SUPPLEMENTARY INFORMATION

Whole-genome alignment between human, mouse, rat and dog

We constructed a whole-genome alignment for the four mammalian genomes using the program Blastz[1] and Multiz[2] in two steps: we first aligned human and dog sequences based on the human/dog syntenic map we generated (to be reported elsewhere, see Lindblad-Toh et al[3]). We then aligned the human/dog sequences to human/mouse/rat three-way alignment downloaded from UCSC genome browser (http://genome.ucsc.edu/) using the profile alignment provided in Multiz package. The assemblies used for this alignment are hg16/mm4/rn2/canFam0. We also tested results of the work in the four-way alignment released on the UCSC genome browser (hg17/mm5/rn3/canFam1). The full alignments are available from the UCSC genome browser.

Aligned promoter and 3'-UTR databases

We constructed the aligned promoter and 3'-UTR database by extracting the portions of the genome-wide alignment whose coordinates correspond to the promoter and 3'-UTR regions respectively. These coordinates were obtained from the annotation of NCBI reference sequences (RefSeq)[4,5].

For promoter regions, we extracted the 4kb segment centered around the annotated transcription start site (TSS) of each human RefSeq gene. If the annotated translation start codon was within 2kb of the TSS, then the shorter region was selected that did not overlap the protein-coding sequence. The 4kb region was chosen to allow sufficient coverage of real promoter sequences, given the small uncertainties in the experimental annotation of TSS sites for the majority of genes. However, choosing such a large segment also increased the percentage of nonfunctional sequences included; in particular, the 4kb region typically includes upstream intergenic segments devoid of regulatory elements, a portion of the 5'-UTR, and possibly portions of the first intron in the case of small non-coding first exons. For genes with alternatively spliced first exons, we included all promoters; when these overlapped, we included the overlapping portion only once. Thus, we ensured that no more than one copy of any promoter was included in the aligned databases, and that our statistical discovery methods were unbiased.

Another advantage of using a 4kb interval for the promoter alignments is that it accounts for any variability in transcription start sites (TSS) across species. To estimate this variability, we used 7878 orthologous human/mouse RefSeq pairs, for which the TSS was mapped in both species. We examined the distance between the two in the aligned promoter database. We found that 92% of TSS pairs were within 500 bp in the alignment, and 82% of them were within 200 bp (see Fig. S1). Since the aligned promoter database covered 2K bp around TSS of each gene, we reasoned that this range should be sufficiently large to include the functional TSS and promoter-proximal regulatory motifs in all four species.

For 3'-UTRs, we extracted the region of the alignment corresponding to the annotated human 3'-UTR in RefSeq, between the translation stop and the transcription stop, excluding introns. For alternatively spliced genes with multiple 3'-UTRs, we included every annotated 3'-UTR segment. When multiple segments overlapped, we included the overlapping portions only once. Thus, we ensured that no more than one copy of any 3'-UTR was included in the aligned databases, and that our statistical discovery methods were unbiased.

Properties of the multiple alignment

We measure the proportion of aligned bases as the number of human bases participating in a local alignment across the four species (possibly with gaps), divided by the total number of nucleotides in the human. The overall proportion of aligned bases across the four genomes is 28% of the human, which is lower than the 40% aligned between human and mouse[6]. Thus, 70% of the sequences conserved between human and mouse are also conserved across the four species. This decrease in coverage is expected, and likely results from the transposable elements gained and lost in each of the lineages. A small fraction of the unaligned regions can also be attributed to possible sequence gaps in the dog genome, which will be described in detail elsewhere[3].

In promoters and 3'-UTRs, the proportion of aligned bases was considerably higher, respectively ~51% for promoters and ~73% for the 3'-UTRs. This higher proportion compared to the rest of the genome is likely due to the numerous conserved regulatory elements in these regions. Also, the lower proportion in promoters as compared to 3'-UTRs is likely due to the inclusion of a higher percentage of nonfunctional sequences in the promoter database, due to the large region aligned around the TSS; this region includes both non-functional intergenic sequence, as well as less-conserved first introns.

Within aligned bases, we measured the evolutionary divergence of the four species, and constructed evolutionary trees, both in promoters and 3'-UTRs. Given the alignment, we used the program ClustalW[7] (http://www.ebi.ac.uk/clustalw/) to infer phylogenetic trees.

Motif conservation score (MCS)

We represent regulatory motifs as consensus sequences (profiles), over an alphabet of 11 characters, consisting of the four nucleotides A,C,G,T, the six two-fold degenerate characters S=[CG], W=[AT], Y=[CT], R=[AG], M=[AC], K=[GT], and the four-fold degenerate character N=[ACGT]. An occurrence of a motif ***m*** is a sequence (over the alphabet ACGT) which matches the consensus of motif ***m*** at every position, namely contains one of the nucleotides allowed by the degenerate code at that position. The terms 'motif occurrence', 'motif instance', or 'match to the motif' are equivalent.

We define a conserved occurrence of a motif **m** as an instance of the motif in the human genome, for which an exact match to the motif is present in each of the four species. For fully specified motifs, this implies that the sequences are identical across the four species. For motifs with degenerate positions (containing ambiguity codes), all sequences need to match the motif, but they do not need to be identical to each other; the four species can contain different variants of the degenerate positions.

We define the conservation rate of a motif **m** as the number of human occurrences of **m** which are conserved across all four species, divided by the total number of human occurrences of **m**. We compute this conservation rate in aligned promoter regions, 3'-UTRs, and introns. Hence, the total number of human occurrences is computed only within these regions, and only for aligned human segments.

We evaluate the Motif Conservation Score (MCS) of a motif **m** of given length and degeneracy, by comparing its conservation rate **p** to the expected rate $\mathbf{p_0}$, estimated using similar random motifs of the same length and degeneracy (see below). Given the rate $\mathbf{p_0}$**,** we evaluate the binomial probability of observing **K** conserved instances out of total **N** instances in the human sequence for motif **m**. We report the MCS as a Z-score defined as MCS = $(\mathbf{K}\text{-}\mathbf{Np_0})/[\mathbf{Np_0}(1\text{-}\mathbf{p_o})]^{1/2}$, which measures the number of standard deviations of conserved instances away from what is expected by chance when the null model is assumed to be binomial. Motifs with high motif conservation scores, are both highly conserved and frequently occurring, resulting in both an increased rate, and sufficient statistical significance given the large counts.

To estimate the conservation rate $\mathbf{p_0}$ expected for a motif **m** of given length and redundancy, we observe the average conservation rate of 1000 random motifs of the same length and redundancy. To account for nucleotide compositional biases in the human genome, we generate these motifs by sampling the human genome. Namely, we select 1000 loci in the four-way species alignment, and extract the human sequences for each of these loci. Based on the degeneracy levels of **m**, we generate a motif for each of these sequences, selecting a degeneracy code for each position matching the sequence of the human locus, and the degeneracy level of **m** at that position. For example, if the first character of **m** is two-fold degenerate and the first nucleotide at the selected locus is A, we pick a two-fold degenerate base containing A (W, R or M), and so on for every character of **m**. We then evaluated, for every locus, whether the resulting random motif is conserved in the other three species, and summed across the 1000 loci. This total number of conserved motifs, divided by the 1000 randomly constructed motifs, was used to estimate the expected conservation rate $\mathbf{p_0}$, under a random model.

We evaluated the MCS separately for each type of region (promoters, introns, 3'-UTR). This ensured that we match the specific nucleotide composition of each type of region, and therefore do not introduce biases in our scoring scheme. Additionally, for promoter motifs, we evaluated the MCS separately for sequences inside CpG islands and those outside CpG islands, to account for their radically different nucleotide compositions. Majority (~80%) of aligned promoter sequences were located outside CpG islands. To boost signal-to-noise-ratio for those sequences, we further used a sliding window of 50 bp and masked those with average nucleotide percent identify less than 60% across the four aligned species, and searched motifs and evaluated MCS only in nonmasked sequences. We defined CpG islands based on their coordinates that were downloaded from the UCSC genome browser.

Identifying conserved motifs through extensive consensus search

We developed a method for identifying conserved motifs by exhaustive enumeration and testing of short sequence patterns. We enumerated all motifs of length between 6 and 26, over an alphabet of 11 characters (the four bases A, C, G, T, the six two-fold degenerate IUB codes R=[AG], Y=[CT], K=[GT], M=[AC], S=[GC], W=[AT], and the four-fold degenerate character N=[ATCG]. The number of motifs that can be formed by combining the 11 letters with various lengths is enormous, but it was still possible to screen most of them because only a small subset of them actually occurred in the database. We started by hashing the positions of all 6-mer motifs, possibly with gaps, and then searched and computed the MCS score for all possible extensions of these 6-mers. The method consisted of the following steps:

(a) We first search and index all positions in the human genome containing a fully-specified 6-mer seed, possibly with a central gap between 0 and 10 non-specified bases. These seeds are of the form UVW-gap-XYZ, where U,V,W,X,Y,Z can be any nucleotide. This resulted in a total number 45,056 six-mers.

(b) For each of these seeds, we extracted the four-way aligned sequence containing the aligned seeds and their neighboring sequences extending 5 nucleotides on each end.

(c) We then enumerated all motifs that contain one of these seeds and have more than one instance in the aligned genomes.

(d) We finally tested the conservation statistics of each of the resulting motifs and selected all motifs with MCS above 6.0

Choosing an MCS cutoff

We chose MCS>6 as a cutoff for motif discovery. This cutoff was selected based on the excess conservation shown in red in figures 1 for promoters and 3'-UTRs. It was selected to capture most of the distribution in the excess conservation (red), while minimizing the non-excess motifs (white) above this cutoff. The table shows the number of 'red' and 'white' motifs, above and below the MCS cutoff of 6.

	MCS<6	MCS>6	Total
Red (excess conservation motifs)	FN=3.94	TP=8.85	12.79
White (expected in a Gaussian model)	TN=87.03	FP=0.18	87.21
	N=90.97	P=9.03	100

For MCS>6, we capture 69.2% of the excess conservation (sensitivity=8.85/12.79=69%), while ensuring that the vast majority of motifs above this cutoff are indeed 'red' motifs (specificity=8.85/9.03=98.1%).

Thus, MCS>6 is a highly specific cutoff (98.1% specificity). Increasing the cutoff would result in lower the sensitivity, missing many real motifs.

Motif clustering

After the motif enumeration and selection step, the resulting motifs with MCS > 6 were highly redundant, since similar motifs could be derived by extending different 6-mers. We clustered these to obtain a non-redundant set.

We grouped the discovered motifs into clusters using two steps: genome-wide co-occurrence, and sequence similarity. We first used the genome-wide co-occurrence step to eliminate motifs that are largely redundant. Namely, if the genome-wide occurrences of two motifs overlapped by more than 80% of their sites, then we only kept the motif with the highest MCS score, and ignored the lower-scoring motif. We then clustered the remaining motifs based on their pairwise sequence similarity.

We evaluate the sequence similarity between two motifs as the Pearson correlation of their equivalent position weight matrices[8]. We first convert the consensus representation of each motif to the equivalent positional weight matrix, representing the frequencies of the four bases at each position of a motif. For example, if the first position of a motif was Y=[CT], the first column of the weight matrix would be [A, C, G, T]=[0, 1/2, 0, 1/2], and so on for each position. We then represent each motif of length L using a single vector, by concatenating the columns of its weight matrix (obtaining a vector of length $4*L$). We then compute the Pearson correlation[9,10] between every alignment of two motifs, as they are scanned past each other, in both strands. At each alignment offset, we extended the motif vectors using nucleotide background frequencies so that all positions of two aligned motifs are matched. We then report the similarity score as the the highest Pearson correlation across all alignments. This score ranges from –1 to 1 and is maximal when the two motifs are exactly the same.

To form the clusters, we visited every motif in the order of decreasing MCS score, and compared each of them with the previous motifs visited. If a match was found between the current motif **m** and a previously visited motif **n** above similarity score 0.75, then motif **m** was considered as a variant of motif **n**, and grouped with it. We continued thus until all motifs with MCS > 6 were grouped into clusters. For each cluster, we selected a representative motif as the one with the highest MCS. Finally, to reduce redundancy of motifs contained in the same cluster, we removed motifs that shared more than 0.85 similarity score with the cluster representative.

Coping with nucleotide compositional biases

Genome-wide motif discovery in the human poses a number of challenges, especially stemming from the widely varying compositional biases found in the human. Importantly, CpG islands in human promoters have widely different sequence composition, di-nucleotide composition, and conservation properties than the rest of the genome. In this section, we specifically evaluate concerns about how these biases have affected our motif discovery.

(1) To account for the important variations in sequence composition that stem from CpG islands, we partitioned each promoter region into a portion associated with CpG-islands (if any), and the remainder of the promoter. We then calculated the MCS separately in three types of region (3'-UTRs, CpG-associated promoters, non-CpG-associated promoters). Thus, high-scoring motifs within CpG islands were those that showed significantly conservation when compared to other motifs in CpG islands.

(2) To account for the di-nucleotide, tri-nucleotide, and higher-order markov properties of the human genome, we constructed random motifs by directly sampling from the genome itself. For every motif m of length L, we sample 1000 regions of the genome (each of length L), hence capturing the di-nucleotide composition of these regions (this holds for 3'-UTRs, CpG-promoters, and non-CpG-promoters). Hence, when estimating the expected conservation rate of random motifs, we take into account the specific di-nucleotide properties of the human genome, in the particular region studied.

(3) Additionally, we asked whether motifs containing CpG have different conservation properties, but a first examination shows that in fact it is not the case. We considered the top 50 motifs (ranked by MCS), and counted the representation of CG di-nucleotides. We found that CG appears 23 times, out of 394 di-nucleotides in these motifs (6% of occurrences), which is nearly identical to what one would expect if all di-nucleotides were equally likely (394/16=24 times). Hence, our computational algorithm is not favoring CG di-nucleotides in the most high scoring motifs.

(4) We also addressed similar concerns regarding the 3'-UTR regions. We compared the di-nucleotide counts of the top 30 and bottom 30 3'-UTR motifs, and found a correlation of R^2=0.9. Again, nucleotide composition doesn't seem to affect the MCS. We have also addressed these comments in the supplementary information.

(5) We then considered whether Transfac motifs may score poorly due to their different motif composition. We compared the di-nucleotide compositions of high-scoring Transfac motifs (MCS > 5) with the di-nucleotide composition of low-scoring Transfac motifs (MCS < 5). We found that the two sets have indeed very similar compositions. We quantified this observation by calculating the auto-correlation of the 16 di-nucleotide counts for each of the two distributions, and we found R^2=0.65, which is remarkably strong. Hence, high-scoring and low-scoring Transfac motifs have largely the same di-nucleotide composition. For CpG di-nucleotides in

particular, the counts were 22 and 19, respectively.

In summary, our computational and statistical methods were designed to capture the variability in sequence composition, both at the regional level, as well as the nucleotide level in the human genome. The end result is a motif discovery algorithm which is unbiased with respect to at least the most apparent sequence artifacts of the human genome.

Evaluating MCS for TRANSFAC motifs

We extracted 460 mammalian transcriptional regulatory motifs from the TRANSFAC database (version 7.4, http://www.gene-regulation.com/), represented by positional weight matrices[8].

We first collapsed the highly redundant set of motifs, using the same method and thresholds as for the discovered motifs (see Motif clustering section). This resulted in a smaller set of 123 motifs (shown in Table S1) using the weight matrix similarity measure described above (see Motif Clustering section).

To evaluate the MCS conservation score of Transfac motifs, we used the same method described earlier (see Motif Conservation Score (MCS) section), in terms of its excess conservation ***K/N*** as compared to the expected background rate $\boldsymbol{p_0}$. The increased challenge was that Transfac motifs are described in terms of position weight matrices (PWM), not consensus sites. We therefore developed ways to (1) determine the conserved and non-conserved occurrences of a PWM motif, to obtain K and N, and (2) determine the expected neutral conservation rate for random similar matrices.

(1) We developed a computational method to evaluate whether a site matched a motif described by its position weight matrix. To do so, we used a log ratio test, comparing the likelihood that a site was generated by a given Transfac motif, as compared to the likelihood that the site was generated by a neutral background model. If the log ratio score between the two probabilities was above a given threshold, we counted the site as a match. Summing all the matches gave us ***N***, the number of total occurrences in the human. If additionally, the site matched in all species, the site was counted as a conserved occurrence, and the sum of these gave us ***K***. We used the following formula to determine the threshold of log ratio score: $\theta = \min L + 0.7(\max L - \min L)$, where max*L* and min*L* were the maximum and minimum log ratio scores the weight matrix could possibly achieve, depending on its total information content and its nucleotide composition.

(2) To compute the neutral conservation rate of a weight matrix, we used a sampling method. For each weight matrix, we randomly permuted the columns representing weights for each of four bases at each position, independently, to generate a set of control weight matrices[11]. The control set preserved the overall information content of the original weight matrix, but changed the nucleotide preferences at each column. We searched the control weight matrices counting total and conserved matches determined by log ratio score with the same threshold as the original weight matrix. The conserved number divided by the total number was used as an estimation of neutral conservation rate $\boldsymbol{p_0}$.

We generated a database describing these matches of Transfac motifs. For every Transfac motif, the annotated occurrences, their corresponding log ratio scores, and the conservation were superimposed with the aligned promoter sequences. The data can be downloaded from: http://www.broad.mit.edu/seq/HumanMotifs/.

Comparing the discovered motifs to the TRANSFAC motifs

We compared the 174 discovered motifs in the promoter database to TRANSFAC motifs using the motif comparison method described above (see the Motif clustering section), first converting the discovered motifs to position weight matrices, and then computing the corresponding Pearson correlation. If a motif matched one of the TRANSFAC motifs with similarity score above 0.85, we marked it as a strong match; otherwise, if one of its co-clustered motifs had a strong match to TRANSFAC motifs, we marked it a weak match. Overall, 72% of the 123 known TRANSFAC motifs showed matches to the highly conserved motifs. To estimate the probability of producing the observed number of matches by chance, we generated a TRANSFAC-like database of control motif, using the same procedure as for generating random motifs (see Motif Conservation Score section). For every TRANSFAC motif, we sampled a random human promoter segment, and constructed a random motif which matches the human segment, and whose degeneracy levels match the Transfac motif used. This procedure ensures that the random control motifs preserve the di-nucleotide composition of human motifs, and the same levels of degeneracy as Transfac motifs.

Motif gene set enrichment analysis for expression data

We evaluated the tissue-specificity of each regulatory motifs by calculating the tissue-specificity of its target gene set, in a gene expression atlas of 75 human tissues[12]. We first preprocessed the expression data by normalizing the expression of each gene across all tissues to be mean zero and variance 1. We then ranked the genes based on their normalized expression values for each tissue, giving rise to 75 ranked gene lists.

For each motif **m**, we generated three gene sets: a target gene set S_1, and two control gene sets S_2 and S_3, with the same number of genes.

(1) We first generated the motif gene set S_1 of 'conserved instances', consisting of the inferred target genes for each motif. This set consisted of all genes whose promoters contained at least one conserved instance of the motif **m**.

(2) We then generated a control gene set S_2 of 'non-conserved instances', by randomly sampling from genes containing non-conserved instances of the motif, until S2 contained the same number of genes as S1.

(3) We also generated a second control gene set S_3 of 'shuffled conserved instances', by randomly sampling genes from the union of all conserved gene sets (S_1), for all motifs.

We used the two control gene sets to evaluate the statistical significance of the tissue enrichment observed in the target gene set S_1, as compared to two similar but random gene sets with the same cardinality S_2 and S_3.

