controlling for UTR conservation, and then plotted the signal-to-background ratio for conserved sites that fall within ten different levels of local conservation (Supplemental Fig. 3, upper panel). These results illustrate how previous methods, which use a single tree and have a single estimate of the background distribution for all UTRs, overestimate the preferential conservation of sites in highly conserved UTRs and can miss the preferential conservation of sites in poorly conserved UTRs. Indeed, when using a single tree and single background estimate, miRNA seed matches in the poorly conserved UTRs appear to be actively avoiding conservation when compared to the background, an observation that is very unlikely to be biological. Therefore, previous methods, which do not control for local conservation, reliably show that many seed-matched sites are preferentially conserved, but they are not reliable in distinguishing individual sites that are preferentially conserved from those that are conserved by chance.

There are multiple reasonable ways to control for local conservation. Our method of separating the UTRs into bins based on their conservation rates raised the question of how to combine the data from these bins. It was not obvious *a priori* that the bins could be treated in the same way. If the bins were not equivalent with respect to the relevant measurement (in this case, branch-length cutoff), the complete analysis might have to be performed separately for each of the ten UTR bins. However, if the bins were equivalent, then the signal and background values for each bin could be safely combined by simply summing at each cutoff to create the aggregate signal and background distributions.

In order to make the UTR bins more comparable, we recalculated the phylogenetic trees separately for each bin. This approach allowed for fine-tuning of the relative branch-lengths, which may provide additional benefits beyond a uniform rescaling of all branch lengths. We reasoned that after this kind of scaling, the signal-to-background ratio would be close to equivalent for all ten bins at a given branch-length cutoff. Indeed, the signal above background reached a maximum at about the same branch-length cutoff (1.0) for each of the bins, after recalculating the trees for each bin (data

not shown). Moreover, the large variability in signal-to-background ratio observed with a single tree was greatly reduced upon recalculating the trees for each bin (Supplemental Fig. 3). Most of the remaining variability could be attributed to edge effects in bin 1 and bin 10. Hence, after recalculating the phylogenetic trees for each bin, we could safely combine the results for the ten UTR bins into one estimate of signal and background, since our confidence in the preferential conservation of a site at any particular branch length was largely independent of the UTR bin to which it was assigned.

Nested seed matches

As schematically depicted in Figure 1D, we nested smaller seed matches within larger ones, which led to a substantial increase in sensitivity. Because many miRNA sites have species-specific differences in seed-match type, a sensitive method was required for determining the largest conserved unit. Our approach was to begin with the largest seed-match class (8mers), and subtract both the signal and the background of the larger seed matches from the signal and background of the smaller ones. Hence, a 6mer conserved to branch length 1.0 contributed to the number of conserved 6mers only if it was the largest functional seed-match unit that was conserved to that branch length; if the 6mer was subsumed in an 8mer conserved to branch length 1.0, it was not counted as a conserved 6mer. In this way, we classified seed match conservation as the longest seed-match type possible, but also allowed for species-specific differences without losing sensitivity.

6mer signal above background

Given the observation of many conserved 6mer seed matches above background, it is natural to ask whether these sites are being selectively maintained or whether there could be other, technical reasons for their preferential conservation. Indeed, because of the methodology discussed above, conserved 6mers may appear in some species as 7mer or 8mer seed matches. This leaves open the possibility that the observed preferential conservation of 6mers could be due to mutation from conserved 7mers, i.e., it could be due to preferential conservation in the 7mer form, with only chance conservation as 6mers.

We have performed two analyses to test whether the 6mer conservation observed could be attributed to decay of conserved 7mers. First, we tested the possibility that 7mer sites in human contribute to 6mer conservation signal through decay in orthologous species by examining only the subset of 6mers that were not part of a 7mer in the human UTR. We found in this subset that there