We evaluated the enrichment of a motif **m** in a given tissue, as the enrichment of its gene set ***S*** in the ranked list for that tissue. We used the Mann-Whitney rank sum statistic[13] to evaluate the non-randomness of the ranks of ***S***, in the list ***L*** specific to that tissue. We sum the ranks of genes in ***S*** that appear in list ***L***. The significance of the rank sum is tested against rank sums of random subsets of the list ***L***, randomly permuted. Let μ and σ^2 be the mean and variance of the control rank sums. We define the Motif Gene Set Enrichment (MGSE) score to be $(\mu\text{-}S)/\sigma$, that is, the number of standard deviations smaller than the mean. This statistic is strongest when the items in ***S*** are ranked at the top of the list ***L***.

For each motif, we computed the MGES for S_1, S_2, and S_3 in all 75 tissue-specific ranked gene lists. For the motif target list S_1, the best MGES among all tissues is annotated in Table 2 (if the score was above 4.0 SD). We also computed the best MGES scores for the two control sets S_2 and S_3, and we found that their scores were indeed much less than the target gene sets S_1 (Fig. S2). Only a few non-conserved control sets S_2 in the beginning of the motif list show enrichment score significantly higher than those from randomly permuted sets. The motifs corresponding to those sets have consistently high conservation rates. It is likely that the consensus sequences of these motifs are specific enough to indicate functionality, regardless of conservation.

Motif positional bias in promoters

For every motif m, we tested the presence of a positional bias in the distance distribution between its instances and the TSS. We identified all sites where a motif occurred in human promoters (without requiring conservation) and recorded their positions relative to TSS. We then divided the region (-2000, 2000) bp around TSS into 100 bins, and counted the number of sites located in each of the bins. We computed the mean and variance on the distribution of the number of sites in different bins, and converted the number of sites in each bin to a Z-score measuring the number of standard deviations away from the mean. Positional clustering of the motif was counted as significant if there existed a bin with Z-score above 5.0, in which case the biased position was determined by the location of the bin.

Conserved 8-mer motifs in 3'-UTR

We evaluated the conservation rates of all 8-mers (total 65,536) in the 3'-UTR, and selected 540 8-mers (0.8% of all) with conservation rate above 0.18 (vs 0.076 for random 8-mers) and having at lease 6 conserved instances. Many of these 8-mers were highly similar to each other, and we clustered them based on their sequence similarity. We used a stringent criterion for clustering, requiring that all 8-mers in a cluster share at least six consecutive nucleotides with the cluster representative (the 8-mer with the highest conservation rate). This resulted in 72 clusters of 8-mers, each with a cluster representative. We used the representative motif to refer to the set of motifs contained in the same cluster.

Estimation of the number of miRNA targets

We observed that about 40% of human 3'-UTRs contain at least one copy of the conserved 8-mers. We used a control set of random 8-mers, with equal number of motifs, to estimate the number of 3'-UTRs that could be hit by chance because of basal conservation rates of random control motifs. Since most of the conserved 8-mers discovered in 3'-UTRs had strong strand-bias (Fig. 4A), we used their reverse complemented sequences as our controls, which preserved CG content and basic nucleotide compositions of the conserved 8-mer motifs. We found about 25% of human 3'-UTRs contained one of the conserved control motifs.

Let p be the proportion of 3'-UTRs with a biologically meaningful miRNA target. Then, 1-p is the proportion without biologically meaningful target. Since the frequency of conserved control occurrences is 25%, the proportion of these genes with a conserved site is (1-p)*0.25. We thus have p + 0.25 (1-p) = 0.40, so p = 0.20. This estimated that about 20% of human genes were targeted by miRNAs.

MicroRNA datasets and pairing of conserved 8-mers to miRNAs

A set of 207 human miRNAs representing 222 human miRNA genes was downloaded from Rfam miRNA registry (Release 5.1, http://www.sanger.ac.uk/Software/Rfam/mirna/).

For each of the miRNAs, we identified all matching 8-mers with Watson-Crick (W-C) pairing from the list of 540 most conserved 8-mers discovered in 3'-UTR. We found that 90 of these miRNAs (43%) have matches within these 8-mers. For comparison, we evaluated the pairing of miRNAs to three control sets of 8-mers with equal number of motifs (540): a random set, the set of most conserved 8-mers from 5'-UTRs and the set of most conserved 8-mers from coding exons. These matched to 2%, 3.7% and 9.6% of miRNAs respectively.

Moreover, we found that when the 8-mer motifs matched known miRNAs genes, they matched specifically the first two positions from the 5' end of the miRNA in 95% of the time (Fig. 4d). To reduce the chance of random pairing, we further identified 8-mers that matched to 5' miRNAs by restricting W-C pairing at only the first two positions (Table S6). For miRNAs that did not match to a conserved 8-mer, we relaxed the requirement of strict W-C pairing and allowed one mismatch. The list of miRNAs with one-base mismatched 8-mers is shown in Table S6 with mismatched bases indicated.

Identification of new miRNAs

We then sought to identify new miRNA genes based on the 540 highly-conserved 8-mers discovered in 3'-UTRs. We first identified conserved occurrences of the 8-mer motifs in the entire human genome, searching both strands for motifs reverse complementary to each 8-mer. In this search, we excluded genomic positions that overlapped annotated genes.

We then searched for stable stem-loops in neighborhoods of these alignments. We extracted the aligned neighborhoods of these conserved sites with 100 bp on each side. A sliding window of 110 bp with an increment of 3 bp was scanned along the extracted sequences. The windows containing the motif sites were folded using the program RNAfold[14], and those with a folding free energy of at least 25 kcal/mol in all aligned species were selected. Each identified window was further examined for pairing and alignment of the core 22-mer sequence containing the original motif at 5' end. We selected the windows whose core sequences were located only in one stem of the folded RNA structure, formed at least 16 base-pairings, and had at least 18 bases conserved in four species. The regions passing these criteria were selected as conserved stable stem-loops.

A total number of 440 conserved stable stem-loops were identified, including 124 known miRNA genes (56% of the total 222). The list included almost all known miRNAs that matched to conserved 3'UTR 8-mer motifs in the previous search, except a small number that was missing due to sequence gaps in one of the mammalian genomes.

We further evaluated these 440 stem-loops using the program MiRScan[15]. For each stem loop, we compared the sequence of human to the aligned sequences of mouse, rat and dog, and scored the three pairs using MiRScan. We further selected only those predictions with a threshold score of at least 13 for all three pairs, narrowing down the predictions to a list of 258 candidate miRNA genes (Table S8). These included 114 known human miRNA genes and 144 candidate novel miRNA genes.

Experimental verification of predicted miRNA genes

We selected 12 of these 144 predicted miRNA genes for experimental validation. These were selected at a range of MiRScan scores, and a range of folding free energy, such that they are representative of the set of all 258 predicted miRNAs (Table S3).

We used a method of PCR amplification followed by sequencing verification (see Lau et al[16]) on a pool of adaptor-ligated 18-26mer RNAs to verify the expression of the predicted miRNAs. This experimental procedure is carefully designed to ensure that there is no contamination with genome DNA:

1. Its includes three steps of rigorous PAGE purification of small RNA fractions in the process of microRNA cloning/PCR verification. Large RNA and genomic DNA are purified away.
2. PCR is done with one primer complementary to the artificial adaptors used in ligating the microRNAs, and a gene specific primer. Genomic DNA thus would not contain the ligation-specific adaptor sequence, and hence would not be amplified.
3. Finally, the sequencing of the resulting clones shows that the product has the precise expected sequence and precise expected junction. This would not occur with genomic DNA.

Small RNAs (18 to 26-mer) from 10 human tissues (breast, pancreas, prostate, colon, stomach, uterus, lung, brain, liver and kidney) were purified through a 15% denaturing polyacrylamide gel. Purified small RNAs were subjected to two steps of adaptor ligations to both the 5' and the 3' ends of miRNAs, with denaturing PAGE purifications after each ligation step, as described by Miska et al[17]. The sequences for the adaptors were artificially designed (5' adaptor: acggaattcctcactAAA; 3' adaptor: pUUUaaccgcgaattccagidT, where p: phosphate; upper-case: RNA base; lower case: DNA base; idT: inverted dT). Ligation products were reverse-transcribed using a primer specific to the 3' adaptor sequence. These cDNAs were pooled and diluted 1000-fold as the substrate for the subsequent PCR reactions. The substrate was amplified in PCR reactions with a common 5' primer (0.1 μM, 5'-CAACGGAATTCCTCACTAAA-3'), corresponding to the 5' adaptor sequence, and miRNA-specific 3' primers (1 μM), for 25 cycles at 50 °C of annealing temperature. The miRNA-specific 3' primers were designed to match 3' end of the predicted miRNAs, but allowing 6-7 bases in the 5' end for sequencing verification. The products of the first-round PCR reactions were diluted 20-fold and amplified for a second-round of 25 cycles with the same reaction conditions. The products of the second-round PCR were cloned into the TOPO PCR4 vector (Invitrogen), following the manufacturer's protocol. The inserts of the clones were PCR-amplified with M13-forward and M13-reverse primers and sequenced both directions with M13-forward and M13-reverse primers to verify the 5'-end of predicted miRNAs. A predicted miRNA was verified only if the sequenced 5' end had the same length and exactly matching sequence as the predicted miRNAs. The list of used primers is shown in Table S6.

Supplementary References

1. Schwartz, S. et al. Human-mouse alignments with BLASTZ. *Genome Res* **13**, 103-7 (2003).
2. Blanchette, M. et al. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* **14**, 708-15 (2004).
3. Lindblad-Toh, K. & al, e. Initial sequencing and analysis of the dog genome (In preparation). (2005).
4. Maglott, D. R., Katz, K. S., Sicotte, H. & Pruitt, K. D. NCBI's LocusLink and RefSeq. *Nucleic Acids Res* **28**, 126-8 (2000).
5. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence project: update and current status. *Nucleic Acids Res* **31**, 34-7 (2003).
6. Waterston, R. H. et al. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520-62 (2002).
7. Thompson, J. D., Higgins, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673-80 (1994).
8. Stormo, G. D. DNA binding sites: representation and discovery. *Bioinformatics* **16**, 16-23 (2000).
9. Pietrokovski, S. Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucleic Acids Res* **24**, 3836-45 (1996).
10. Hughes, J. D., Estep, P. W., Tavazoie, S. & Church, G. M. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol* **296**, 1205-14 (2000).
11. Kellis, M., Patterson, N., Birren, B., Berger, B. & Lander, E. S. Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. *J Comput Biol* **11**, 319-55 (2004).
12. Su, A. I. et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**, 6062-7 (2004).
13. Hollander, M. & Wolfe, D. A. *Nonparametric statistical methods* (J. Wiley, New York, 1999).
14. Fontana, W. et al. RNA folding and combinatory landscapes. *Physical Review. E. Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* **47**, 2083-2099 (1993).
15. Lim, L. P., Glasner, M. E., Yekta, S., Burge, C. B. & Bartel, D. P. Vertebrate microRNA genes. *Science* **299**, 1540 (2003).
16. Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* **294**, 858-62 (2001).
17. Miska, E. A. et al. Microarray analysis of microRNA expression in the developing mammalian brain. *Genome Biol* **5**, R68 (2004).

Figure S1. Distribution of transcription starting sites (TSS) differences between 7878 orthologous human/mouse gene pairs. TSS difference between a gene pair was the distance in the mouse genome between the position aligned to human TSS and the annotated mouse TSS. The annotations of TSS for both human and mouse were based RefSeq[4,5].

Figure S2. Tissue specificity of expression for genes containing discovered motifs. For each discovered motif, three gene sets are generated: S_1 contains all genes with conserved occurrences of the motif and two equal-sized control sets S_2 and S_3. S_2 is a control for the specific motif, containing a random subset of genes in which the motif occurs in the human genome but is not conserved. S_3 is a general control, containing a random set of genes randomly drawn from the union of the sets S_1 for all motifs. Tissue-specific enrichment of gene sets was tested using a database of 75 RNA expression in human tissues[12]. For each set, an enrichment score was calculated for each tissue. Shown here are enrichment scores of 175 discovered motifs, represented in pseudo color, for conserved motif gene set (**a**) and non-conserved motif gene set (**b**). Similar to S_2, the control set S_3 also showed little enrichment in the same tissues.

Table S1. List of 123 promoters motifs in the TRANSFAC database, ranked by MCS, and related discovered motifs. Matching bases shown in bold. Known motif: consensus of the TRANSFAC motif. Discovered motif: consensus of the discovered motifs from the aligned promoter database. MCS: Motif conservation score.

Table S2. List of 174 discovered promoter motifs, ranked by MCS. MCS: Motif conservation score. Known factor: name of best matching motif in TRANSFAC database, if any. Maximum Tissue Enrichment Score (see legend to Figure S2). Position bias: Mode of position for highly clustered motifs, shown for cases with positional clustering score above 5 standard deviations. Weak matches to known motifs are indicated by '*'.

Table S3. List of 174 discovered promoter motifs and motif variants grouped in the same clusters. Conserved num: Number of conserved instances. Total num: Number of total instances. MCS: motif conservation score. Known factor: name of the best matching motif in TRANSFAC database, if any.

Table S4. List of 106 motifs discovered in 3'-UTR regions and motif variants grouped in the same cluster. Conserved num: Number of conserved instances. Total num: Number of total instances. MCS: motif conservation score.

Table S5. List of 72 known 8-mer motifs discovered in 3'-UTRs. Motifs in each cluster share at least six consecutive nucleotides as the most conserved 8-mer in the cluster, which is chosen as a representative of the cluster. Matched miRNA: known and predicted miRNAs that match to the conserved 8-mers (the predicted miRNAs start with 'MIR' followed by a number without dash.).

Table S6. List of 90 known miRNA sequences that can form W-C pairing to the conserved 8-mer motifs discovered in 3'-UTR. Matched sequences in miRNAs are highlighted in lower cases. C: Number of the conserved instances of the motif. N: Number of the total number of instances of the motif. Pc: Conservation rate of the motif. The table also included an additional list of 27 miRNAs that can pair to the conserved 8-mers when one mismatch was allowed. The list of one-base mismatched miRNAs was grouped into three categories, including 4 miRNAs containing T-G pairing, 10 miRNAs with mismatched first 5' nucleotide to letter 'A' of the conserved 8-mer motifs, and other mismatches.

Table S7. List of 60 3'-UTR motifs not related to miRNA regulations. The motifs are ranked by MCS. MCS: motif conservation score. Total Num: the total number of sites found in 3'-UTR. Conserved Num: the number of conserved sites found in 3'-UTR. Pc: conservation rate.

Table S8. List of 258 predicted miRNA genes. Please refer to online link:
http://www.broad.mit.edu/seq/HumanMotifs/

Table S9. List of 13 tested miRNA genes. Please refer to online link:
http://www.broad.mit.edu/seq/HumanMotifs/

Table S10. List of 3' primers used for PT-PCR amplification. 12 predicted novel miRNAs were tested. The ~22 bp sequences downstream of the seed 8-mers were predicted as mature products, and were used to design 3' primers.

Table S11. List of 11 predicted miRNAs that show high sequence similarity to known miRNAs.

Dataset D1: Aligned promoter database.
http://www.broad.mit.edu/seq/HumanMotifs/

Dataset D2: Aligned 3'-UTR database.
http://www.broad.mit.edu/seq/HumanMotifs/

Dataset D3: Genomic locations of conserved motifs.
http://www.broad.mit.edu/seq/HumanMotifs/

Dataset D4: Sites of TRANSFAC motifs annotated in aligned promoter database.
http://www.broad.mit.edu/seq/HumanMotifs/

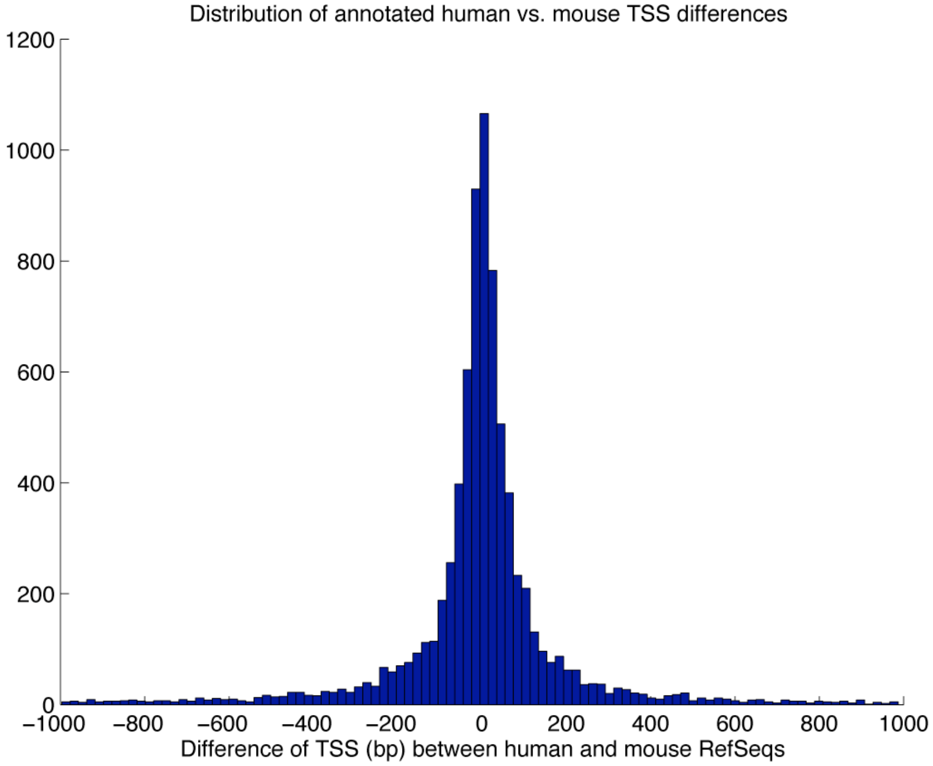
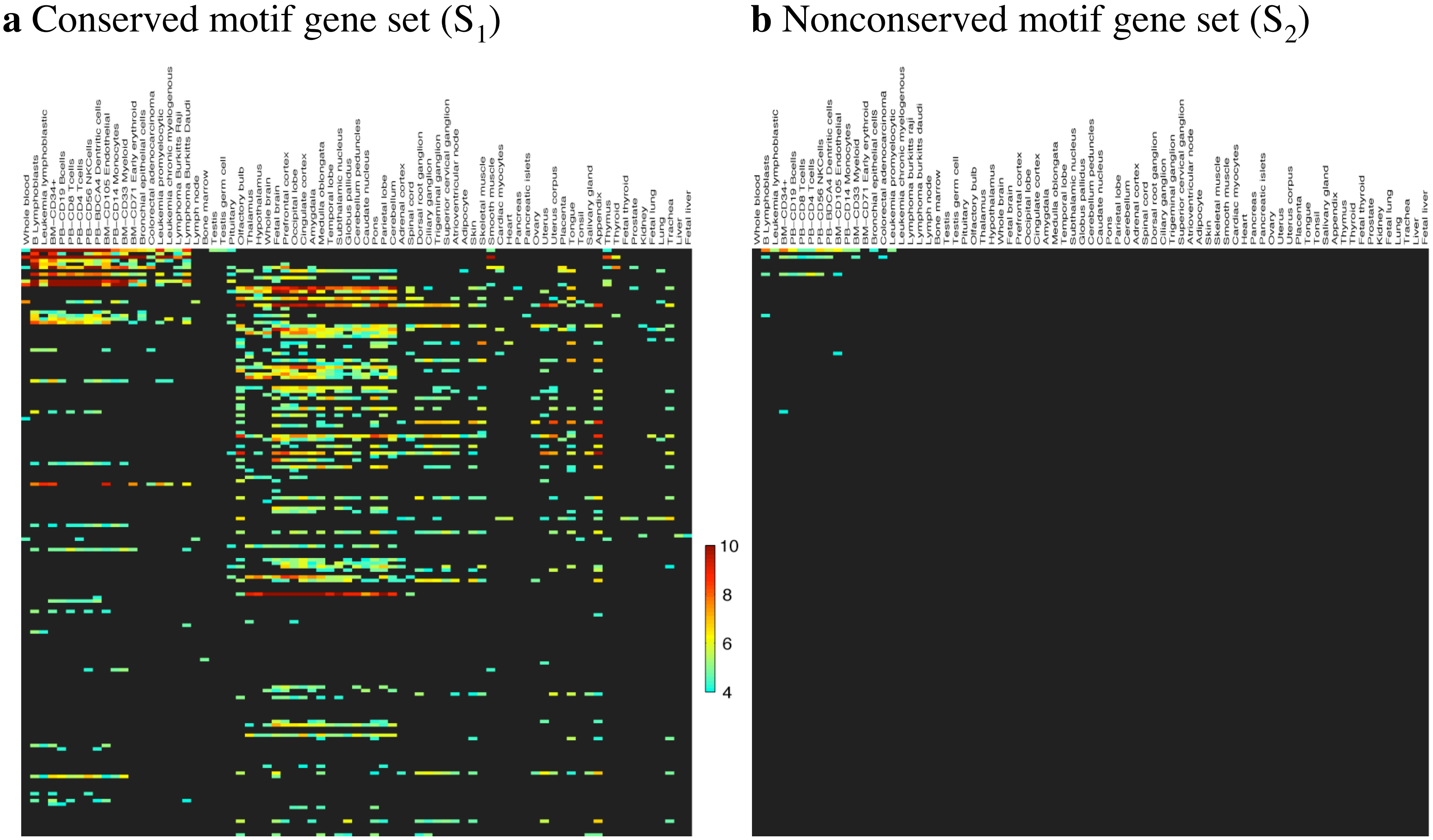