was significant enrichment in conservation for both canonical 6mer seed matches and offset 6mers (Supplemental Figure 2B). The converse possibility is that the seed match is conserved as a 7mer in other species, but is a 6mer in human. Because of technical issues, this possibility was difficult to evaluate directly. But to get a sense of its impact we reasoned that the number of 6mers in human preferentially conserved as a 7mer in other species that include mouse, would mirror the number of 6mers in mouse preferentially conserved as a 7mer in other species that include human. With this in mind, we performed an analysis examining conserved 7mers that have decayed to 6mers in the mouse UTRs. Testing all possible branch-length cutoffs, we counted the number of such sites, scaled by the proportion of 7mer sites conserved above background (given by (S-B)/S). The cutoff yielding the most 6mer decay (i.e., capturing the most sites above background) was 0.9, corresponding to selectively maintained 7mer sites that are conserved to a branch length of 0.9 but would appear as a 6mer conserved to 1.0 in a mouse-centric analysis. For all 87 broadly conserved families combined, there were 888 decayed 6mers and 362 decayed offset 6mers above background. Symmetrically, one would expect that roughly the same number of conserved 6mers we observed in human are conserved because the site is a selectively maintained 7mer or 8mer in other species. Combining these two sources of error in an aggregate estimate, we still predict 77 6mer seed matches conserved above background per miRNA, and 69 per miRNA for offset 6mers.

By eliminating all 6mer conservation when the human site has a 7mer, some sites that are preferentially conserved as 6mers were surely lost, causing the first analysis to overestimate the number of 6mers conserved due to decay of a human 7mer. In fact, when allowing for single mismatches in 7mers that create 6mer seed matches, extra conservation is added equally to the signal and to the background, yielding roughly the same number of predicted targets as a 7mer conservation analysis. This suggests that when our methods detect a preferentially conserved 6mer that is a 7mer in human, the preferential conservation of the site is due to its presence as a 6mer in other vertebrates and not due to its presence as a 7mer. In the second analysis, in all likelihood there are selectively-maintained 7mers conserved to branch-lengths less than 0.9 that are not accounted for because this preferential conservation is difficult to detect with our methods, potentially leading to an underestimate of the number of human 6mers preferentially conserved only because of their activity as 7mers in other species. In balance, we believe that the overestimate of the first analysis outweighs the underestimate of the second analysis, making our aggregate estimate conservative. Thus, we cannot explain the conservation of 6mer seed matches by their relationship to conserved 7mers.

It is worth noting that in cases found by the above analysis, in which the 6mer preferential conservation might be attributable to conserved 7mers, our estimate for the number of sites conserved

above background and the number of preferentially conserved miRNA targets remains the same. The difference in our two estimates of 6mer conservation above background merely reflects the difficulty of assigning the preferential conservation we observe to individual seed-match types. In cases in which a seed-match site is broadly conserved, but exists as different seed-match types in different species, it is not obvious which type should be assigned the preferential conservation, and in many cases orthologous sites with different seed match types are sure to be simultaneously selectively maintained. Despite this uncertainty, our methods find such preferential conservation with high sensitivity without double-counting, and we have shown that even the weakly effective 6mer and offset 6mer seed matches have substantial conservation independent of the other types.

P_{CT} values reported on the TargetScan website

While calculating P_{CT} values for the TargetScan website, we observed a high variability of P_{CT} values for sites with close branch-length values. This variability was observed for only a subset of miRNAs, and even for those in which it was observed, the strong underlying trend of higher PCT at higher branch lengths was clear, which explains the correlation with the experimental data when looking at the P_{CT} scores in aggregate (Figure 6). Nonetheless, we considered it prudent to implement a smoothing procedure when deriving P_{CT} values reported at the TargetScan website. Thus, for each miRNA and for each seed match type, we fit a modified sigmoid function to the P_{CT} scores using a least squares estimator. The function was given by:

$$P_{CT} = \max\left(0, \beta_0 + \frac{\beta_1}{1 + \exp(-\beta_2 x + \beta_3)}\right)$$

where, x is the branch length value for a particular site. In other words, the P_{CT} score reported on the Targetscan website was either the output of a modified sigmoid function given by $\beta_0 + \beta_1/(1 + \exp(-\beta_2 x + \beta_3))$, or zero if that function was negative. The values of β_0, β_1, β_2, and β_3, were fit so that the function would best match the raw P_{CT} values. β_0 and β_1 can be interpreted as offsetting and scaling the P_{CT}, respectively, whereas β_2 and β_3 can be interpreted as scaling and offsetting the influence of the branch-length value, respectively. There is no special significance to the particular form of the function or the value of the parameters — rather, we observed that the P_{CT} values closely followed this modified sigmoid function, and that in all cases the modified sigmoid function closely matched the data but substantially smoothed curves plotting P_{CT} with respect to the branch-length.

Supplemental Table 1: Broadly conserved miRNA families.

Seed + nt 8	Human miRNAs in family	8mer signal-to-background ratio in 23 vertebrates (cutoff 1.0)	8mer signal above background* in 23 vertebrates (cutoff 1.0)
GAGGUAG	let-7a;let-7b;let-7c;let-7d;let-7e;let-7f;miR-98;let-7g;let-	6.33	235
GGAAUGU	miR-1;miR-206;miR-613	3.08	182
GGAAGAC	miR-7	2.38	79
CUUUGGU	miR-9	3.77	229
ACCCUGU	miR-10a;miR-10b	1.86	25
AGCAGCA	miR-15a;miR-16;miR-15b;miR-195;miR-424;miR-497	3.51	222
AAAGUGC	miR-17;miR-20a;miR-93;miR-106a;miR-106b;miR-	4.59	304
AAGGUGC	miR-18a;miR-18b	2.81	44
GUGCAAA	miR-19a;miR-19b	3.06	316
AGCUUAU	miR-21;miR-590-5p	2.07	49
AGCUGCC	miR-22	3.19	61
UCACAUU	miR-23a;miR-23b	1.86	253
GGCUCAG	miR-24	2.66	130
AUUGCAC	miR-25;miR-32;miR-92a;miR-363;miR-367;miR-92b	4.18	222
UCAAGUA	miR-26a;miR-26b	3.28	247
UCACAGU	miR-27a;miR-27b	2.95	260
AGCACCA	miR-29a;miR-29b;miR-29c	4.22	205
GUAAACA	miR-30a;miR-30c;miR-30d;miR-30b;miR-30e	3.59	514
GGCAAGA	miR-31	1.77	56
UGCAUUG	miR-33a;miR-33b	1.62	29
GGCAGUG	miR-34a;miR-34c-5p;miR-449a;miR-449b	3	158
UUGGCAC	miR-96	3.83	175
ACCCGUA	miR-99a;miR-100;miR-99b	13	Near zero
ACAGUAC	miR-101	3.15	196
GCAGCAU	miR-103;miR-107	2.18	92
GGAGUGU	miR-122	1.7	28
AAGGCAC	miR-124;miR-506	5.63	212
CCCUGAG	miR-125b;miR-125a-5p	4.08	206
CGUACCG	miR-126	7.5	Near zero
CACAGUG	miR-128a;miR-128b	3.96	203
UUUUUGC	miR-129-5p	0.89	Near zero
AGUGCAA	miR-130a;miR-301a;miR-130b;miR-454;miR-301b	3.27	133
UUGGUCC	miR-133a;miR-133b	3.38	130
AUGGCUU	miR-135a;miR-135b	3.09	131
UAUUGCU	miR-137	3.12	258
GCUGGUG	miR-138	3.58	144
CUACAGU	miR-139-5p	1.13	Near zero
AGUGGUU	miR-140-5p	2.34	44
AACACUG	miR-141;miR-200a	2.43	135
GUAGUGU	miR-142-3p	4.45	94
GAGAUGA	miR-143	1.86	70
ACAGUAU	miR-144	1.44	75
UCCAGUU	miR-145	2.19	148
GAGAACU	miR-146a;miR-146b-5p	0.9	Near zero
CAGUGCA	miR-148a;miR-152;miR-148b	3.23	162
CUCCCAA	miR-150	0.89	Near zero
UGCAUAG	miR-153	3.73	135
UAAUGCU	miR-155;	1.62	56
ACAUUCA	miR-181a;miR-181b;miR-181c;miR-181d	2.14	232
UUGGCAA	miR-182	2.67	228
AUGGCAC	miR-183	3	66
GGACGGA	miR-184	2.06	Near zero
CGUGUCU	miR-187	0.65	Near zero

Supplemental Table 1 (continued)