Distribution of annotated human vs. mouse TSS differences
1200
1000
800
600
400
200
0
−1000
−800
−600
−400
−200
0
200
400
600
800
1000
Difference of TSS (bp) between human and mouse RefSeqs

Fig. S1



a Conserved motif gene set (S_1)

b Nonconserved motif gene set (S_2)

Fig. S2

Supplementary Table S1 **Known promoter regulatory elements and related discovered motifs**

Factor	Known motif	MCS	Discovered motif
SP-1	G**GGGCGGG**GC	46.8	**GGGCGGR**
YY1	**GCCATnTT**	34.7	**GCCATnTT**G
MYC	S**CACGTG**	32.7	**CACGTG**
NF-Y	Y**SATTGGY**Y	31.2	**GATTGGY**
AP-1	C**TGASTCA**	30.8	**TGAnTCA**
MAZ	G**GGGAGGG**	29.7	**GGGAGGR**R
CREB	**TGACGTMA**	29.5	**TGACGTMR**
NF-MUE1	**CGGCCATCT**	26.0	**CGGCCATYK**
MYOD	Rn**CAGGTG**	24.7	**CAGGTG**
ELK-1	**CCGGAART**	22.6	**CCGGAARY**
NRF-1	Y**GCGCATGCG**	20.9	**RCGCAnGCG**Y
TEL-2	**CAGGAAGT**AR	20.8	**SMGGAAGT**
GABP	v**CCGGAAGn**GCR	19.8	**SCGGAAGY**
STAT1	**CAnTTCCS**	17.9	**CATTTCCK**
CAC-BP	GR**GGSTGG**G	15.0	**GGGTGG**
AP-4	G**CAGCTG**nY	14.9	**CAGCTG**
SRY	K**TWGTTT**	14.6	**TTGTTT**
TBP	**TATAAA**TW	14.2	**TATAAA**
FOXO1	**RWAAACA**A	14.1	**RTAAACA**
TFII-I	R**GAGGKAGG**	13.9	**GnGGGAGG**
PEA3	**MGGAWGT**	13.6	S**MGGAAGT**
SF-1	**TGRCCTTG**	12.6	**TGACCTTG**
SOX-5	**ATTGTT**	12.5	YY**ATTGTT**
SREBP-1	A**TCACGTG**AY	12.4	**TCACGTG**
OCTAMER	**ATGCAAAT**nA	12.2	Y**ATGYAAAT**
P65	**GGGRATTTCC**	11.9	**GGGnnTTTCC**
ATF6	**TGACGTGG**	11.7	**TGACGTGK**
E4BP4	R**TTACRTAA**Y	11.0	**TTAYRTAA**
SRF	Gn**CCAWATAWGG**M	10.7	**CCAWWnAAGG**
MEF-2	Y**TAAAWATAG**CY	10.7	**TAAWWATAG**
POU3F2	**ATTARCAT**AA	10.5	**ATTARCAT**
ELF-1	RnW**MbAGGAART**	9.5	**RGAGGAARY**
BACH2	SR**TGAGTCA**nC	9.3	**TGAGTCA**
MEIS1	**TGACAG**	9.2	**TGACAG**
E2F	**GCGCSAAA**	9.2	S**GCGSSAAA**
ETS	AnnC**ACTTCCTG**	9.0	**RYTTCCTG**
RFX1	GTTRCYWnGYnAC	8.7	-
RORALPHA2	**TGACCTAnW**TW	8.6	**TGACCTAnW**
PU.1	W**GAGGAAG**	8.4	**GAGGAAG**Y
EGR	**GTGGGSGCRR**S	8.4	**GTGGGCGnR**
NRSF	TTCAGCACCACGGACAGMGCC	8.3	-
NF-E2	R**TGACTCAGC**A	7.8	**TGASTMAGC**
LEF1	**CTTTGA**	7.6	**CTTTGA**
HNF-3	**TRTTTRY**TYW	7.6	**TGTTTGY**
ALPHA-CP1	CAG**CCAATGAG**	7.1	**YYAATGAG**
STATX	**TTMCGGGAA**	6.8	**TTCYnRGAA**
RP58	TC**CAGATGTT**	6.4	**CRGATGTT**
TEF-1	**GRRATG**	6.3	W**GGAATG**Y
NKX2-5	C**WTAATTG**	6.1	**TTAATTG**
CHX10	G**CTAATTA**	5.9	**CTAATTW**
TCF-4	**WTCAAAG**S	5.8	**WTCAAAG**
STAT5A	**AWTTCY**	5.7	M**ATTTCC**
IY	**AWTTTCC**	5.5	**MATTTCC**
NKX6-2	WAd**TAAWTA**	5.4	**TAATTA**
ATF-1	**TGACGTCA**RRG	5.3	**TGACGTCA**
POU1F1	**ATGAATAA**WT	5.2	**ATGAATRR**
PBX-1	W**TGATTGnT**	5.0	**TGATTGRY**
LHX3	**TTAATTAA**TT	5.0	**YTAATTAA**
ICSBP	CA**GTTTCAYTTY**	5.0	**STTTCRnTTT**
FOX	W**AA**A**YAAACA**ATM	5.0	**AA**G**YAAACA**
HOXA4	**CYAATTW**T	4.8	**CTAATTW**
CDP	RnT**AATCGATn**W	4.8	**RATCRATA**
IRF	bnCR**STTTCAnTTY**Y	4.6	**STTTCRnTTT**
GATA	**WGATAR**	4.6	**WGATAA**GR
MYB	GnCnGTT	4.4	-
MIF-1	**GTTGCWWGGYAAC**nGS	4.3	**RYTGCnnRGnAAC**
HSF2	**GAAnnWTC**K	4.0	R**GAAnnTTC**
HNF-1	**GGTTAATnWTT**AMC	4.0	**RGTTAMWnATT**
AREB6	W**CAGGTG**WnW	3.8	**CAGGTG**
C-REL	S**GGRnTTTCC**	3.6	**GGGnnTTTCC**
TAL-1ALPHA/E47	A**ACAGATGK**T	3.4	**MCAGATGK**
POU6F1	GCA**TAAWTTAT**	3.4	**TAATTTAT**K
FREAC-4	CTW**AWGTAAACA**nWG	3.4	**RRGTAAACA**
BRN-2	YKnATTWYSnATG	3.4	-
AFP1	GTGYARTTAAT	3.3	-
TCF-1(P)	GKCRGKTT	3.2	-
HNF-4	**TGAMCTTT**GMMCYT	3.1	**TGAMCTTT**
STAT	**TCCMAGAA**	3.0	**TCCCRGAA**R
IRF1	**AAGTGAA**	2.8	**AAGTGAA**
E4F1	**GTGACGT**ARS	2.7	**GTGACGY**
NF-AT	**WGGAAA**nW	2.6	**TGGAAA**
CDC5	GATTTAACATAA	2.6	-
AML1	**ACCACA**	2.6	R**ACCACA**R
IPF1	KG**TCATTAnn**dC	2.5	**TCATTAnY**
FAC1	TnYG**TGTTKTG**	2.5	**TGTTGTK**
AHR	TnGCGTG	2.5	-
C/EBPBETA	Kn**TTGCnYAAY**	2.3	**TTGCWCAAY**
AP-2ALPHA	SCYnnnGGC	2.3	-
ER	Rnnn**TGACCT**	2.1	**TGACCT**
CR1	SC**GATCGAT**	1.9	**RATCRAT**A
DBP	**GTdTGCT**	1.8	Y**GTTTGCT**Y
HSF	**TTCCMGARGYTTC**	1.7	**TTCnRGnnnnTTC**
TEF	TTATRTWAACAT	1.5	-
PAX-4	AA**WAATTA**nS	1.4	**TAATTA**
HNF-6	hWAAATCAATAW	1.4	-
CDX-2	GGYMATAAAAnTnT	1.4	-
NKX2-5	TYAAGTG	1.3	-
TCF11	Wnn**ATGAC**	1.2	**ATGAC**A
HIF-1	CCGCACGTMnnC	1.1	-
OLF-1	CMnvYT**CYCTRGGGAv**ThG	0.9	**CCCnnGGGAR**
NKX6-1	TWT**TTAATTG**GTT	0.9	**TTAATTG**
AREB6	AbW**CAGGTR**nR	0.9	**CAGGTA**
IRF-7	An**TTTCGnWTT**CSnA	0.8	S**TTTCRnTTT**
IRF-1	G**GTTTCRCTTT**TS	0.8	**STTTCRnTTT**
HP1	CTGTTGAAWATT	0.8	-
FREAC-3	TGTT**TATTTAC**	0.8	**TRTTTAC**T
RREB-1	GGGGKKGTTTGGGG	0.6	-
CHOP-C/EBPALPHA	RTGCAATMCCC	0.5	-
CRX	KG**RGATTA**nnnR	0.3	**GGATTA**
AMEF-2	KKRG**TTATTTTTAR**hCMG	0.3	**YTATTTWTAA**
PITX2	YTG**GGATTA**nW	0.2	**GGATTA**
NCX	**GTAAKT**nG	0.1	**GTAATT**
PTF1-BETA	SCTGWvvKTTTCYC	0.0	-
T3R	Mn**TGWCCT**	-0.4	**TGACCT**Y
SMAD-3	TGTCTGTCT	-0.4	-
NF-1	**TGGnnnnnnGCCAA**	-0.5	**TGGnnnnnnKCCAR**
CP2/LBP-1C/LSF	GCTGGnTnGnnCYnG	-0.5	-
PAX	GTKAGTTCCAG	-0.7	-
ZID	GGCTCYATCAYC	-0.8	-
MTF-1	TbTGCAChCGGCCC	-0.8	-
IK-1	GGYA**TTCCCAnd**	-1.2	T**TTCCCAnR**
LYF-1	YCTCCCAAA	-2.5	-
P300	GGGAGTnnnnS	-4.8	-

Supplementary Table S2 **Discovered motifs in human promoters**

No.	Discovered motif	MCS	Known factor	Conservation in promoters	Conservation in introns	Maximum Tissue enrichment score	Position bias
1	RCGCAnGCGY	107.8	NRF-1	0.49	0.09	15.0	-62
2	CACGTG	85.3	MYC	0.47	0.01	8.8	-62
3	SCGGAAGY	80.4	ELK-1	0.44	0.02	22.4	-24
4	ACTAYRnnnCCCR	69.5	-	0.61	0.06	8.1	-89
5	GATTGGY	64.6	NF-Y	0.51	0.04	9.8	-63
6	GGGCGGR	63.9	SP1	0.21	0.02	11.4	-63
7	TGAnTCA	62.8	AP-1	0.38	0.08	6.5	-
8	TMTCGCGAnR	55.7	-	0.64	0.08	9.4	-62
9	TGAYRTCA	55.7	ATF3	0.50	0.07	6.1	-66
10	GCCATnTTG	54.7	YY1	0.72	0.03	12.2	-
11	MGGAAGTG	51.6	GABP	0.43	0.02	13.9	-23
12	CAGGTG	47.6	E12	0.26	0.06	9.9	-
13	CTTTGT	46.0	LEF1	0.42	0.05	13.6	-
14	TGACGTCA	44.8	ATF3	0.44	0.07	4.2	-22
15	CAGCTG	43.9	AP-4	0.27	0.08	8.9	-
16	RYTTCCTG	43.0	C-ETS-2	0.32	0.06	7.4	-24
17	AACTTT	42.1	IRF1(*)	0.43	0.04	11.1	-
18	TCAnnTGAY	40.4	SREBP-1	0.47	0.04	4.9	-64
19	GKCGCnnnnnnnTGAYG	40.1	-	0.35	0.00	5.6	-62
20	GTGACGY	38.4	E4F1	0.34	0.02	6.6	-56
21	GGAAnCGGAAnY	37.7	-	0.68	0.00	7.0	-33
22	TGCGCAnK	37.4	-	0.24	0.02	8.2	-17
23	TAATTA	37.3	CHX10	0.29	0.13	7.1	-
24	GGGAGGRR	33.5	MAZ	0.16	0.03	9.4	-
25	TGACCTY	33.4	ESRRA	0.30	0.07	7.7	-
26	TTAYRTAA	32.6	E4BP4	0.34	0.05	6.1	-
27	TGGnnnnnnKCCAR	32.3	-	0.27	0.07	4.5	-
28	CTAWWWATA	32.3	RSRFC4	0.36	0.05	7.6	-
29	CTTTAAR	30.8	-	0.43	0.05	5.4	-
30	YGCGYRCGC	30.5	-	0.19	0.00	5.2	-31
31	GGGYGTGnY	30.0	-	0.24	0.04	5.4	-63
32	TGASTMAGC	27.2	NF-E2	0.39	0.07	5.4	-66
33	YTATTTTnR	26.4	MEF-2	0.21	0.05	7.1	-
34	CYTAGCAAY	26.1	-	0.50	0.06	5.2	-142
35	GCAnCTGnY	25.7	MYOD	0.25	0.06	8.2	-
36	RTAAACA	25.6	FREAC-2	0.46	0.07	7.0	-
37	GTTRYCATRR	25.3	-	0.54	0.11	7.6	-56
38	TGACCTTG	25.2	ERRALPHA	0.37	0.06	8.1	-
39	TCCCRnnRTGC	24.3	-	0.30	0.03	6.8	-60
40	TTCYnRGAA	24.3	STAT5A	0.19	0.05	-	-
41	TGACAGnY	24.1	MEIS1	0.27	0.07	6.9	-
42	TGACATY	23.8	-	0.23	0.06	5.8	-
43	GTTGnYnnRGnAAC	23.7	-	0.47	0.13	4.7	-57
44	YATGnWAAT	23.5	OCT-X	0.53	0.06	6.9	-
45	CCAnnAGRKGGC	23.4	-	0.47	0.20	-	-101
46	WTTGKCTG	23.0	-	0.25	0.04	5.0	-63
47	TGCCAAR	22.9	NF-1	0.25	0.08	7.0	-
48	GCGnnAnTTCC	22.8	C-REL(*)	0.30	0.00	6.0	-12
49	CATTGTYY	22.5	SOX-9	0.43	0.04	5.8	-
50	RGAGGAARY	22.4	PU.1	0.22	0.04	4.0	-
51	TATAAA	22.1	TATA	0.47	0.05	8.6	-23
52	YYCATTCAWW	21.6	POU1F1(*)	0.61	0.03	5.8	-
53	RYTGCnnRGnAAC	21.3	MIF-1	0.33	0.13	-	-
54	TAAWWATAG	21.1	RSRFC4	0.31	0.05	4.5	-
55	TGGAAA	21.1	NF-AT	0.18	0.05	8.8	-
56	GGGTGGRR	20.9	PAX-4	0.20	0.03	7.5	-
57	ACCTGTTG	20.7	-	0.38	0.03	4.1	-
58	YCATTAA	20.3	IPF1(*)	0.24	0.08	6.2	-
59	WCTCnATGGY	19.9	-	0.41	0.02	-	-66
60	TTGTTT	19.8	FOXO4	0.27	0.06	9.6	-
61	YTAATTAA	19.8	LHX3	0.28	0.13	4.1	-
62	SMTTTTGT	19.1	-	0.37	0.03	8.0	-
63	AAGWWRnYGGC	19.1	-	0.38	0.02	5.4	-
64	TTAnTCA	18.8	AP-1(*)	0.20	0.06	7.0	-
65	ARGGGTTAA	18.7	FXR(*)	0.41	0.10	4.1	-104
66	RACTnnRTTTnC	18.5	-	0.36	0.03	-	-67
67	TGAnnYRGCA	17.5	TCF11/MAFG	0.24	0.04	5.3	-
68	RGAAnnTTC	17.4	HSF1	0.18	0.04	5.6	-
69	SGCGSSAAA	17.3	E2F-1/DP-2	0.24	0.01	9.1	-21
70	CGTSACG	17.2	PAX-3	0.18	0.04	-	-25
71	SYATTGTG	17.1	-	0.40	0.03	4.2	-
72	TTCYRGAA	17.1	-	0.20	0.05	-	-
73	CTTTGA	17.0	LEF1	0.19	0.07	6.4	-
74	GGAMTnnnnnTCCY	16.7	-	0.21	0.01	4.0	-104
75	TnCATnTCCYR	16.5	STAT1(*)	0.35	0.03	-	-62
76	CAGGTA	16.3	AREB6	0.22	0.05	6.3	-
77	AAAYRnCTG	16.3	-	0.18	0.04	5.2	-
78	GCTnWTTGK	16.2	-	0.24	0.03	-	-104
79	WGGAATGY	16.1	TEF-1	0.21	0.05	6.5	-
80	SnACAnnnYSYAGA	15.8	-	0.31	0.02	-	-68
81	CGGAARnGGCnG	15.7	-	0.24	0.07	5.3	-25
82	CTGYnnCTYTAA	15.5	-	0.41	0.04	-	-120
83	TGTTTGY	15.1	HNF-3	0.19	0.05	6.6	-
84	RGTTAMWnATT	15.0	HNF-1	0.31	0.03	5.3	-
85	STTTCRnTTT	14.9	IRF	0.24	0.03	4.7	-
86	GGGnnTTTCC	14.9	NF-KAPPAB	0.21	0.02	-	-