GAUAUGU	miR-190;miR-190b	1.54	23
AACGGAA	miR-191	3.75	Near zero
UGACCUA	miR-192;miR-215	0.99	Near zero
ACUGGCC	miR-193a-3p;miR-193b	2.17	29
GUAACAG	miR-194	1.89	48
AGGUAGU	miR-196a;miR-196b	3.62	33
CCAGUGU	miR-199a-5p;miR-199b-5p	3.5	136
AAUACUG	miR-200b;miR-200c;miR-429	3.14	199
UGAAAUG	miR-203	1.33	52
UCCCUUU	miR-204;miR-211	1.44	81
CCUUCAU	miR-205	1.61	56
UAAGACG	miR-208;miR-208b	2.19	Near zero
UGUGCGU	miR-210	2.42	Near zero
AACAGUC	miR-212;miR-132	2.33	54
CAGCAGG	miR-214	2.1	61
AAUCUCU	miR-216b	1.09	Near zero
AAUCUCA	miR-216a	0.78	Near zero
ACUGCAU	miR-217	1.35	22
UGUGCUU	miR-218	3.25	230
GAUUGUC	miR-219-5p	3.79	82
GCUACAU	miR-221;miR-222	1.78	68
GUCAGUU	miR-223	1.69	41
AAGUGCU	miR-302a;miR-302b;miR-302c; miR-302d;miR-372;miR-373;miR-	2.24	70
CCAGCAU	miR-338-3p	1.14	Near zero
AAUGCCC	miR-365	2.11	30
UUGUUCG	miR-375	2.73	Near zero
GAUCAGA	miR-383	1.03	Near zero
AUGACAC	miR-425	1.07	Near zero
AACCGUU	miR-451	3.08	Near zero
AUGUGCC	miR-455-5p	1.25	Near zero
AACCUGG	miR-490-3p	0.88	Near zero
UAAGACU	miR-499-5p	1.19	16
AGCAGCG	miR-503	5.15	Near zero
CGACCCA	miR-551a;miR-551b	3.75	Near zero

*Values for signal above background of <16 sites were designated as “near zero.” This cutoff was chosen because the most poorly performing 8mer had a background estimate 16 sites greater than the signal, putting a conservative upper limit on the ability to distinguish preferentially conserved sites from background.

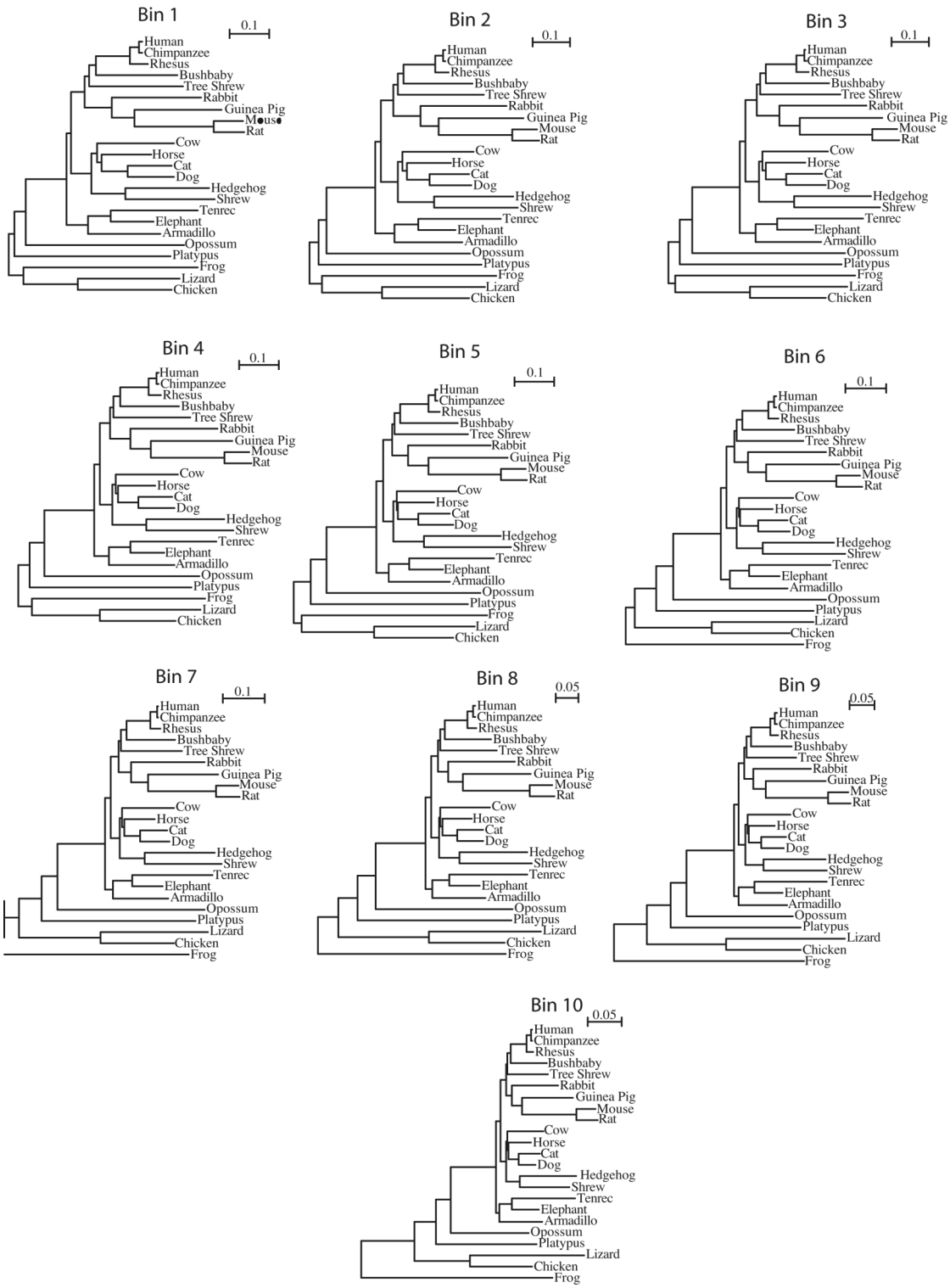
Supplemental Table 2. Mammalian-specific miRNA families.