Supplementary Table S2 **Discovered motifs in human promoters**

87	RYTGCnWTGGnR	14.6	-	0.26	0.06	5.6	-
88	GGCnKCCATnK	14.3	-	0.30	0.03	5.9	-
89	GTTnYYnnGGTnA	14.3	-	0.26	0.06	-	-
90	YAATnRnnnYnATT	14.3	CART-1(*)	0.22	0.05	-	-
91	GTGGGTGK	14.1	-	0.20	0.03	5.9	-
92	TGCTGAY	14.0	-	0.21	0.05	5.9	-
93	GGATTA	14.0	PITX2	0.22	0.05	6.7	-
94	TGATTTRY	13.9	GFI-1	0.19	0.08	5.6	-
95	GCCnnnWTAAR	13.7	-	0.29	0.04	-	-69
96	YGCAnTGCR	13.7	-	0.18	0.02	8.5	-
97	YATTnATC	13.7	CDP(*)	0.19	0.05	6.5	-
98	GTCnYYATGR	13.6	-	0.31	0.03	-	-
99	ATCMnTCCGY	13.3	-	0.42	0.01	-	-275
100	CRGAARnnnnCGA	13.3	-	0.23	0.00	-	-
101	CTGCAGY	13.2	-	0.18	0.03	11.6	-
102	ATGGYGGA	13.2	-	0.29	0.02	4.1	-
103	ACAWnRnSRCGG	13.1	-	0.29	0.00	5.0	-
104	CCAATnnSnnnGCG	13.0	-	0.23	0.00	-	-87
105	ACTWSnACTnY	13.0	-	0.25	0.01	-	-66
106	CCGnMnnTnACG	12.9	-	0.19	0.00	5.1	-48
107	RTTTnnnYTGGM	12.8	-	0.18	0.06	4.3	-
108	AACWWCAAnK	12.7	FAC1(*)	0.35	0.03	-	-105
109	YGTCCTTGR	12.7	-	0.26	0.04	4.7	-
110	MCAATnnnnnGCG	12.5	-	0.21	0.00	4.6	-62
111	RACCACAR	12.3	AML	0.21	0.04	-	-
112	KTGGYRSGAA	12.3	-	0.26	0.02	5.1	-
113	AACYnnnnTTCCS	12.3	-	0.24	0.01	-	-53
114	YTCCCRnnAGGY	12.2	-	0.17	0.03	-	-63
115	YRTCAnnRCGC	12.2	-	0.20	0.02	-	-36
116	KMCATnnWGGA	12.2	-	0.33	0.02	-	-
117	TGTYnnnnnRGCARM	12.1	-	0.19	0.03	-	-
118	GGCnRnWCTTYS	12.0	-	0.17	0.02	-	-21
119	GGGnRMnnYCAT	11.9	-	0.19	0.02	-	-
120	KRCTCnnnnMAnAGC	11.8	-	0.28	0.01	4.7	-
121	CCAWWnAAGG	11.7	SRF	0.25	0.03	4.4	-
122	RnTCAnnRnnYnATTW	11.7	-	0.21	0.04	-	-
123	GGCnnMSMYnTTG	11.6	-	0.21	0.01	5.1	-30
124	CCAWYnnGAAR	11.5	-	0.22	0.04	-	-103
125	RAAGnYnnCTTY	11.5	-	0.17	0.03	-	-
126	WYAAAnnRnnnGCG	11.4	-	0.26	0.02	-	-
127	WWTAAGGC	11.3	-	0.26	0.02	-	-
128	RYCACnnRnnRnCAG	11.3	-	0.23	0.06	-	-
129	RRAGTTGT	11.2	-	0.20	0.02	5.6	-
130	CCCnnGGGAR	11.2	OLF-1	0.18	0.03	5.2	-
131	GATAAGR	11.2	GATA-X	0.18	0.04	5.8	-
132	TCCATTKW	11.1	-	0.27	0.02	4.9	-
133	RYTAAWnnnTGAY	11.1	-	0.24	0.03	-	-
134	CATRRAGC	11.1	-	0.26	0.03	-	-
135	AGCYRWTTC	11.1	-	0.19	0.04	-	-
136	TAAYnRnnTCC	11.0	-	0.21	0.05	-	-
137	GAAnYnYGACnY	11.0	-	0.22	0.02	-	-
138	MYAATnnnnnnnGGC	11.0	-	0.19	0.03	-	-66
139	AAAYWAACM	11.0	HFH-4	0.34	0.05	5.6	-
140	RnGTGGGC	10.9	-	0.19	0.03	7.1	-
141	TTCnRGnnnnTTC	10.9	HSF	0.19	0.03	-	-
142	ACAWYAAAG	10.9	-	0.27	0.03	-	-
143	CAGnWMCnnnGAC	10.8	-	0.24	0.02	6.7	-
144	AAAnWWTGC	10.8	-	0.28	0.03	5.4	-
145	YKACATTT	10.7	-	0.32	0.04	-	-
146	RRCCGTTA	10.5	-	0.30	0.02	5.1	-
147	YAATnAnRnnnCAG	10.5	-	0.24	0.04	-	-
148	GATGKMRGCG	10.5	-	0.27	0.07	4.2	-
149	YGACnnYACAR	10.4	-	0.26	0.02	-	-68
150	YTTCCnnnGGAMR	10.4	-	0.22	0.04	-	-
151	RYAAAKnnnnnnTTGW	10.4	-	0.17	0.03	-	-
152	WCAAnnnYCAG	10.3	-	0.22	0.02	-	-
153	CTGRYYYnATT	10.3	-	0.21	0.03	4.3	-
154	RnCTGnYnRnCTGnY	10.2	-	0.20	0.03	-	-
155	WGTTnnnnnAAA	10.2	-	0.21	0.03	6.8	-
156	YRCCAKnnGnCGC	10.2	-	0.19	0.10	-	-65
157	KCCGnSWTTT	10.2	-	0.22	0.02	6.8	-
158	CCCnnnnnnAAGWT	10.2	-	0.22	0.02	-	-
159	GGCKCATGS	9.9	-	0.18	0.01	-	-21
160	CAGnYGKnAAA	9.9	-	0.20	0.03	-	-
161	TTAnWnAnTGGM	9.8	-	0.18	0.03	-	-
162	TAAnnYSGCG	9.8	-	0.21	0.04	4.3	-
163	GGARnTKYCCA	9.8	-	0.23	0.03	-	-
164	GCGSCMnTTT	9.8	-	0.22	0.01	5.2	-18
165	CCAWnWWnnnGGC	9.8	-	0.21	0.01	4.2	-
166	YnTTTnnnAnGCARM	9.6	-	0.22	0.03	5.0	-
167	CCTnTMAGA	9.6	-	0.21	0.02	-	-
168	YTAAYnGCT	9.5	-	0.18	0.06	-	-
169	TTTnnAnAGCYR	9.5	-	0.18	0.04	-	-
170	YnGTTnnnATT	9.1	-	0.20	0.04	6.6	-
171	CTCnAnGTGnY	9.1	-	0.23	0.02	-	-
172	TTGCWCAAY	9.0	C/EBPBETA	0.23	0.01	-	-
173	YWATTWnnRGCT	8.8	-	0.18	0.04	-	-
174	WTGAAAT	8.1	-	0.22	0.05	5.1	-

Supplementary Table S3 **Motifs discovered in promoters in clusters**

Motif	Conserved num	Total num	Conservation rate	MCS	Known factor
>cluster_1					
RCGCANGCGY	1013	2117	0.48	107.8	V$NRF1_Q6
RNGCATGCNY	506	1304	0.39	53.3	-
CGCNTGYGCANT	73	263	0.28	21.8	V$NRF1_Q6
RCGCANNCKCAG	86	464	0.19	21.4	-
GNGCANGYNCAGY	60	311	0.19	17.8	-
SAGCATGY	99	409	0.24	15	-
RNGCANSNKMAGT	54	286	0.19	14.6	-
CATGNNCAGY	70	266	0.26	14.3	-
GCGCANRCTC	59	349	0.17	11.9	-
MGCATGTR	52	212	0.25	11.4	-
>cluster_2					
CACGTG	1588	3378	0.47	85.3	V$MYC_Q2
TCACGTG	459	954	0.48	47.8	V$USF_Q6_01
GCCACGTS	174	634	0.27	25.7	-
GCCACGYS	340	1972	0.17	23.5	V$MYCMAX_B
AGCASGTG	158	586	0.27	17.6	V$USF_C
RCGCAYGTG	58	241	0.24	13.8	-
RYTTCMNGTG	65	321	0.20	13	-
>cluster_3					
SCGGAAGY	1023	2566	0.40	80.4	V$ELK1_02
CCCGGAWR	451	1635	0.28	40.5	-
SWTCCGGGTC	50	133	0.38	25.3	-
GACMYGGAAR	54	167	0.32	21.2	-
CCGGAARY	171	482	0.35	20.9	V$ELK1_02
ACATMCGG	53	169	0.31	16.5	-
RGTTCCGG	145	744	0.19	15.5	V$ELK1_02
AAGTYCCGS	63	348	0.18	13.4	-
AACWTCCG	62	287	0.22	13	V$PEA3_Q6
YTCCRKMTGT	68	303	0.22	11.5	-
CRGATGTT	93	454	0.20	11.1	V$RP58_01
YYGGTTCCG	57	316	0.18	10.9	-
>cluster_4					
ACTAYRNNNCCCR	317	520	0.61	69.5	-
ACTACNNNNCCC	384	646	0.59	65	-
ACTACNNNTCCCR	91	131	0.69	38.2	-
RRACTACA	240	581	0.41	33.3	-
RRACTNCATNT	58	190	0.31	15.7	-
GACKNCATY	95	415	0.23	14.8	-
RAACYRCNNNNCCC	54	193	0.28	14.1	-
GGAYTAC	98	560	0.18	9.1	-
>cluster_5					
GATTGGY	857	1740	0.49	64.6	V$NFY_Q6_01
RGCCAATNR	593	1527	0.39	50.5	V$NFY_01
RNCCAATGR	385	1005	0.38	42.2	V$NFY_01
TGATTGRY	268	822	0.33	23.3	V$PBX1_02
YNATTGGT	184	845	0.22	16.3	V$NFY_Q6_01
TCCAATNA	58	247	0.23	12	-
RCCCAATNR	77	371	0.21	11.6	-
CCAATAR	113	546	0.21	11.6	V$CDP_01
YAATTGGNY	124	611	0.20	10.5	-
YGACYAAT	89	410	0.22	9.3	-
>cluster_6					
GGGCGGR	3067	15905	0.19	63.9	V$SP1_Q6
GGGCGG	4303	24152	0.18	55.9	V$SP1_Q6
RGGCGKGGC	810	4190	0.19	47.9	V$SP1_Q6
GGGNGGG	2711	11113	0.24	40.9	V$SP1_Q4_01
RGGCGGAGY	250	1601	0.16	19.5	-
RGGCGGGNY	286	1131	0.25	18.8	V$SP1_Q6
YAGGKGGCGC	76	308	0.25	18.7	-
RGGTGKGGC	190	843	0.23	18.4	-
GCCMCTCCY	193	993	0.19	14.6	-
AGGNGKCGCTS	56	298	0.19	13.6	-
GGGGGCG	224	1261	0.18	8.3	V$EGR_Q6
>cluster_7					
TGANTCA	1924	5048	0.38	62.8	V$AP1_C
TGAGTCA	535	1406	0.38	39.9	V$BACH2_01
TGCRTCA	164	616	0.27	19.9	-
TGACKCAC	66	230	0.29	14.7	V$BACH2_01
TGANTMATC	106	455	0.23	14.1	-
TGTNANTCA	236	1240	0.19	13.4	-
TGANYCAGA	150	798	0.19	13.4	-
TGANTNRCAG	57	234	0.24	11.7	-
AATKANTCA	160	854	0.19	11.7	-
WGACNCACCY	59	286	0.21	10.7	-
TGCNGCA	592	3362	0.18	10.2	-
TGAYNCAA	74	344	0.22	9.5	-
>cluster_8					
TMTCGCGANR	236	368	0.64	55.7	-
MTCGCGAGA	202	321	0.63	49.1	-
>cluster_9					
TGAYRTCA	466	924	0.50	55.7	V$ATF3_Q6
GTGANNNCAC	119	615	0.19	15.4	-
TGANNWMATC	62	217	0.29	14.5	-
YKTCATCA	59	222	0.27	13.1	-
GCTGANNTCA	90	424	0.21	12.8	-
TGAnnTSACA	130	735	0.18	12.6	-
CATKANNTCANY	58	297	0.20	12.1	-
TGATGTMR	54	172	0.31	11.4	-
ATGRNNTCAT	102	621	0.16	10.2	-
>cluster_10					
GCCATNTTG	316	452	0.70	54.7	V$YY1_Q6
ANATGGCG	459	821	0.56	50.1	V$YY1_Q6
CCAWNWTGG	238	590	0.40	31.1	-
GCCATTKT	176	345	0.51	26.8	V$YY1_Q6
KCCATTTTRT	51	80	0.64	20.3	-
MAGATGGY	283	1165	0.24	18.9	V$TAL1BETAE47_01
AAANATGGM	76	211	0.36	18.9	V$YY1_Q6
KCCATNTTA	60	130	0.46	17.3	V$YY1_Q6

RNGGCCATNT	69	233	0.30	14.6	V$NFMUE1_Q6
TCAMNATGG	57	179	0.32	14.2	-
GGCNNYTTTRW	74	331	0.22	13.6	-
GCCYWYGTG	80	463	0.17	12.1	-
CCAWNKTTG	69	317	0.22	11.8	-
GGCNNSNTTAW	56	319	0.18	11.6	-
RNAGCCATNT	54	201	0.27	11.4	V$YY1_Q6
GGAMNATGSY	58	264	0.22	10.9	-
WNATGRCTG	59	252	0.23	10.1	-
RGAANWTGGC	72	339	0.21	9.9	-
GGCYMTTTW	59	297	0.20	9.9	-
GGCNATTKK	78	403	0.19	9.8	-
CGGCCATYK	51	226	0.23	9.8	V$NFMUE1_Q6
GCCMWYKTGC	51	260	0.20	9.6	-
GCCWNATTK	58	310	0.19	8.3	-
>cluster_11					
MGGAAGTG	474	1184	0.40	51.6	V$GABP_B
GGAARTGAYR	128	221	0.58	39.9	-
SMGGAAGT	430	1412	0.30	34	V$ETS_Q4
CATTTCCK	181	913	0.20	14.6	V$STAT1_02
RCCACWYCCT	73	331	0.22	13.6	-
MGGAARYGAG	57	192	0.30	13.4	-
>cluster_12					
CAGGTG	1151	4397	0.26	47.6	V$E12_Q6
CAGNTGG	1626	6895	0.24	35.7	V$MYOD_Q6
CCANNTGGY	540	2361	0.23	25	-
RYAGGTGG	326	1323	0.25	24	V$E12_Q6
AGGTGA	501	2336	0.21	23	V$AREB6_02
RNCAGNWGGT	252	1368	0.18	15.4	-
YRGGTGGC	213	1119	0.19	11.5	-
RNCAGRKGGCA	67	339	0.20	11.4	-
RTCAMCTT	56	244	0.23	10.5	-
AACMANMTGG	63	334	0.19	9.3	-
CAANNTGAY	61	320	0.19	8.6	-
>cluster_13					
CTTTGT	763	1887	0.40	46	V$LEF1_Q2
YYTTTGTC	358	1509	0.24	20.6	-
TTTGTG	1145	6210	0.18	18.2	-
YTTTRTCT	362	1964	0.18	16.9	-
YCTTTKRTCT	55	256	0.21	10	-
TTTGTA	132	666	0.20	7.7	-
>cluster_14					
TGACGTCA	316	712	0.44	44.8	V$ATF3_Q6
TGACGTMR	436	1008	0.43	50.7	V$CREB_01
>cluster_15					
CAGCTG	2468	9256	0.27	43.9	V$AP4_Q5
CAGNTGT	1286	5437	0.24	31.4	V$E47_01
CCANNTGTY	506	2401	0.21	21.6	-
RACANSTGT	254	1086	0.23	20.6	-
TGAYWNATG	213	1082	0.20	16.1	-
ACASRTGGY	61	201	0.30	15.1	V$TAL1BETAE47_01
TGAYWNNTGA	191	1108	0.17	14	-
CCASNTGTG	67	406	0.17	10.7	-
TGACRNNTGT	72	394	0.18	9.6	-
>cluster_16					
RYTTCCTG	807	2616	0.31	43	V$ETS2_B
YTTCCKGTT	135	522	0.26	17.9	V$ELK1_02
GCCARGAA	176	1019	0.17	12	-
YAATTTCCT	56	318	0.18	10.2	V$HMGIY_Q6
>cluster_17					
AACTTT	617	1471	0.42	42.1	-
AAGTTT	1414	6480	0.22	28.2	-
GAAGTT	367	1694	0.22	19.3	-
GAACTT	784	3928	0.20	16.8	-
AAAGTG	252	1546	0.16	10.8	-
KNAACTTGRY	84	486	0.17	9.3	-
AAGTGAA	53	327	0.16	6.1	V$IRF1_Q6
>cluster_18					
TCANNTGAY	317	704	0.45	40.4	V$SREBP1_01
YCACRTGAY	112	347	0.32	17.6	V$SREBP1_01
RTCACATGNY	61	249	0.25	12.8	-
CAGNTGAC	208	1054	0.20	12.7	-
RTCACATK	52	189	0.28	9.9	-
>cluster_19					
GKCGCNNNNNNNTGAYG	50	154	0.32	40.1	-
RTCATNNNNNNGCG	53	206	0.26	15.1	-
YMATCNNNNNGCGM	53	327	0.16	12.1	-
KNCATNNNNNNGCGC	56	345	0.16	9.9	-
>cluster_20					
GTGACGY	543	1699	0.32	38.4	V$E4F1_Q6
RCGTCATY	115	389	0.30	19	V$CREB_02
RGGTGACNY	213	957	0.22	16.3	V$AP1FJ_Q2
RAGTGACNY	122	538	0.23	15.8	-
TGTGAC	256	1426	0.18	13.7	-
TGAYGTGNY	63	212	0.30	12.9	V$ATF6_01
GCGYYATTK	62	208	0.30	12.7	-
GGTNACNTTG	54	210	0.26	11.8	-
GGTGACNT	115	637	0.18	11.8	V$CREB_02
TGACGTGK	51	188	0.27	11.4	V$ATF6_01
ACGTSACT	62	326	0.19	11.3	-
GTGATG	184	1087	0.17	8.8	-
>cluster_21					
GGAANCGGAANY	82	140	0.59	37.7	-
MGGAANCGGAA	87	150	0.58	36.3	-
>cluster_22					
TGCGCANK	600	2481	0.24	37.4	-
TGCGCAGGC	99	450	0.22	21.7	-
CMTGCKYAGT	53	208	0.25	17.7	-
>cluster_23					