Seed + nt 8	Human miRNAs in family	8mer signal-to-background ratio in placental mammals (cutoff 0.85)
CAGGUGA	miR-125a-3p	0.61
CGGAUCC	miR-127-3p	2.94
GUGACUG	miR-134	0.85
CUCCAUU	miR-136	0.84
AGGUUAU	miR-154	0.93
GGAGAGA	miR-185	1.10
UCACCAC	miR-197	1.24
AGGAGCU	miR-28-5p;miR-708	1.40
AGGGUUG	miR-296-3p	0.68
AUGUGGG	miR-299-3p	0.85
GCAUCCC	miR-324-5p	0.69
UGGCCCU	miR-328	1.14
CUCUGGG	miR-330-5p;miR-326	1.80
CAAGAGC	miR-335	1.67
CCCUGUC	miR-339-5p	1.01
UAUAAAG	miR-340	1.46
CUCACAC	miR-342-3p	0.69
GUCUGCC	miR-346	1.10
UAUCAGA	miR-361-5p	1.34
ACACACC	miR-362-3p;miR-329	1.33
CCUGCUG	miR-370	1.29
CUCAAAC	miR-371-5p	0.71
UAUAAUA	miR-374a;miR-374b	0.88
UCAUAGA	miR-376a;miR-376b	1.14
ACAUAGA	miR-376c	0.69
UCACACA	miR-377	1.16
CUGGACU	miR-378;miR-422a	0.99
GGUAGAC	miR-379	1.46
AUACAAG	miR-381;miR-300	1.38
AAGUUGU	miR-382	1.16
UUCCUAG	miR-384	0.84
AUAUAAC	miR-410	1.05
UCAACAG	miR-421	1.19
GUCUUGC	miR-431	1.27
UCAUGAU	miR-433	1.30
UUUGCGA	miR-450a	0.59
GAGGCUG	miR-485-5p	0.92
AUCGUAC	miR-487b	0.77
UGAAAGG	miR-488	1.02
GUGGGGA	miR-491-5p	0.75
GAAACAU	miR-494	0.89
AACAAAC	miR-495	1.28
GAGUAUU	miR-496	1.01
GACCCUG	miR-504	1.25
GUCAACA	miR-505	1.29
GAGAAAU	miR-539	1.16
GUGACAG	miR-542-3p	1.14
AACAUUC	miR-543	1.38
UUCUGCA	miR-544	1.28
UGUUGAA	miR-653	0.85
UUGUGAC	miR-758	1.27
UGCCCUG	miR-874	1.30
GGAUUUC	miR-876-5p	1.01