TAATTA	1614	5584	0.29	37.3	V$CHX10_01
TATTTAW	137	366	0.37	18.5	V$TBP_01
CTAATTW	514	2537	0.20	17.7	V$CHX10_01
CTAWTTANR	55	146	0.38	13.8	V$CHX10_01
ATTTAANK	84	318	0.26	10.4	-
ATATTTR	66	222	0.30	10.1	-
AAAYATT	71	315	0.23	9.6	V$FOXJ2_02
STGTMATTA	57	294	0.19	8.7	-
GTAATT	94	470	0.20	8.3	V$NCX_01
>cluster_24					
GGGAGGRR	1393	8773	0.16	33.5	V$MAZ_Q6
CCCYTCCCCC	297	1102	0.27	31.2	-
YYCCTCCCYY	608	3350	0.18	28.6	V$MAZ_Q6
GNGGGAGG	799	4166	0.19	19.6	V$TFIII_Q6
RGGAGGAG	456	2615	0.17	17.3	-
GTGGGAGG	166	917	0.18	12.1	-
>cluster_25					
TGACCTY	727	2418	0.30	33.4	V$ERR1_Q2
TGACCT	1137	4121	0.28	32	V$ER_Q6_02
GTGACCY	438	1900	0.23	18.7	V$ER_Q6_02
TGAMCTTT	191	909	0.21	15.2	V$COUP_01
YYTTGACCY	120	605	0.20	13.4	-
GAAGGTMR	75	401	0.19	12.5	-
WGAGSTCAY	52	228	0.23	12	-
YTTGAMCTT	90	450	0.20	11.9	V$GNCF_01
TRACCYNNTTT	67	365	0.18	11.4	-
AGGTNAGT	148	797	0.19	11	-
RGATCAARK	94	503	0.19	10.1	-
GGTRAGT	119	650	0.18	10.1	-
CTGWCCTTNR	74	411	0.18	9.4	V$T3R_Q6
TGACCTANW	66	339	0.19	8.8	V$RORA2_01
>cluster_26					
TTAYRTAA	436	1288	0.34	32.6	V$E4BP4_01
>cluster_27					
TGGNNNNNNKCCAR	421	1659	0.25	32.3	-
GGCNNNNNKCCAR	394	1470	0.27	31.2	-
YNGGCNNNNNNYCAAR	97	452	0.21	15.7	-
TTGRNNNNNNTCCAR	71	357	0.20	13.7	-
>cluster_28					
CTAWWWATA	351	1054	0.33	32.3	V$RSRFC4_Q2
TATNNATA	112	260	0.43	19.2	-
>cluster_29					
CTTTAAR	299	790	0.38	30.8	-
KNCCCTTTAA	74	168	0.44	23.5	-
CCCYKKAAG	131	682	0.19	16.8	-
CCCYTTTRW	91	350	0.26	14.9	-
GCCYNTTAA	59	221	0.27	13.2	-
WWTAAAGT	54	184	0.29	11	-
>cluster_30					
YGCGYRCGC	431	2232	0.19	30.5	-
>cluster_31					
GGGYGTGNY	349	1640	0.21	30	-
AGGYGTG	324	1630	0.20	13.6	-
>cluster_32					
TGASTMAGC	139	359	0.39	27.2	V$NFE2_01
GCTGWGTCAY	59	125	0.47	22.8	V$NRF2_Q4
GCTRANNCAGS	71	394	0.18	9.5	-
>cluster_33					
YTATTTTNR	661	3255	0.20	26.4	V$MEF2_02
YTATTTWTAA	91	408	0.22	14.6	V$AMEF2_Q6
TTATTT	231	1018	0.23	14.2	-
TRTTTTGG	58	210	0.28	11.5	-
YRGAAATARM	91	541	0.17	10.1	-
CTAWWTTARS	58	301	0.19	8.9	-
YTATTTNTGG	58	306	0.19	8.5	-
>cluster_34					
CYTAGCAAY	73	165	0.44	26.1	-
GTTRCYAGG	64	211	0.30	19	-
GCARCCAWT	71	285	0.25	14.9	-
GCTAAT	500	2446	0.20	14.6	-
RYTGCYAAGR	68	312	0.22	12.3	-
MTTAGCAW	55	199	0.28	11.9	-
GGTTGCYA	65	292	0.22	11.7	-
GCTRATGR	156	920	0.17	9.5	-
YRGCAACCR	66	340	0.19	8.5	-
>cluster_35					
GCANCTGNY	671	2881	0.23	25.7	V$MYOD_Q6
CAGATG	1145	5153	0.22	25.1	V$TAL1BETAE47_01
GCASSTGC	418	2482	0.17	23.6	-
RACANSTGC	113	600	0.19	14.9	V$E47_01
GCANCWGCT	189	1005	0.19	13.1	-
GCANCANCTG	82	452	0.18	9.9	-
MCAGATGK	72	391	0.18	9.8	V$TAL1BETAE47_01
>cluster_36					
RTAAACA	204	481	0.42	25.6	V$FREAC2_01
RTAAATA	852	3557	0.24	26.3	V$TBP_01
RRGTAAACA	94	387	0.24	13.6	V$FREAC4_01
TRTTTACT	154	828	0.19	11.3	V$FREAC3_01
CTCRNRTTTC	59	301	0.20	10.8	-
MNGTAANCAGR	57	280	0.20	10.4	-
GTAANYNGAG	58	297	0.20	10.3	-
TCTNTTTA	57	242	0.24	9.4	-
TRTTTACCW	60	310	0.19	8.8	V$FREAC2_01
>cluster_37					
GTTRYCATRR	69	142	0.49	25.3	-
MYATGRNNACC	54	208	0.26	13.4	-
>cluster_38					

TGACCTTG	162	434	0.37	25.2	V$SF1_Q6
GTGWMCTT	82	390	0.21	12.5	-
GTGNCMTTG	54	278	0.19	11.9	-
YSACCWTGG	54	319	0.17	10.6	-
GTGRNYTTGG	77	478	0.16	10	-
>cluster_39					
TCCCRNNRTGC	138	502	0.27	24.3	-
TCCCRNCATNC	56	191	0.29	16.1	-
YNCCANNWTGC	132	645	0.20	15	-
SCATNNTRGGA	52	261	0.20	9.4	-
>cluster_40					
TTCYNRGAA	468	2414	0.19	24.3	V$STAT5B_01
GCGYSGGAAR	61	370	0.16	13.3	-
TCCCRGAAR	84	499	0.17	12.5	V$STAT_Q6
CGCNCMGGAA	53	325	0.16	12.2	-
KTCCTNGAA	136	781	0.17	11.8	V$STAT5B_01
YYCTGGGAAA	59	299	0.20	11.5	-
TTTYYNNGAANK	55	259	0.21	10.7	V$STAT5B_01
YNTCTRGGAAW	58	341	0.17	10.5	-
TTCYCACRS	65	415	0.16	10.5	-
TKCTGRGAA	52	311	0.17	10.4	-
TTTCCCANR	79	503	0.16	9.6	V$IK1_01
>cluster_41					
TGACAGNY	594	2363	0.25	24.1	V$MEIS1_01
TGACAG	486	2007	0.24	23.8	V$MEIS1_01
CTGNCAGNY	635	3598	0.18	17.3	-
GTGACA	349	1642	0.21	16.9	V$MEIS1_01
KGACAGCTS	68	355	0.19	12.8	V$TGIF_01
TTGACA	160	671	0.24	12.2	-
CTGACARY	211	1105	0.19	12.2	V$TGIF_01
ATGACA	712	3938	0.18	12.2	V$TCF11_01
RNTTGTCA	90	394	0.23	10.9	-
TGTCATKK	51	191	0.27	9.9	V$TCF11_01
RNTGTNAAA	72	345	0.21	9	-
>cluster_42					
TGACATY	557	2425	0.23	23.8	-
RYGATGTCAY	79	185	0.43	23.9	-
RATGNCATY	238	1226	0.19	16.4	-
GAAKKTCA	63	280	0.23	12.6	-
RATGWCAG	235	1403	0.17	11.8	-
GATRNYATC	100	518	0.19	11.4	-
GTGWCATY	61	236	0.26	10.9	-
SCTGMCATNK	52	245	0.21	10	-
>cluster_43					
GTTGNYNNRGNAAC	69	152	0.45	23.7	-
GNTGCYNNRKNAAC	54	192	0.28	22.2	-
GTTGNYNNNNNGAC	55	244	0.23	11.8	-
KTTGNYNNRGAANY	74	399	0.19	10.7	-
TTGNYNNNNTGAYR	68	400	0.17	9.8	-
>cluster_44					
YATGNWAAT	105	199	0.53	23.5	V$OCT_C
ATTWSCAT	109	248	0.44	20.1	V$OCT_C
YATGYAAAT	149	501	0.30	19.9	V$OCT_Q6
TTTNAAT	251	879	0.29	19.7	-
RTTTGAAY	75	207	0.36	16.5	-
YNATTTGCRW	50	124	0.40	15.1	V$OCT1_B
RTTTNMATG	59	210	0.28	12.2	-
TGTNAATNR	75	289	0.26	11.6	-
TTTNAAC	161	838	0.19	11.5	-
RRATGNNAATTNR	50	242	0.21	9.4	-
YAATTTGY	200	1213	0.16	9.2	-
ATTARCAT	103	588	0.18	8	V$POU3F2_02
>cluster_45					
CCANNAGRKGGC	75	185	0.41	23.4	-
CACNAKRKGGC	54	152	0.36	18.3	-
CNGCANRKGNCGC	50	244	0.20	15.4	-
CACNWGRGGGY	48	168	0.29	14.1	-
GCCWNYWKGTG	54	206	0.26	13.5	-
GCCNYNNNGTGRY	86	399	0.22	13.1	-
CCASYAGGKG	56	260	0.22	12.9	-
CAGYRSRNGGC	57	258	0.22	12.7	-
CCATYKGCT	82	381	0.22	11.5	-
KMCATNNNGTG	53	222	0.24	10.5	-
CCAYCYRCTG	53	229	0.23	10.3	-
>cluster_46					
WTTGKCTG	235	932	0.25	23	-
CAGMSAATG	57	274	0.21	13	-
GATTKSCTG	83	350	0.24	12.4	-
WTTGKYTGG	72	416	0.17	10.9	-
WWTGTCTG	56	272	0.21	8.6	-
>cluster_47					
TGCCAAR	652	2843	0.23	22.9	V$NF1_Q6
TTGNCAAR	464	2597	0.18	15.4	-
GTGCCAR	366	2015	0.18	14.1	-
>cluster_48					
GCGNNANTTCC	98	350	0.28	22.8	-
GCGKNAYTTCC	62	171	0.36	22.3	-
RNGGGANTTCC	54	183	0.30	17.3	V$CREL_01
RGGANNYCCC	251	1512	0.17	12.7	V$NFKAPPAB_01
SGGAYTTCYY	55	270	0.20	10	-
>cluster_49					
CATTGTYY	119	296	0.40	22.5	V$SOX9_B1
YWTTGTTC	181	439	0.41	25.5	V$SOX9_B1
YYATTGTT	103	220	0.47	20	V$SOX9_B1
YYATTGTCY	63	146	0.43	19.2	-
CCCWTTGTNY	70	210	0.33	14.6	-
GGCNTTGTNY	97	521	0.19	10.1	-
RNACAAWGAC	59	298	0.20	9	-
>cluster_50					
RGAGGAARY	333	1612	0.21	22.4	V$PU1_Q6

GAGGAAGY	251	1184	0.21	16.8	V$PU1_Q6
RYTTCCTTY	287	1646	0.17	14.1	-
GAGRAAGTK	61	341	0.18	10	-
>cluster_51					
TATAAA	209	513	0.41	22.1	V$TATA_01
TTTNNAAA	400	1650	0.24	22.1	-
AATAAA	315	1068	0.29	21.5	-
YNATTNATA	97	185	0.52	19.2	-
TATATA	128	290	0.44	18.7	V$TATA_01
YTTGYAAA	436	2515	0.17	16.5	-
TGTAAAY	547	2686	0.20	16.2	-
ATAAAA	208	805	0.26	15.9	V$TATA_C
YNATAAAG	103	351	0.29	15.5	-
YATAWAAG	63	185	0.34	14.4	V$TATA_01
YTATWAAA	62	176	0.35	13.5	V$TATA_C
YNCATAAANM	51	157	0.32	11.1	-
TTTRYMAACA	65	342	0.19	10.1	-
>cluster_52					
YYCATTCAWW	66	122	0.54	21.6	-
ATGAATRR	110	237	0.46	19.8	V$POU1F1_Q6
YYATTYATT	68	179	0.38	15.4	-
TGAYNGACR	72	273	0.26	13.1	-
TGANTGA	211	1044	0.20	13.1	-
CATWMATT	54	169	0.32	11.1	-
ATGRRTGA	52	201	0.26	9.9	-
CCATNCAT	57	220	0.26	9.7	-
TAATTTATK	59	325	0.18	7.8	V$POU6F1_01
>cluster_53					
RYTGCNNRGNAAC	52	164	0.32	21.3	V$MIF1_01
GTTNSNNRNCAAY	55	165	0.33	18.9	-
GTTNMNNNNNAAC	110	513	0.21	15.5	-
TTAnnnnKTAACY	69	296	0.23	14.1	-
RTTGYNNNGCA	58	279	0.21	12.2	-
TTAMNNNKNRACC	58	328	0.18	9.5	-
>cluster_54					
TAAWWATAG	147	534	0.28	21.1	V$RSRFC4_Q2
CTANRTTTMS	57	157	0.36	15.4	-
GCTRNTTTNR	266	1474	0.18	14.4	-
TATWWWTAAC	63	261	0.24	12.6	-
TAAWAATM	53	173	0.31	10.5	-
AATAAT	104	401	0.26	10.1	V$CDP_01
>cluster_55					
TGGAAA	1659	9001	0.18	21.1	V$NFAT_Q4_01
MATTTCC	218	1181	0.18	13.9	V$HMGIY_Q6
RGAAACTY	296	1710	0.17	13.7	-
GTTTCC	930	5470	0.17	12.6	V$NFAT_Q4_01
GGAANSTTT	64	324	0.20	10.6	-
RGGAAACTK	56	342	0.16	10.2	-
YNATGGAARC	56	295	0.19	9.6	-
>cluster_56					
GGGTGGRR	822	4394	0.19	20.9	V$PAX4_03
GGGNGGGG	2362	12769	0.19	59.1	V$MAZR_01
GGGNGGAGYY	143	568	0.25	17.2	-
GGGTGG	1589	8365	0.19	16.5	V$PAX4_03
AGCYMCNCCC	152	783	0.19	15.6	-
GGGYGNNGTC	82	402	0.20	11.7	-
>cluster_57					
ACCTGTTG	99	258	0.38	20.7	-
YRACAGGT	110	336	0.33	22.5	-
CACNTSYTGC	74	325	0.23	15.6	-
CYTGTTGC	79	275	0.29	14.3	-
ACTTSNTGC	62	254	0.24	13.2	-
CACWWNCTGY	55	260	0.21	11.6	-
CAGRANRTGA	93	533	0.17	11	-
CAAYWRGTG	66	300	0.22	11	-
MACCKGTTT	58	273	0.21	10.3	-
GTCAYNKCCT	57	301	0.19	9.8	-
CAAYARATG	58	315	0.18	8.8	-
GCCTGTTKM	55	304	0.18	8.4	-
>cluster_58					
YCATTAA	522	2362	0.22	20.3	-
TCATTANY	389	2156	0.18	13	V$IPF1_Q4
ATTAAT	110	334	0.33	10.9	-
KCATNAAT	73	219	0.33	10.7	-
GCCWTTAM	64	301	0.21	10.4	-
YYAATGAG	80	397	0.20	10.1	V$ALPHACP1_01
GCCYWTAA	60	282	0.21	10.1	-
TAAWGGCnS	66	333	0.20	9.5	-
TTAANGAG	56	280	0.20	8.1	-
YCATGAA	60	294	0.20	7.9	-
>cluster_59					
WCTCNATGGY	51	127	0.40	19.9	-
CATGGCNR	258	1005	0.26	22.4	-
CCATGGM	246	1073	0.23	18.1	-
KCTATGGY	54	203	0.27	12.9	-
CCGYYATGW	61	169	0.36	12.4	-
CTCTATRR	54	187	0.29	12.2	-
>cluster_60					
TTGTTT	391	1430	0.27	19.8	V$FOXO4_01
TTGNTTT	686	4000	0.17	12.84	-
TGTTTTGR	52	206	0.25	11.4	V$FAC1_01
WMAACAAG	69	260	0.27	11.3	V$FOXO4_01
MNTTGTTA	60	211	0.28	10.5	-
AAAYAATR	56	181	0.31	9.6	V$SOX5_01
>cluster_61					
YTAATTAA	245	901	0.27	19.8	V$LHX3_01
TTAATT	1627	7181	0.23	30.4	V$NKX61_01
TTTNATT	259	1053	0.25	16.6	-
AATNAAG	1049	6119	0.17	14.7	-
ACTNYATTNC	58	179	0.32	13.9	-
TTAATTG	238	1204	0.20	12.5	V$NKX61_01

YNTAAWYAAC	111	562	0.20	11.8	-
KMAATWWAGC	63	299	0.21	10.3	-
TTTWAWTAGY	57	276	0.21	9.2	-
ATTNNATT	93	436	0.21	8.7	-
TAANMAAG	72	366	0.20	8.5	-
>cluster_62					
SMTTTTGT	150	410	0.37	19.1	-
ACAAWAGC	65	163	0.40	17.2	-
MTTTTGT	124	374	0.33	15.6	-
CCCYWTTGT	56	163	0.34	14	-
CCTWTTGT	55	161	0.34	12.4	-
TCTWTTGT	151	749	0.20	11	-
CGCYWTTGY	64	329	0.19	10.5	-
GCCNWTTGTY	55	219	0.25	9.9	-
GCTKWYGTC	54	310	0.17	8.8	-
>cluster_63					
AAGWWRNYGGC	90	260	0.35	19.1	-
GCCNNNWTGTT	54	238	0.23	11.3	-
AAGWWNNYNGCG	54	280	0.19	9.9	-
CGCNNNNWWGTT	60	374	0.16	9.7	-
>cluster_64					
TTANTCA	841	4288	0.20	18.8	-
CTTWGTCAY	57	286	0.20	9.4	V$AP1_Q4
RTTTMRTCA	88	524	0.17	9.2	-
>cluster_65					
ARGGGTTAA	85	209	0.41	18.7	-
RRGGTTAA	242	972	0.25	16.6	V$PXR_Q2
RGTTAAA	443	2460	0.18	14.2	-
RNAGTTAAY	188	1059	0.18	10.9	-
WTTAACCTY	57	303	0.19	8.9	-
>cluster_66					
RACTNNRTTTNC	62	199	0.31	18.5	-
>cluster_67					
TGANNYRGCA	230	1045	0.22	17.5	V$TCF11MAFG_01
TGACWYRGCA	55	169	0.33	14.1	V$TCF11MAFG_01
RTGANNNWGCA	68	300	0.23	12.8	V$TCF11MAFG_01
TTGNYNNRNCAA	134	810	0.17	11	-
YTGCNNYRNCAA	96	509	0.19	10.2	-
TGAYNYNRCAA	79	417	0.19	9.3	-
>cluster_68					
RGAANNTTC	177	1006	0.18	17.4	V$HSF1_01
>cluster_69					
SGCGSSAAA	141	683	0.21	17.3	V$E2F1DP2_01
SGCGCCAWW	81	337	0.24	12	V$E2F_03
KTTGRCGC	84	505	0.17	9.9	V$E2F1DP1RB_01
>cluster_70					
CGTSACG	256	1410	0.18	17.2	V$PAX3_B
>cluster_71					
SYATTGTG	87	232	0.38	17.1	-
YWTTGTGT	116	315	0.37	16.9	-
SCATTNTGG	78	253	0.31	16.9	-
GNATTSTGGG	56	182	0.31	14.7	-
YWTTGTGC	106	376	0.28	14.5	-
CTTYSTGT	123	639	0.19	14.3	-
YWTTGTGA	65	238	0.27	11.5	-
YYATTGTG	160	887	0.18	10	-
SCATGNTGG	58	271	0.21	9.6	-
WTTGYGTG	59	308	0.19	7.5	-
>cluster_72					
TTCYRGAA	144	734	0.20	17.1	-
TTTNNKGAA	118	728	0.16	9.7	-
>cluster_73					
CTTTGA	1029	5587	0.18	17	V$LEF1_Q2
CCTYTGAYY	124	745	0.17	9.7	-
WTCAAAG	94	494	0.19	9.2	V$TCF4_Q5
RCCTNTRATT	55	299	0.18	9.2	-
>cluster_74					
GGAMTNNNNNTCCY	86	498	0.17	16.7	-
GACKNYNNNTCCYR	52	233	0.22	11.3	-
>cluster_75					
TNCATNTCCYR	66	220	0.30	16.5	-
YATTTCCCR	51	228	0.22	12.2	V$STAT1_02
YNGGANTTGYA	63	296	0.21	10.5	-
>cluster_76					
CAGGTA	227	1137	0.20	16.3	V$AREB6_01
AGGTAA	162	783	0.21	12.8	-
YYCAGRTAA	106	628	0.17	9.5	-
>cluster_77					
AAAYRNCTG	297	1657	0.18	16.3	-
CAARYRTTT	175	1055	0.17	14.1	-
AAAYNGYTGA	62	316	0.20	10.4	-
CAAWNYTTT	65	272	0.24	9.5	-
>cluster_78					
GCTNWTTGK	141	622	0.23	16.2	-
CGCNNATTGG	83	407	0.20	14.3	-
CTCWTTGT	56	216	0.26	11.1	-
GCGNNNATTGKY	51	286	0.18	10.8	-
CCTNWTTGK	82	524	0.16	10.5	-
GCTSTTTRY	53	261	0.20	9	-
RNCCANTSAGC	60	297	0.20	8.7	-
>cluster_79					
WGGAATGY	293	1398	0.21	16.1	V$TEF1_Q6
TGGAATKY	302	1623	0.19	15.6	-
RMATTCCT	94	386	0.24	14.5	-

RAATTCC	459	2656	0.17	12.6	-
GGACTTY	337	1883	0.18	11.9	-
SGGACTTYC	54	313	0.17	11.8	V$NFKB_C
RKGAAGTC	196	1127	0.17	11	-
WGGAAYGTG	62	345	0.18	8.7	-
GGAWTGT	95	515	0.18	7.9	-
WKGAATTT	50	229	0.22	7.7	-
>cluster_80					
SNACANNNYSYAGA	50	195	0.26	15.8	-
YKACANNNNNCAGA	61	322	0.19	11.5	-
>cluster_81					
CGGAARNGGCNG	50	235	0.21	15.7	-
>cluster_82					
CTGYNNCTYTAA	55	170	0.32	15.5	-
CTGNNNYTTTRA	52	170	0.31	15.3	-
YTGTNNMWTTAA	72	339	0.21	13	-
>cluster_83					
TGTTTGY	590	3090	0.19	15.1	V$HNF3_Q6
TGTTTGTNY	61	226	0.27	11.6	V$HNF3_Q6
TGTYKRYTTT	128	790	0.16	10.9	-
AAGYAAACA	70	367	0.19	9.8	V$FREAC4_01
YGTTTGCTY	64	364	0.18	6.9	V$DBP_Q6
>cluster_84					
RGTTAMWNATT	63	228	0.28	15	V$HNF1_01
GTTWNWNATTAR	55	248	0.22	12	-
AGTNNNYNRTTAR	77	445	0.17	10	-
>cluster_85					
STTTCRNTTT	120	568	0.21	14.9	V$IRF_Q6
GAANNMGGAARY	61	240	0.25	18.1	-
RNAAGNRGAARY	144	760	0.19	14.8	-
RGAANNGAAANY	139	829	0.17	13.7	V$IRF_Q6
WTCCKSGTT	55	303	0.18	12.7	-
AAANNGGAANY	68	383	0.18	10.6	-
AAANMRNAAGY	62	337	0.18	10.6	-
RNTTCNTYATT	79	464	0.17	9.4	-
>cluster_86					
GGGNNTTTCC	91	420	0.22	14.9	V$NFKB_Q6_01
GGAAYNTCCY	65	296	0.22	11	V$NFKB_Q6
TGGNAWNYCCC	59	328	0.18	10.3	V$NFKAPPAB_01
KGAANTTCC	58	287	0.20	10.3	V$NFKAPPAB65_01
GGANTTTSC	117	693	0.17	10.1	V$NFKAPPAB65_01
TGGRRWTTCY	84	473	0.18	9.5	-
>cluster_87					
RYTGCNWTGGNR	49	201	0.24	14.6	-
YTGCYNNGGTRW	54	255	0.21	8.6	-
>cluster_88					
GGCNKCCATNK	80	307	0.26	14.3	-
CGGNNRYCATNK	78	306	0.25	13.7	-
CGCNKNNGYCATNK	50	203	0.25	13.5	-
GGCNNYNKMCAT	53	309	0.17	9.5	-
ATGRMNKYGGC	46	231	0.20	9.3	-
>cluster_89					
GTTNYYNNGGTNA	60	263	0.23	14.3	-
YYGCCNNRRNAAC	72	422	0.17	15.2	-
GTTNYCNNNGAGNY	62	258	0.24	13.2	-
KYTCCNNRRNAAC	52	292	0.18	12.6	-
>cluster_90					
YAATNRNNNYNATT	128	624	0.21	14.3	-
YAATYRNNNYAAT	62	278	0.22	13.1	-
WAATNNYNATTAR	53	286	0.19	9.2	V$CART1_01
>cluster_91					
GTGGGTGK	213	1062	0.20	14.1	-
>cluster_92					
TGCTGAY	409	2089	0.20	14	-
YGCTGACTY	65	250	0.26	13.2	-
GCTGACRK	143	794	0.18	13.2	-
WTCARCAC	63	289	0.22	11.2	-
RTGCTGAMR	52	242	0.21	11.1	-
GGTGCTKA	67	380	0.18	10.3	-
TGCWRATT	232	1399	0.17	9.8	-
>cluster_93					
GGATTA	501	2421	0.21	14	V$PITX2_Q2
>cluster_94					
TGATTTRY	252	1365	0.18	13.9	V$GFI1_01
TGATTTANR	147	753	0.20	12.5	-
AAAYCACNR	234	1342	0.17	11.6	-
YNGTGATTNR	151	801	0.19	10.9	V$GFI1_01
TGATYTATNR	55	255	0.22	10.1	-
GTGATTRR	80	355	0.23	10.1	-
GTAAATM	58	220	0.26	8.9	V$FREAC3_01
>cluster_95					
GCCNNNWTAAR	70	303	0.23	13.7	-
GCCNNTWWAAG	52	173	0.30	15.4	-
CCCNWWTAA	59	277	0.21	10.6	-
>cluster_96					
YGCANTGCR	124	710	0.17	13.7	-
CYGCANTGC	99	428	0.23	11.3	-
RNAATGCA	88	514	0.17	7.8	-
>cluster_97					
YATTNATC	330	1772	0.19	13.7	-
RATCRATA	114	453	0.25	12.2	V$CDP_02
>cluster_98					
GTCNYYATGR	54	200	0.27	13.6	-

GTCKCCATNK	56	211	0.27	12.6	-
GGTYNYNATG	61	304	0.20	11.6	-
YYATGSAGAY	53	263	0.20	9.5	-
GTCNNNNTGGYR	96	528	0.18	9.5	-
TGTYNSYATGR	55	303	0.18	9.4	-
>cluster_99					
ATCMNTCCGY	49	129	0.38	13.3	-
KRTCCNTCCSC	51	254	0.20	11.2	-
>cluster_100					
CRGAARNNNNCGA	53	285	0.19	13.3	-
>cluster_101					
CTGCAGY	693	4061	0.17	13.2	-
>cluster_102					
ATGGYGGA	58	212	0.27	13.2	-
>cluster_103					
ACAWNRNSRCGG	55	213	0.26	13.1	-
>cluster_104					
CCAATNNSNNNGCG	55	294	0.19	13	-
>cluster_105					
ACTWSNACTNY	54	232	0.23	13	-
>cluster_106					
CCGNMNNTNACG	68	380	0.18	12.9	-
>cluster_107					
RTTTNNNYTGGM	144	813	0.18	12.8	-
>cluster_108					
AACWWCAANK	56	208	0.27	12.7	-
TGTTGTK	112	542	0.21	10.1	V$FAC1_01
GAAYNACANY	59	287	0.21	10	-
GAANKWCAA	57	294	0.19	9.2	-
>cluster_109					
YGTCCTTGR	69	307	0.22	12.7	-
>cluster_110					
MCAATNNNNNGCG	66	320	0.21	12.5	-
>cluster_111					
RACCACAR	222	1332	0.17	12.3	V$AML_Q6
AACYRNAAC	55	313	0.18	9.8	-
>cluster_112					
KTGGYRSGAA	52	228	0.23	12.3	-
TTGNYRGGAR	64	351	0.18	12.3	-
>cluster_113					
AACYNNNNTTCCS	53	281	0.19	12.3	-
>cluster_114					
YTCCCRNNAGGY	50	327	0.15	12.2	-
TCCMANNWGGC	59	348	0.17	11.7	-
>cluster_115					
YRTCANNRCGC	56	309	0.18	12.2	-
>cluster_116					
KMCATNNWGGA	52	181	0.29	12.2	-
>cluster_117					
TGTYNNNNNRGCARM	69	373	0.19	12.1	-
GTCNYNNNRGCAMS	50	268	0.19	11.5	-
GCCNNNTGACR	57	253	0.23	9.4	-
GGCNNNRTGAC	67	343	0.20	9.3	-
GTCAYNNNGGC	55	298	0.18	8	-
>cluster_118					
GGCNRNWCTTYS	51	320	0.16	12	-
>cluster_119					
GGGNRMNNYCAT	56	295	0.19	11.9	-
ATGGYNRCGC	51	231	0.22	11.5	-
MAATRGNNGCG	50	253	0.20	10.1	-
WAATNNCNGCG	50	223	0.22	9.4	-
>cluster_120					
KRCTCnnnnMAnAGC	48	185	0.26	11.8	-
GCTMTNWWNAGA	50	162	0.31	13.4	-
STCTNNWNRGAGNC	52	209	0.25	13	-
YTCTNNNNARNGCCNY	50	215	0.23	11.8	-
RNGGCTNYNNNNAGAS	50	220	0.23	11.8	-
GCTNWNNNNAGAGYM	52	203	0.26	11.5	-
TCCNMYAAT	62	311	0.20	10.4	-
CTCNNYNAATNR	56	326	0.17	10.1	-
RCTCNNNWMAGANS	49	222	0.22	9.3	-
>cluster_121					
CCAWWNAAGG	62	270	0.23	11.7	V$SRF_Q4
TCCWWWNWTGG	61	300	0.20	11.4	V$SRF_Q4
YTCCRKNTTG	56	321	0.17	9.4	-
MCAATNNNGAG	49	258	0.19	9.2	-
CTTWNWAGG	57	310	0.18	9.1	-
YNAATNAGG	156	876	0.18	9.04	-
GCCNWWWAAG	27	100	0.27	9.02	-
SYAAAYRAGG	53	258	0.21	8.7	-
TCCNNATTRR	57	312	0.18	8.3	-
>cluster_122					
RNTCANNRNNYNATTW	62	352	0.18	11.7	-
>cluster_123					
GGCNNMSMYNTTG	54	314	0.17	11.6	-
GGCNNNNNNATTGK	55	301	0.18	10.4	-

>cluster_124					
CCAWYNNGAAR	60	320	0.19	11.5	-
MCAATNRGAG	74	362	0.20	16.1	-
RCCAAYNGGAR	61	265	0.23	15.4	-
TTTNNNWATGR	53	170	0.31	11.4	-
TCTNRYTGGY	101	593	0.17	9.9	-
>cluster_125					
RAAGNYNNCTTY	144	878	0.16	11.5	-
>cluster_126					
WYAAANNRNNNGCG	52	268	0.19	11.4	-
TYAAANNNNNCGC	61	252	0.24	15.9	-
>cluster_127					
WWTAAGGC	58	238	0.24	11.3	-
MNTTAMGGC	55	280	0.20	9.5	-
>cluster_128					
RYCACNNRNNRNCAG	61	325	0.19	11.3	-
>cluster_129					
RRAGTTGT	87	473	0.18	11.2	-
YKACANCTCSM	50	296	0.17	11	-
AAANWTGT	73	330	0.22	9.7	-
RGGAGTTRW	55	292	0.19	9.4	-
RNAAASYTGTNR	68	406	0.17	8.8	-
>cluster_130					
CCCNNGGGAR	177	1095	0.16	11.2	V$OLF1_01
>cluster_131					
GATAAGR	251	1379	0.18	11.2	V$GATA_C
WGATAAGR	170	973	0.17	10.3	V$GATA_C
>cluster_132					
TCCATTKW	57	221	0.26	11.1	-
RNAAAYNRAGGC	52	222	0.23	15.7	-
AAAYWGAGRY	44	188	0.23	14	-
TCCWTTGT	136	691	0.20	12	-
CTCYATTNW	62	262	0.24	11.1	-
>cluster_133					
RYTAAWNNNTGAY	54	242	0.22	11.1	-
>cluster_134					
CATRRAGC	61	257	0.24	11.1	-
GGCKCCATNW	54	194	0.28	14.1	-
GGCNCMATG	54	288	0.19	9.7	-
>cluster_135					
AGCYRWTTC	99	547	0.18	11.1	-
>cluster_136					
TAAYNRNNTCC	132	719	0.18	11	-
GAARKNGTTAR	56	194	0.29	14.3	-
RKCTGNNNNNRMTTA	59	238	0.25	12.8	-
YRTCTGNNNNNNATT	53	269	0.20	11.8	-
GAANSRRTTA	91	452	0.20	11.2	-
TAATKRNNNCCA	60	281	0.21	11.1	-
STAATNRNNNNCAG	49	191	0.26	11	-
AATNNNNNNCAGCNG	52	251	0.21	10.2	-
>cluster_137					
GAANYNYGACNY	54	263	0.21	11	-
>cluster_138					
MYAATNNNNNNNGGC	63	339	0.19	11	-
>cluster_139					
AAAYWAACM	54	206	0.26	11	V$HFH4_01
>cluster_140					
RNGTGGGC	363	2106	0.17	10.9	-
TGTGGGYR	199	1223	0.16	10.4	-
YRCGTGGG	88	397	0.22	9.1	-
>cluster_141					
TTCNRGNNNNTTC	59	335	0.18	10.9	V$HSF_Q6
>cluster_142					
ACAWYAAAG	85	373	0.23	10.9	-
>cluster_143					
CAGNWMCNNNGAC	51	256	0.20	10.8	-
YRGCAMNNNNGAC	58	302	0.19	10.5	-
SNACCNNRRACAR	52	256	0.20	10.4	-
>cluster_144					
AAANWWTGC	63	250	0.25	10.8	-
CGCYWWGTT	65	270	0.24	11.9	-
CGCNAWNTTT	52	210	0.25	11.5	-
>cluster_145					
YKACATTT	58	217	0.27	10.7	-
>cluster_146					
RRCCGTTA	59	230	0.26	10.5	-
>cluster_147					
YAATNANRNNNCAG	58	307	0.19	10.5	-
YYAATnAnnnnCCA	55	278	0.20	10.4	-
TGGYNNNNRATTNR	78	406	0.19	10.3	-
>cluster_148					
GATGKMRGCG	58	230	0.25	10.5	-
>cluster_149					
YGACNNYACAR	62	291	0.21	10.4	-
GGTKRNGTCA	57	271	0.21	9.5	-
YGACCNCAC	61	319	0.19	8.1	-

>cluster_150					
YTTCCNNNGGAMR	51	240	0.21	10.4	-
>cluster_151					
RYAAAKNNNNNNTTGW	71	435	0.16	10.4	-
CAAWGRNNNYTTT	57	322	0.18	9.2	-
RWTGGNNNNTTT	53	300	0.18	9.1	-
>cluster_152					
WCAANNNYCAG	85	514	0.17	10.3	-
TGGRRNTTGYR	66	366	0.18	10.3	-
>cluster_153					
CTGRYYYNATT	65	364	0.18	10.3	-
>cluster_154					
RNCTGNYNRNCTGNY	67	387	0.17	10.2	-
AGCSWRTCAS	58	225	0.26	10.8	-
TGAYWRRCTG	64	307	0.21	8.8	-
>cluster_155					
WGTTNNNNNAAA	88	467	0.19	10.2	-
TTTRYNYNRACA	56	321	0.17	9.9	-
>cluster_156					
YRCCAKNNGNCGC	51	296	0.17	10.2	-
>cluster_157					
KCCGNSWTTT	81	405	0.20	10.2	-
>cluster_158					
CCCNNNNNNAAGWT	49	272	0.18	10.2	-
>cluster_159					
GGCKCATGS	52	307	0.17	9.9	-
>cluster_160					
CAGNYGKNAAA	68	386	0.18	9.9	-
YAGCYRNYAGC	55	308	0.18	8.7	-
>cluster_161					
TTANWNANTGGM	61	335	0.18	9.8	-
SCATYRNNTAA	65	365	0.18	8.9	-
>cluster_162					
TAANNYSGCG	61	340	0.18	9.8	-
>cluster_163					
GGARNTKYCCA	50	269	0.19	9.8	-
>cluster_164					
GCGSCMNTTT	51	275	0.19	9.8	-
>cluster_165					
CCAWNWWNNNGGC	53	281	0.19	9.8	-
>cluster_166					
YNTTTNNNANGCARM	63	328	0.19	9.6	-
>cluster_167					
CCTNTMAGA	51	255	0.20	9.6	-
>cluster_168					
YTAAYNGCT	130	763	0.17	9.5	-
>cluster_169					
TTTNNANAGCYR	92	545	0.17	9.5	-
>cluster_170					
YNGTTNNNATT	71	366	0.19	9.1	-
>cluster_171					
CTCNANGTGNY	52	255	0.20	9.1	-
>cluster_172					
TTGCWCAAY	48	249	0.19	9	V$CEBPB_02
>cluster_173					
YWATTWNNRGCT	62	363	0.17	8.8	-
RWTTAYAGCY	55	265	0.21	8.8	-
ATTNWNAGC	70	392	0.18	8.6	-
>cluster_174					
WTGAAAT	61	295	0.21	8.1	-

Supplementary Table S4 **Motifs discovered in 3' UTR**

Motif	Conserved num	Total num	Conservation rate	MCS
>motif 1				
AATAAA	6617	14266	0.46	135.7
AAATAAA	2078	6041	0.34	66.6
CAATAAA	1111	2632	0.42	57.7
TAATAAA	1210	3510	0.34	50.9
GAATAAA	559	2148	0.26	26.4
CAAWWAAA	638	3120	0.20	21.1
TGAMYAAA	374	1789	0.21	19.5
GCAWTWAAA	127	571	0.22	12.4
>motif 2				
TATTTAT	1758	3706	0.47	79.4
TATTTAA	978	3158	0.31	41.6
CTATTTWW	622	2533	0.25	23.6
GTAAATAG	75	265	0.28	12.4
>motif 3				
TGTAnATA	1528	2968	0.51	70.4
TGTRnWTAT	744	2799	0.27	30.6
TGTRnnTWTAT	216	1032	0.21	15.0
TTGTRTATWT	112	411	0.27	13.2
CTGTnTnTAT	129	635	0.20	10.3
>motif 4				
TATTTTT	2068	6861	0.30	58.9
TATATTT	1033	3800	0.27	37.6
CTAKWTTT	402	1995	0.20	15.8
TATWTWTGA	170	804	0.21	14.0
WATATWTTTG	106	433	0.24	13.3
>motif 5				
TTTGTA	2777	8873	0.31	53.4
TTTnTAC	1623	6207	0.26	33.6
TTTnCTA	1378	6740	0.20	25.8
TTTCTA	1487	7235	0.21	24.3
TTTTGATA	163	589	0.28	18.0
KYTGATAAnR	110	545	0.20	8.9
>motif 6				
GTGCCTT	606	1266	0.48	46.9
AAGTGCCT	173	375	0.46	27.7
GGTGCCWW	140	659	0.21	14.0
>motif 7				
TTTTATA	1185	3861	0.31	45.4
ATTTTTAT	393	1742	0.23	23.0
CTTTTTAYR	119	575	0.21	9.5
>motif 8				
TGCATG	1234	3608	0.34	42.3
YYGCATGT	213	706	0.30	17.9
>motif 9				
TTTTGT	3185	13094	0.24	41.7
TTTnTGT	2490	12301	0.20	32.4
TATTYTTGTA	110	237	0.46	20.5
TTAnWYTTGTR	108	429	0.25	12.8
ATAnTYnTGTR	122	547	0.22	11.5
>motif 10				
WGCCTTA	669	1821	0.37	40.0
YGCCTTAA	171	381	0.45	25.1
TGCMnTAA	494	1855	0.27	23.5
TTGYMTTAA	112	514	0.22	10.9
TCGCCTTA	10	33	0.30	10.0
>motif 11				
GGTGCT	989	3190	0.31	38.6
GCTGCT	1157	5295	0.22	28.5
YGGTGCTA	147	278	0.53	24.8
TTGSTGCW	320	1444	0.22	18.3
ATGSTGCW	256	1164	0.22	17.7
GSTGCTAA	120	355	0.34	17.2
GSTGCTAT	115	366	0.31	15.0
>motif 12				
RCCAAAG	700	1951	0.36	37.5
ACCRWAGA	158	457	0.35	20.3
GCCRWAGA	111	425	0.26	13.6
TCCAAAGR	129	630	0.20	13.3
CCAAAGAT	80	289	0.28	12.6
CCAAAGAC	72	264	0.27	11.8
>motif 13				
TGTRnnTTT	1575	6878	0.23	36.9
WGTATTTW	1034	3981	0.26	33.2
TGTRnnnnTTT	1326	6418	0.21	31.5
TGTRnTTT	1373	6560	0.21	29.1
TGTAnATT	674	2325	0.29	26.7
TGTAnnnTTT	736	3201	0.23	24.4
TTGYAnnTTT	444	1834	0.24	22.2
WTTGYRnnnTTT	404	1976	0.20	20.8
TGTRnnTTA	703	3509	0.20	18.0
TGTnnATWTTT	246	1049	0.23	17.8
ATGTATWT	377	1794	0.21	15.9