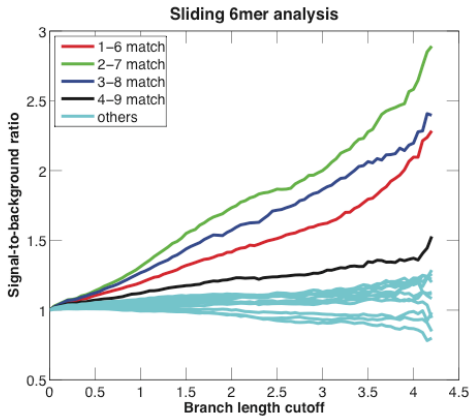
Supplemental Table 3: miRNA families with intermediate conservation. These families were not found in enough non-mammalian genomes to be placed in the broadly conserved set (Supplemental Table 1), yet they were found in too many non-placental animals to be part of the mammalian-only set (Supplemental Table 2).

Seed + nt 8	Human miRNAs in family	8mer signal-to-background ratio in placental mammals (cutoff 0.85)	Notes
CUGGCUC	miR-149	2.14	non-placental mammal conservation
AAAGAAU	miR-186	1.06	found in opossum, platypus
GAGGUAU	miR-202	4.30	seed changes in mammals and in fish
AAGUCAC	miR-224	1.46	non-placental mammal conservation
AAAGCUG	miR-320	2.23	non-placental mammal conservation
AGUAGAC	miR-411	1.79	non-placental mammal conservation
AUGACAC	miR-425	1.00	found in opossum, platypus, lizard
UGCAUAU	miR-448	1.91	non-placental mammal conservation
CCUGUAC	miR-486-5p	1.35	non-placental mammal conservation
AGCAGCG	miR-503	5.45	non-placental mammal conservation
CGACCCA	miR-551a;miR-551b	1.67	found in opossum, platypus, chicken
AAUUUUA	miR-590-3p	0.76	found in platypus
UUGUGUC	miR-599	1.01	found in chicken, platypus, opossum
CCGAGCC	miR-615-3p	1.72	chicken, platypus, lizard alignment
CAGGAAC	miR-873	1.06	found in opossum, platypus
AUACCUC	miR-875-5p	0.97	found in opossum, platypus, lizard

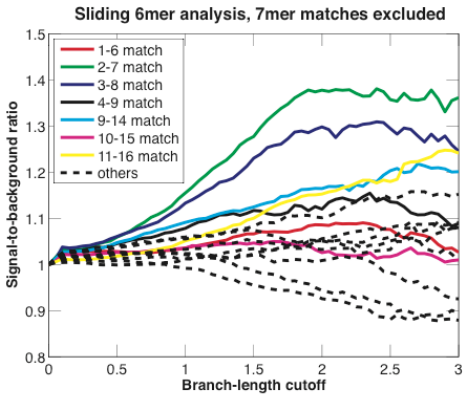
Supplemental Figure 1. Trees for the ten UTR bins, with bin 1 being least conserved and bin 10 being most conserved.