TGTAnnTWnTTA	110	506	0.22	12.9
TGTRnYATTTW	121	586	0.21	11.8
TGTAnWGTT	209	944	0.22	11.4
TGTRYRnWTTA	124	599	0.21	10.5
TKTACAnnTTT	113	557	0.20	9.0
>motif 14				
GCACTTT	503	1337	0.38	35.4
TGCACTTT	209	522	0.40	27.5
TGCACTnW	607	2169	0.28	26.1
TGCRYYTTA	115	504	0.23	11.2
TGCACGTT	15	71	0.21	9.8
>motif 15				
TGTTTAC	518	1418	0.37	35.0
GTTTACAT	145	374	0.39	22.3
ACGTTTAC	13	49	0.27	10.5
CCGTTTAC	7	35	0.20	6.5
>motif 16				
TAATTTAT	331	868	0.38	33.3
TAATTTAA	195	812	0.24	17.2
CTAAnTTAW	144	668	0.22	12.9
YnTAAnTWAAG	116	514	0.23	12.3
TAAGTTAT	81	331	0.24	11.3
>motif 17				
TGTACAKW	712	2048	0.35	33.0
GTACAGA	185	797	0.23	13.4
GTACAGTT	73	282	0.26	11.3
ACGGTACA	7	18	0.39	9.7
>motif 18				
WGCAATA	664	1856	0.36	32.4
GTGCAATA	133	248	0.54	26.9
GTGCMATA	168	397	0.42	21.7
RTGCMnTAT	185	762	0.24	13.9
>motif 19				
TGTATAnW	1133	4075	0.28	31.2
CTGTATWW	380	1899	0.20	13.5
TnTGTATAAM	110	381	0.29	11.9
>motif 20				
TGTRnnnnTGT	1018	4650	0.22	31.1
TGTAnnnTTGY	188	820	0.23	13.0
>motif 21				
CTCAGGRA	231	772	0.30	30.6
>motif 22				
TGCCAAR	658	2450	0.27	30.0
TTGYMAAA	436	2159	0.20	17.1
TTTRCCAAR	112	500	0.22	12.9
>motif 23				
AGCMWTAA	348	1064	0.33	29.0
AAGCCATR	146	494	0.30	17.3
>motif 24				
TTGCACW	676	2203	0.31	28.9
TTTGCAY	826	2978	0.28	28.5
TTTGCACW	372	980	0.38	27.8
TTGYRCAA	237	948	0.25	13.6
TTGCACAA	74	273	0.27	11.9
>motif 25				
TGTGAA	1364	6081	0.22	27.4
ACTGTGA	393	1300	0.30	25.7
ACTGTGAA	173	477	0.36	23.2
AATGTGA	340	1613	0.21	16.1
ACTKYGAAY	116	443	0.26	13.6
ACTGKRAAT	119	502	0.24	12.3
>motif 26				
TATTAAA	665	2800	0.24	26.0
TGTAATWW	538	2345	0.23	21.1
WGTAWTAA	353	1746	0.20	15.4
>motif 27				
ACTKGAA	669	2690	0.25	25.8
TACTTGAA	160	368	0.43	25.5
ATACTTGA	96	269	0.36	17.1
>motif 28				
CTACCTCA	114	209	0.55	25.2
>motif 29				
TGTRnnATA	678	2727	0.25	25.0
TGTnnWRTAAA	235	986	0.24	18.8
TGTAnnRTAA	213	784	0.27	17.1
TGTAnnnTAG	156	760	0.21	10.6
>motif 30				
TATTTATTG	152	344	0.44	24.1
TATTTWnTGT	153	711	0.22	13.0
TATTTWnTGA	116	541	0.21	12.2
>motif 31				

WnTATWTTG	707	3180	0.22	23.4
TnTATnnTGT	507	2369	0.21	17.7
YnTATnnnnTGTA	196	911	0.22	14.1
ATAYWnTGTA	132	608	0.22	10.8
>motif 32				
AAGCACAA	139	335	0.41	23.0
AAGCACA	328	1145	0.29	22.3
TTGYRCTT	292	1296	0.23	15.6
TGGYRCTT	181	854	0.21	14.0
>motif 33				
WGTAWWTATT	229	742	0.31	22.8
TTGTRWWnATT	118	562	0.21	11.5
TGTAnnTnTTG	133	614	0.22	10.9
>motif 34				
TTTnnnnYGTA	685	3322	0.21	22.0
YTTGnAnTGTR	113	524	0.22	10.3
>motif 35				
GTACTGTA	123	293	0.42	21.8
TACTGTAT	101	471	0.21	11.0
GTACTGTG	48	220	0.22	7.8
CACGGTAC	6	26	0.23	6.6
>motif 36				
GTTTACAG	135	359	0.38	21.1
YnCTGTAA	503	2332	0.22	17.4
KTTTRYAGTnW	124	613	0.20	10.1
>motif 37				
TAATATAT	192	641	0.30	20.9
YGTAnTATRW	126	532	0.24	9.2
>motif 38				
WRCCAAAA	359	1481	0.24	20.9
WRCCAAAT	261	1232	0.21	16.4
AGCCAAA	219	1059	0.21	12.6
>motif 39				
TATWTTnnTAC	145	440	0.33	20.7
TATTTWnCTA	120	404	0.30	17.2
TATWTWnnTAG	143	610	0.23	14.7
>motif 40				
YGAATGTA	186	547	0.34	20.5
TGAATGY	556	2622	0.21	18.2
ACATTCC	251	962	0.26	17.7
ACATTCCA	109	343	0.32	16.6
>motif 41				
MAGTATT	688	3065	0.22	20.1
CAGTATTA	112	347	0.32	17.0
ACTACTG	155	646	0.24	12.7
ACTACTGW	109	411	0.27	11.3
>motif 42				
TTTKnnTAC	686	3330	0.21	19.9
WATTTWnTAC	132	582	0.23	13.0
TYGTAMnAAA	110	432	0.25	11.0
GTACCAAA	42	152	0.28	9.1
>motif 43				
WRTAAATG	550	2736	0.20	19.4
TGTRMATG	361	1788	0.20	17.4
TGTAAnTGT	113	542	0.21	8.5
>motif 44				
TGTAnnnTAT	421	1786	0.24	18.6
TGTRnnnWATT	436	2174	0.20	17.9
TTGYRnnnTATWY	116	516	0.22	9.9
>motif 45				
TGTAnnWWnTGTA	113	313	0.36	18.6
TGTnnnnnTTGTA	136	629	0.22	11.4
>motif 46				
TTGnAATAAA	126	411	0.31	18.2
CTTnnWATAAR	118	588	0.20	12.2
>motif 47				
CTATGCAA	83	199	0.42	17.8
TTTKYRTAG	162	781	0.21	10.1
>motif 48				
TTCnnWATAAA	127	511	0.25	17.0
TGTWnnnWATAWA	142	620	0.23	16.0
GTTnnnWRTAAA	142	659	0.22	14.9
>motif 49				
AAGYRYCTT	141	546	0.26	16.8
>motif 50				
ACACTAM	301	1160	0.26	16.7
AACACTAM	125	434	0.29	13.0
>motif 51				
GGACCAR	319	1565	0.20	16.7

>motif 52				
CTTWRTAA	325	1584	0.21	16.4
>motif 53				
CTATKYATT	130	491	0.26	16.1
>motif 54				
YGTAnAKRnTTT	112	353	0.32	15.7
>motif 55				
AACSRAAG	139	462	0.30	15.5
>motif 56				
YACCAGCA	136	536	0.25	15.5
>motif 57				
CTCRnTAAA	117	525	0.22	15.1
YCATTAAA	201	917	0.22	15.0
>motif 58				
YACTGCCR	155	714	0.22	14.9
>motif 59				
TnTATnTGTAnR	139	598	0.23	14.9
>motif 60				
TGCnnWRTAAA	122	513	0.24	14.8
>motif 61				
TGTRCCAW	220	988	0.22	14.7
>motif 62				
TGCKRCTA	128	460	0.28	14.6
>motif 63				
TGTnnnAWTAAA	128	608	0.21	14.6
>motif 64				
AATAWAnnTTG	110	489	0.22	14.5
>motif 65				
WCACYGTGM	104	406	0.26	14.5
ACACTGKR	230	1149	0.20	13.9
>motif 66				
WRTAAnnnnYGTAnW	108	437	0.25	14.3
TKTACRnnnnTTT	141	585	0.24	13.0
>motif 67				
GTTWTnTAT	240	1170	0.21	14.3
CTTWTnTAT	268	1326	0.20	13.2
>motif 68				
AWTAAAnnCTT	109	530	0.21	13.7
AWTAAAnnGTT	108	479	0.23	11.5
>motif 69				
TATTTWnATG	142	610	0.23	13.7
>motif 70				
ATAnTGTAnW	230	989	0.23	13.6
YGTAMAATA	114	445	0.26	10.9
YnTATTGTA	178	833	0.21	9.9
YnTACnRTATnY	110	514	0.21	9.1
>motif 71				
GACAATC	103	330	0.31	13.6
>motif 72				
TGCRMYAAA	115	434	0.26	13.4
RTTTRYTGC	125	594	0.21	10.7
>motif 73				
TTCnAnTAAA	117	553	0.21	13.3
TTTCTRnnAAA	116	570	0.20	12.7
>motif 74				
TGCSRAAA	138	508	0.27	13.3
TGCSRAAG	111	447	0.25	12.7
>motif 75				
TTTnnnRYCAAA	128	616	0.21	12.8
>motif 76				
TTGKAWTTAW	117	480	0.24	12.8
>motif 77				
AATRMAnTGT	165	823	0.20	12.8
YAATRnACTnK	119	571	0.21	9.9
>motif 78				
ATACGGGT	9	19	0.47	12.3
>motif 79				
YYGCACTA	116	400	0.29	12.1
TTGMRCTA	133	584	0.23	10.5
TTTKYRCTA	136	655	0.21	9.8

>motif 80				
ATGYACTKY	147	625	0.24	12.0
ATGTACWG	134	586	0.23	7.7
>motif 81				
TTTCAATA	105	464	0.23	11.9
GTTnYAATA	106	513	0.21	8.4
>motif 82				
TCTRTRnATA	124	531	0.23	11.9
TCTRTWTAT	135	653	0.21	11.0
>motif 83				
AATMWAGTT	117	577	0.20	11.9
WATAAMGTT	108	499	0.22	10.1
>motif 84				
TGTRYMAATR	113	482	0.23	11.8
GTGYnAATW	187	911	0.21	11.0
>motif 85				
YAATRWAGC	106	488	0.22	11.7
ATGTAGCA	54	224	0.24	9.1
TGCTGCAT	69	328	0.21	8.9
>motif 86				
TTGTKKACA	112	457	0.25	11.7
TGTTKMCAA	110	479	0.23	11.1
>motif 87				
AGAnTATTWW	127	632	0.20	11.5
>motif 88				
AGAKnTnTATW	120	582	0.21	11.2
>motif 89				
GTGCnATT	156	676	0.23	10.8
>motif 90				
WKTACWnKAAA	116	580	0.20	10.6
WTTTnTKGTAM	113	532	0.21	8.8
>motif 91				
TGTWnAnAGC	115	572	0.20	10.3
TGTAnAnAGA	115	502	0.23	9.8
>motif 92				
YRAAGYnTTA	123	606	0.20	9.9
>motif 93				
YYGTAnnnnKATT	108	514	0.21	9.7
>motif 94				
YACARTnTTT	120	590	0.20	9.5
>motif 95				
GTTGTAnA	191	927	0.21	9.5
>motif 96				
GGTACGAA	8	25	0.32	9.3
>motif 97				
ATAYGCAR	114	555	0.21	9.2
>motif 98				
TATTKnnnnGTAnW	110	545	0.20	9.2
>motif 99				
TCGCATGA	6	17	0.35	8.5
TCGCATGG	5	19	0.26	6.5
>motif 100				
CTTRYRnATA	111	513	0.22	8.4
CTTKYGTAW	128	621	0.21	8.3
>motif 101				
GTCAATAA	49	214	0.23	8.2
>motif 102				
TAACGGGT	5	14	0.36	7.8
>motif 103				
TRTAAnTAC	116	574	0.20	7.5
>motif 104				
CGCAAAAA	6	23	0.26	7.1
>motif 105				
AAGGGCTA	33	138	0.24	7.1
>motif 106				
GGCAGCTA	33	142	0.23	6.9

Supplementary Table S5 **Conserved 8-mer motifs discovered in 3' UTR**

Motif	Conserved num	Total num	Conservation rate	MCS	matched miRNA
>cluster_1	GTGCAATA				
GTGCAATA	160	291	0.55	30.6	miR-92 miR-32 MIR200 MIR256
AAGCAATA	145	386	0.38	22.3	miR-137
TGTGCAAT	134	377	0.36	20.6	MIR200
AGTGCAAT	90	273	0.33	15.9	miR-367 miR-25 MIR1 MIR228 MIR252 MIR256
ATGCAATA	85	294	0.29	13.8	miR-217(#3)
GGTGCAAT	33	116	0.28	8.5	
AGCAATAG	49	183	0.27	9.8	
AAGTGCAA	81	306	0.27	12.5	
TGCAATAT	90	344	0.26	13.1	
GTGCAATT	62	240	0.26	10.7	MIR173 MIR228
TAGCAATA	60	234	0.26	10.5	MIR189 MIR210
GAGCAATA	41	163	0.25	8.5	
CAGCAATA	66	262	0.25	10.8	
GTGCAATG	46	183	0.25	9.0	
TTGCAATA	80	321	0.25	11.8	MIR45 MIR166 MIR216
AGCAATAT	65	263	0.25	10.5	
GTGCAATC	22	91	0.24	6.0	
TATGCAAT	70	296	0.24	10.5	
TTGTGCAA	77	329	0.23	10.9	MIR200
TGCAATAG	40	171	0.23	7.8	MIR216
TGCAATAC	38	164	0.23	7.6	
GCAATATT	65	291	0.22	9.5	
AGCAATAC	33	154	0.21	6.5	
GCAATACT	31	146	0.21	6.2	
TGGCAATA	44	235	0.19	6.5	MIR150
>cluster_2	GTGCCTTA				
GTGCCTTA	147	271	0.54	29.1	miR-124a
AGTGCCTT	227	473	0.48	33.3	
AAGTGCCT	196	430	0.46	29.8	MIR63
TGCCTTAA	186	409	0.46	29.0	
AGCCTTAA	113	295	0.38	20.0	
GTGCCTTG	133	355	0.38	21.3	miR-224(#3)
GTGCCTTT	179	482	0.37	24.6	
TGTGCCTT	228	620	0.37	27.5	MIR157 MIR182
CGCCTTAA	14	38	0.37	13.5	
GCCTTAAT	84	238	0.35	16.2	
ATGCCTTA	90	258	0.35	16.6	
GGCCTTAA	52	160	0.33	11.9	
GCCTTAAA	102	314	0.33	16.7	
GCCTTAAG	56	183	0.31	11.8	
TTGCCTTA	122	401	0.30	17.3	
AAGCCTTA	79	264	0.30	13.7	
GTGCCTTC	107	361	0.30	15.9	MIR157
GGTGCCTT	82	281	0.29	13.7	
GCCTTAAC	48	165	0.29	10.5	
CTGCCTTA	108	371	0.29	15.7	
TGCCTTAC	63	220	0.29	11.8	
TAGTGCCT	59	209	0.28	11.3	
TCGCCTTA	10	37	0.27	9.5	
TGCCTTAT	99	384	0.26	13.5	
CGCCTTAT	6	24	0.25	7.0	miR-208(#3)
GAGCCTTA	43	174	0.25	8.6	
TGCCTTAG	70	288	0.24	10.7	
TGGCCTTA	59	250	0.24	9.6	
GCCTTATT	65	293	0.22	9.5	
AGTGCCTA	41	185	0.22	7.5	miR-34b(#3)
AGTGCCTG	95	434	0.22	11.3	
GCCTTACT	36	172	0.21	6.6	MIR123
GTTGCCTT	66	319	0.21	8.9	
ACGCCTTA	6	29	0.21	6.2	
TAGCCTTA	33	162	0.20	6.2	miR-9*(#3)
TTGTGCCT	93	464	0.20	10.2	
AGTGCCTC	54	270	0.20	7.7	MIR63 MIR182
CAGTGCCT	102	512	0.20	10.6	miR-34c(#3)
AATGCCTT	90	452	0.20	9.9	MIR240
TGCCTTGC	62	336	0.19	7.6	miR-330(#3)
ATGTGCCT	64	353	0.18	7.5	
TGCCTTGA	63	350	0.18	7.4	MIR77
>cluster_3	CTACCTCA				
				27.8	miR-98 let-7i let-7g let-7f let-7e let-7c let-7b let-7a
CTACCTCA	139	263	0.53		
ATACCTCA	97	240	0.40	19.3	MIR207
ACTACCTC	63	160	0.39	15.2	let-7d
TCTACCTC	97	256	0.38	18.4	
TTACCTCA	99	262	0.38	18.5	MIR250
TACCTCAG	96	283	0.34	16.8	
CTACCTCT	96	284	0.34	16.7	MIR26
CCTACCTC	88	276	0.32	15.3	MIR26
TACCTCAA	59	186	0.32	12.5	MIR250
GCTACCTC	48	152	0.32	11.2	
AATACCTC	55	176	0.31	11.9	MIR8
TACCTCAC	58	187	0.31	12.1	
ATTACCTC	40	156	0.26	8.5	
TACCTCAT	58	239	0.24	9.8	
CTACCTCC	54	243	0.22	8.6	
AACTACCT	43	194	0.22	7.7	miR-196b miR-196a
TTTACCTC	65	298	0.22	9.3	
TATACCTC	35	161	0.22	6.8	
TACCTCTT	61	295	0.21	8.5	
CTACCTCG	10	50	0.20	7.9	