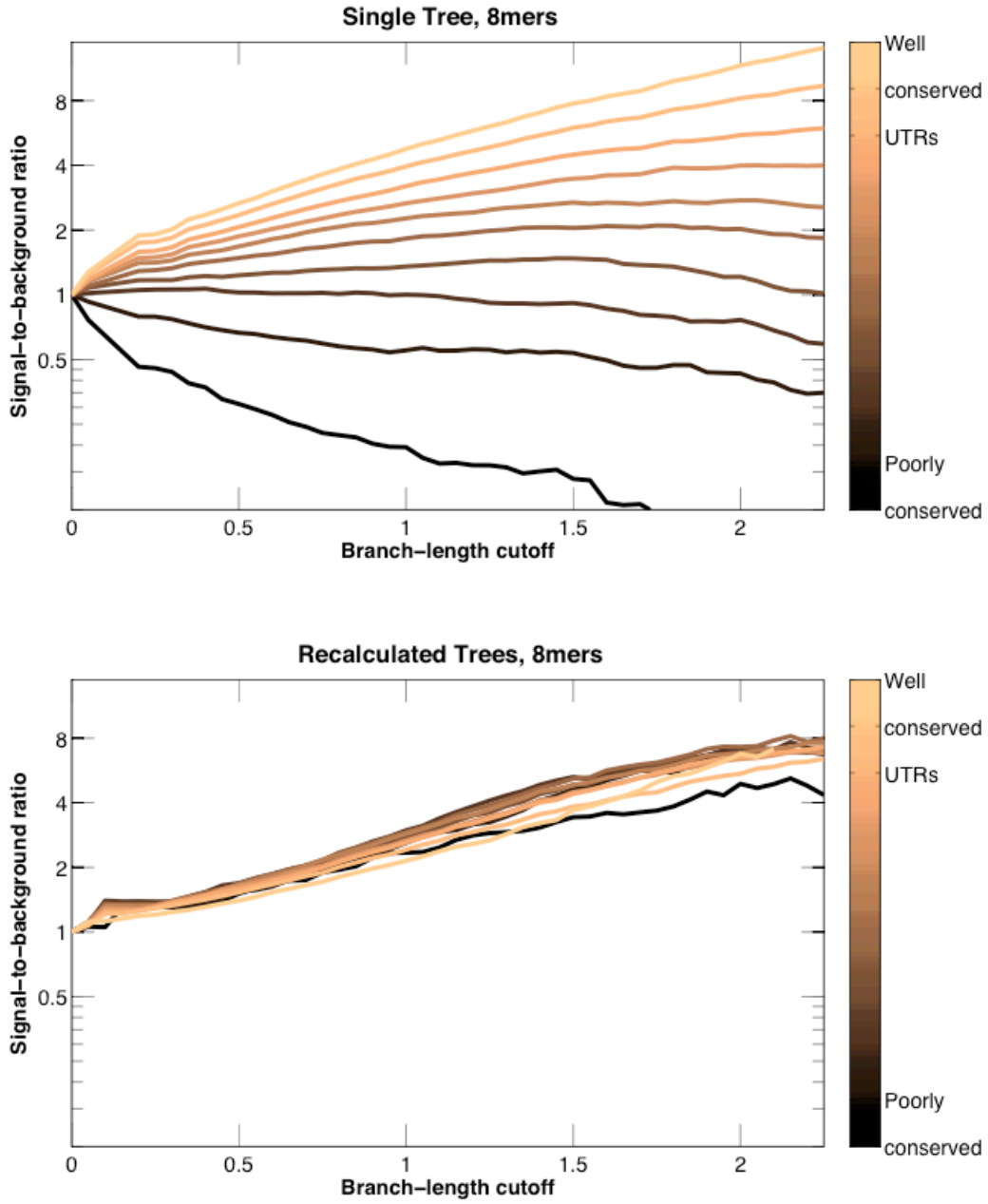
A



B



Supplemental Figure 2. A systematic analysis of matches to each 6-nt segment across the 87 broadly conserved miRNA families. (*A*) Analysis performed without excluding those sites that also possessed canonical 7mer matches. Comparison to panel B indicates that much of the preferential conservation is due to overlap with parts of larger, canonical sites. (*B*) Analysis performed excluding those sites that also possessed canonical 7mer matches, revealing no other segment with appreciable enrichment in conservation.



Supplemental Figure 3. Reduced variability of signal-to-background ratios when using UTR bin-specific trees. Top panel: Analysis performed using a single tree, which was estimated using all UTRs. UTRs were divided into ten equally populated bins, based on their conservation rates, and the signal-to-background ratio for 8mer sites matching the 87 broadly conserved miRNAs is plotted separately for each bin. For each UTR bin, the fraction of conserved sites (the signal) was divided by the fraction of conserved controls (the background), using the same background estimate for all ten bins, which was the overall background, estimated using all UTRs. Bottom panel: As above but signal and background were calculated for each UTR conservation bin with

a unique tree (Supplemental Fig. 1), estimated using only the UTRs from that bin. In other words, sites were compared only with background in the same UTR bin. Note that the bin that deviates most at high cutoffs is the one with the most poorly conserved UTRs. Because at high cutoffs this bin had too few conserved sites for reliable signal-to-background determination (less than one site per miRNA family conserved at branch-length cutoffs exceeding 1.5), its deviation at high cutoffs is not informative, and even if it were informative, it would involve too few sites to be of concern.