TACCTCGG	7	36	0.19	6.5	
>cluster_4	ACCAAAGA				
ACCAAAGA	178	366	0.49	29.7	miR-9 MIR170 MIR188
AACCAAAG	161	412	0.39	24.2	MIR188
TACCAAAG	98	259	0.38	18.4	MIR134 MIR170
GCCAAAGA	112	299	0.38	19.6	MIR221
GACCAAAG	97	270	0.36	17.6	
GGACCAAA	82	256	0.32	14.8	
TGCCAAAG	125	393	0.32	18.2	MIR164
CCAAAGAT	100	337	0.30	15.4	
ACCAAAGT	85	306	0.28	13.4	
TGACCAAA	97	356	0.27	14.1	
ACCAAAGC	74	273	0.27	12.2	
GTACCAAA	50	187	0.27	9.9	MIR134
CCAAAGAC	79	302	0.26	12.2	
CCAAAGAA	116	443	0.26	14.8	MIR170
TCCAAAGA	111	427	0.26	14.4	
GACCAAAA	77	297	0.26	12.0	MIR122
ACCAAAGG	74	290	0.26	11.6	MIR56
AGACCAAA	82	327	0.25	12.0	
CACCAAAG	72	296	0.24	10.9	MIR56
AGCCAAAG	84	354	0.24	11.5	MIR221
TAACCAAA	74	325	0.23	10.4	
GAACCAAA	71	315	0.23	10.1	MIR140
ACCAAATA	75	337	0.22	10.2	
GACCAAAT	56	263	0.21	8.4	
AAACCAAA	155	726	0.21	14.1	MIR143
GCCAAAGT	51	241	0.21	8.0	
CCAAAGAG	81	390	0.21	9.9	
ACCAAAAA	109	524	0.21	11.5	
GCCAAAGC	47	237	0.20	7.1	
CCAAAGTT	65	330	0.20	8.3	
TTACCAAA	77	393	0.20	9.0	MIR228
ATACCAAA	56	290	0.19	7.6	MIR170 MIR245
AACCAAAT	74	390	0.19	8.5	
AACCAAAA	100	544	0.18	9.5	
>cluster_5	TGTTTACA				
TGTTTACA	369	771	0.48	42.3	miR-30e-5p miR-30d miR-30c miR-30b miR-30a-5p MIR183 MIR257
GTTTACAT	167	441	0.38	24.1	
GTTTACAG	146	406	0.36	21.6	
GTTTACAA	134	400	0.34	19.6	
TTGTTTAC	160	501	0.32	20.6	MIR225
ATGTTTAC	147	465	0.32	19.6	
AGTTTACA	112	382	0.29	16.1	
GTTTACAC	45	179	0.25	8.9	
CGTTTACA	18	74	0.24	11.9	
AAGTTTAC	70	292	0.24	10.6	
GGTTTACA	46	198	0.23	8.3	MIR214
ACGTTTAC	13	56	0.23	9.9	
CTGTTTAC	87	383	0.23	11.2	MIR183
CCGTTTAC	8	42	0.19	6.8	
TGTTTACT	102	544	0.19	9.9	MIR225
CAGTTTAC	41	221	0.19	6.2	
GTGTTTAC	53	289	0.18	6.9	MIR257
>cluster_6	GCACTTTA				
GCACTTTA	193	405	0.48	30.5	miR-20 miR-106b miR-18(#2)
TGCACTTT	244	596	0.41	30.8	
AGCACTTT	190	601	0.32	22.3	miR-93 miR-372 miR-17-5p miR-106a MIR103
TTGCACTT	145	464	0.31	19.3	
GCACTTTG	119	398	0.30	16.9	MIR103
ATGCACTT	82	302	0.27	12.9	
GCACTTTT	118	440	0.27	15.3	
AAGCACTT	123	483	0.26	14.9	miR-302d miR-302c miR-302b miR-302a miR-373
TGCACTTA	59	248	0.24	9.7	
GGCACTTT	64	276	0.23	9.8	
CTGCACTT	80	374	0.21	10.1	MIR209
CACTTTAT	84	413	0.20	9.8	
GTGCACTT	47	237	0.20	7.1	MIR153 MIR186
GCACTTTC	57	290	0.20	7.8	
CACTTTAA	77	418	0.18	8.4	
>cluster_7	TGGTGCTA				
TGGTGCTA	116	272	0.43	21.9	miR-29c miR-29b miR-29a MIR196
GGTGCTAA	74	180	0.41	17.0	
GGTGCTAT	59	167	0.35	13.6	
AGGTGCTA	63	192	0.33	13.2	
ATGGTGCT	108	343	0.32	16.8	miR-107(#3) miR-103(#3) MIR139
GGTGCTAG	44	142	0.31	10.6	
GGTGCTAC	31	108	0.29	8.3	
TTGGTGCT	114	408	0.28	15.6	
TGGTGCTT	112	417	0.27	14.9	
AAGGTGCT	76	299	0.25	11.7	
GGTGCTTT	85	340	0.25	12.2	
TGGTGCTC	66	266	0.25	10.6	
CGGTGCTA	9	37	0.24	8.4	
AATGGTGC	57	235	0.24	9.7	
CTGGTGCT	107	449	0.24	13.0	MIR196 MIR198
TGGTGCTG	118	525	0.23	12.9	MIR24 MIR198
GTGCTATT	48	238	0.20	7.4	
TTTGGTGC	62	317	0.20	8.1	
GTGCTAAA	47	240	0.20	7.0	MIR194

ATTGGTGC	29	148	0.20	5.5	MIR219
GGTGCTGA	56	287	0.20	7.7	MIR198
GGTGCTTG	39	208	0.19	6.1	
GGTGCTGT	66	354	0.19	7.9	
>cluster_8	CTATGCAA				
CTATGCAA	97	231	0.42	19.8	miR-153 MIR246
ACTATGCA	57	183	0.31	12.1	
TCTATGCA	63	223	0.28	11.7	MIR246
CTATGCAT	44	169	0.26	9.1	
TATGCAAA	112	446	0.25	14.0	MIR239 MIR242 MIR248 MIR255
GCTATGCA	36	146	0.25	7.8	MIR224 MIR226
ATATGCAA	84	344	0.24	11.8	MIR41 MIR239 MIR248 MIR255
TTATGCAA	67	309	0.22	9.4	MIR242
TATGCATG	50	238	0.21	7.8	MIR74
TATGCACT	46	220	0.21	7.5	
TTTATGCA	85	432	0.20	9.5	
>cluster_9	TACTTGAA				
TACTTGAA	178	425	0.42	26.8	miR-26b miR-26a MIR243
ATACTTGA	105	316	0.33	17.3	
ACTTGAAT	134	461	0.29	17.5	
TTACTTGA	82	315	0.26	12.4	MIR243
AACTTGAA	131	536	0.24	14.8	MIR104 MIR249
ACTTGAAC	53	222	0.24	9.2	MIR131
ACTTGAAA	124	614	0.20	11.8	
AAACTTGA	106	526	0.20	10.9	
GTACTTGA	39	196	0.20	6.5	
CACTTGAA	64	338	0.19	7.9	MIR131
>cluster_10	CGCAAAAA				
CGCAAAAA	17	41	0.42	16.0	MIR161 MIR178
GCCAAAAA	82	362	0.23	10.9	
>cluster_11	GTGCCAAA				
GTGCCAAA	118	291	0.41	21.3	miR-96 MIR217
TTGCCAAA	172	521	0.33	22.0	miR-182 MIR48 MIR253
AGTGCCAA	84	255	0.33	15.3	MIR217
TGTGCCAA	109	351	0.31	16.6	MIR108 MIR206
TGCCAAAA	148	480	0.31	19.3	MIR205
AAGTGCCA	92	305	0.30	14.9	MIR196
ATGCCAAA	103	351	0.29	15.4	MIR204
GTGCCAAT	40	142	0.28	9.3	MIR108 MIR206
TGCCAAAT	110	411	0.27	14.7	MIR19 MIR48 MIR79 MIR199
GTGCCATA	41	170	0.24	8.2	miR-183 MIR87
TGCCAAAC	62	262	0.24	9.9	
CTGCCAAA	88	384	0.23	11.4	MIR19 MIR79 MIR164
GTGCCAAG	62	273	0.23	9.5	
GGTGCCAA	42	189	0.22	7.6	
GTGCCATT	61	277	0.22	9.1	MIR3
TTTGCCAA	109	506	0.22	11.9	
GCCAAATA	53	254	0.21	8.0	MIR48
ATTGCCAA	60	301	0.20	8.1	MIR48 MIR237 MIR253
AAGCCAAA	97	489	0.20	10.3	MIR131
GCCAAACT	39	202	0.19	6.3	
AATGCCAA	68	353	0.19	8.3	
GTTGCCAA	43	225	0.19	6.6	
TGCCAATA	40	214	0.19	6.2	
GCCAAATT	49	263	0.19	6.8	
>cluster_12	GTACTGTA				
GTACTGTA	136	338	0.40	22.7	miR-101
TACTGTGA	99	354	0.28	14.5	MIR233
TACTGTAA	126	461	0.27	16.1	
TGTACTGT	129	506	0.26	15.3	
CTACTGTA	69	283	0.24	10.7	miR-199a*
TACTGTAC	63	274	0.23	9.7	MIR184
ACACTGTA	64	289	0.22	9.4	
ACTACTGT	59	268	0.22	9.0	
AGTACTGT	58	267	0.22	8.8	
ACTGTAAA	122	572	0.21	12.5	
GGTACTGT	36	172	0.21	6.6	
GTACTGTG	55	265	0.21	8.1	
TACTGTAT	108	530	0.20	11.2	
ACTGTACA	72	359	0.20	9.0	
ACTGTATA	81	416	0.20	9.2	
ATACTGTA	80	420	0.19	8.9	miR-144
TACTGTAG	47	254	0.19	6.6	
TTGTACTG	69	375	0.18	7.9	
>cluster_13	ATACGGGT				
ATACGGGT	10	25	0.40	11.9	
TACGGGTT	8	21	0.38	10.4	miR-99a miR-100 miR-99b(#3)
TACGGGTA	7	19	0.37	9.5	
TTACGGGT	9	29	0.31	9.8	
ACGGGTTT	8	44	0.18	6.6	
>cluster_14	AAGCACAA				
AAGCACAA	157	393	0.40	24.3	miR-218 MIR113 MIR197
TTGCACAA	81	315	0.26	12.2	MIR200
AAAGCACA	134	521	0.26	15.7	MIR148 MIR197
AGCACAAT	62	244	0.25	10.5	

TAGCACAA	44	194	0.23	8.0	
AGCACAAA	80	371	0.22	10.2	MIR113 MIR203
AGCACAAC	34	162	0.21	6.5	MIR166 MIR210
GAAGCACA	63	305	0.21	8.7	
CAAGCACA	57	291	0.20	7.8	MIR113 MIR165
TAAGCACA	41	224	0.18	6.1	
AGCACAAG	48	265	0.18	6.5	
>cluster_15	TTTGCACT				
TTTGCACT	209	559	0.37	26.7	
TTGCACTA	89	244	0.37	17.1	
TTTGCACA	207	583	0.36	25.5	miR-19b miR-19a
TTTGCATG	158	456	0.35	21.9	MIR108
TTTTGCAC	167	513	0.33	21.4	MIR208
TTGCACTG	110	399	0.28	15.1	miR-301 miR-130b miR-130a MIR185 MIR228
GTTTGCAC	66	248	0.27	11.3	
ATTTGCAC	99	375	0.26	13.8	
TTGCACGA	9	35	0.26	8.7	
TGCACTGA	103	403	0.26	13.7	miR-152 miR-148b miR-148a MIR238
TTGCACAT	93	371	0.25	12.8	
TGCACTAA	53	213	0.25	9.6	MIR5
TGCACTAC	29	122	0.24	6.8	
TTTGCACG	19	81	0.24	12.0	
TGCACTAT	44	195	0.23	7.9	
ATTGCACT	57	262	0.22	8.7	MIR24 MIR71 MIR228
TTGCACGT	13	60	0.22	9.4	
TGCACTGT	90	418	0.22	10.8	miR-139 MIR228
GTTGCACT	38	194	0.20	6.3	
AATGCACT	56	285	0.20	7.7	
ATGCACTG	56	301	0.19	7.2	MIR226
AATTGCAC	42	228	0.18	6.2	MIR201
TTGCACAG	63	347	0.18	7.5	
>cluster_16	TGTACATA				
TGTACATA	276	762	0.36	29.9	
TGTACAGA	178	523	0.34	22.9	
TGTACAGT	176	524	0.34	22.5	
TTGTACAG	156	483	0.32	20.6	
TTTGTACA	236	771	0.31	24.2	
TGTACATT	162	590	0.28	18.3	
ATGTACAG	93	363	0.26	13.0	
TTGTACAT	160	641	0.25	16.7	
CTGTACAG	90	371	0.24	12.2	
TTGTACAA	99	430	0.23	12.1	
GTTGTACA	53	230	0.23	8.9	
CTTGTACA	64	293	0.22	9.2	
TGTACAAA	105	484	0.22	11.8	
GTACATAA	60	279	0.22	8.8	
CTGTACAT	94	437	0.22	11.0	
AATGTACA	101	493	0.21	10.9	
GTACATTT	107	525	0.20	11.1	
GTGTACAG	48	237	0.20	7.4	
GTACATTA	40	199	0.20	6.7	
GTACATAG	46	232	0.20	7.1	MIR167
TATGTACA	82	422	0.19	9.2	
ATGTACAT	101	528	0.19	10.1	
TGTACAAT	67	364	0.18	7.8	
GTACATAT	66	362	0.18	7.7	
TCTGTACA	76	419	0.18	8.2	
>cluster_17	AAGCCATA				
AAGCCATA	92	261	0.35	16.9	miR-135b miR-135a
AAAGCCAT	108	438	0.25	13.5	
AAGCCATG	72	321	0.22	10.1	
GAAGCCAT	67	334	0.20	8.6	
AGCCATAA	35	177	0.20	6.1	
>cluster_18	ACTGTGAA				
ACTGTGAA	192	551	0.35	24.2	miR-27b miR-27a MIR192
CACTGTGA	132	414	0.32	18.7	miR-128b miR-128a MIR192
TGTGAATA	128	498	0.26	15.3	
AACTGTGA	111	450	0.25	13.7	
AATGTGAA	207	872	0.24	18.1	miR-23b(#1) miR-23a(#1)
GCTGTGAA	91	411	0.22	11.2	MIR177
CTGTGAAT	107	484	0.22	12.1	
ACTGTGAT	77	370	0.21	9.6	MIR233
ACACTGTG	84	407	0.21	10.0	MIR21
TTGTGAAT	122	626	0.20	11.2	
TCACTGTG	101	522	0.19	10.2	
CTGTGAAA	122	643	0.19	10.9	
TCTGTGAA	114	603	0.19	10.5	
ACTGTGAC	52	285	0.18	6.8	MIR247
>cluster_19	AGACAATC				
AGACAATC	44	134	0.33	11.1	
TGACAATC	41	130	0.32	10.3	MIR149 MIR218
GACAATCA	46	155	0.30	10.4	miR-219 MIR149
GTACAATC	20	78	0.26	6.0	
GACAATCT	31	121	0.26	7.5	
TACAATCA	33	155	0.21	6.5	
ACAATCAT	40	207	0.19	6.4	
>cluster_20	TGCTGCTA				

TGCTGCTA	124	380	0.33	18.5	miR-195 miR-16 miR-15b miR-15a MIR117
GCTGCTAA	60	223	0.27	10.9	
GCTGCTAT	69	262	0.26	11.5	MIR56
ATGCTGCA	75	332	0.23	10.4	miR-338(#3)
TTGCTGCT	145	664	0.22	13.9	miR-424 MIR116 MIR117
TGCTGCAT	76	357	0.21	9.8	
AAGCTGCT	130	621	0.21	12.5	MIR129
TTTGCTGC	104	522	0.20	10.7	MIR116
GCTGCTAG	34	182	0.19	5.7	MIR28
AGCTGCTA	44	241	0.18	6.3	
>cluster_21	TTTTGTAC				
TTTTGTAC	244	753	0.32	25.8	
CTTTTGTA	178	748	0.24	16.8	MIR156
TTTTGTAG	147	641	0.23	14.7	
GTTTTGTA	154	676	0.23	15.0	
TTTGTACT	134	614	0.22	13.4	
TTTGTATG	133	627	0.21	12.9	
CTTTGTAC	78	370	0.21	9.8	
CTTTGTAA	146	693	0.21	13.5	
TTTGTAAC	103	492	0.21	11.2	
TTTGTACC	66	324	0.20	8.7	
ACTTTGTA	119	582	0.20	11.8	
ATTTGTAC	86	426	0.20	9.9	
TTTGTAGC	74	368	0.20	9.1	
TTTTGTGC	103	524	0.20	10.5	
TTTGTAAG	108	559	0.19	10.5	
TTTGTACG	14	73	0.19	9.1	
TCTTTGTA	122	651	0.19	10.8	
CATTTGTA	111	614	0.18	9.9	
CTTTGTAT	112	623	0.18	9.8	
>cluster_22	ACATTCCA				
ACATTCCA	127	402	0.32	18.2	miR-206 miR-1 miR-122a(#2)
AACATTCC	80	315	0.25	12.0	
TACATTCC	58	229	0.25	10.2	
ACATTCCT	85	405	0.21	10.2	
>cluster_23	TGAATGTA				
TGAATGTA	183	594	0.31	21.4	miR-181b(#1)
GAATGTAT	107	461	0.23	12.7	
TTGAATGT	123	586	0.21	12.3	miR-181c miR-181a
GAATGTAG	47	228	0.21	7.5	
GAATGTAA	98	476	0.21	10.8	
GAATGTAC	40	214	0.19	6.2	
AGAATGTA	88	479	0.18	9.0	
>cluster_24	ACGGTACA				
ACGGTACA	7	23	0.30	8.5	
CGGTACAG	7	24	0.29	8.3	
CACGGTAC	6	27	0.22	6.5	
>cluster_25	CAGTATTA				
CAGTATTA	121	400	0.30	17.2	miR-200c miR-200b MIR115
ACAGTATT	142	551	0.26	16.2	
CAGTATTT	179	811	0.22	15.6	
TCAGTATT	124	577	0.22	12.7	
CAGTATTG	61	301	0.20	8.3	
>cluster_26	TTGCATGT				
TTGCATGT	112	383	0.29	16.1	
TGCATGCT	78	269	0.29	13.3	MIR105
TTGCATGC	44	170	0.26	9.0	
TGCATGTT	118	456	0.26	14.8	
TTGCATGA	72	279	0.26	11.5	
AATGCATG	83	330	0.25	12.1	MIR236
GTGCATGC	49	196	0.25	9.2	
CTGCATGC	67	270	0.25	10.7	MIR152
TGCATGTC	60	245	0.25	10.0	
GTTGCATG	45	184	0.25	8.7	
TGCATGTA	80	340	0.24	11.1	
TGCATGAG	50	217	0.23	8.6	
ATGCATGC	45	196	0.23	8.2	MIR74 MIR105
GCATGTTT	96	419	0.23	11.9	
TGCATGAA	75	332	0.23	10.4	
GTGCATGA	38	170	0.22	7.3	MIR212
TGCATGCC	47	211	0.22	8.1	
AGTGCATG	47	211	0.22	8.1	MIR212
GCTGCATG	58	265	0.22	8.8	
TAGCATGT	50	230	0.22	8.1	
CTGCATGT	77	361	0.21	9.9	
TGTGCATG	99	468	0.21	11.1	
ATTGCATG	46	217	0.21	7.6	
TGCATGCA	64	306	0.21	8.8	MIR74 MIR193 MIR236
TTGCATGG	52	255	0.20	7.8	
TGCATGGA	52	256	0.20	7.7	
ATGCATGT	82	413	0.20	9.5	
TCTGCATG	71	361	0.20	8.7	
GATGCATG	39	200	0.20	6.4	MIR251
ATGCATGA	43	220	0.20	6.7	
GCATGTAA	43	224	0.19	6.6	
CTGCATGA	53	276	0.19	7.3	
CTTGCATG	44	234	0.19	6.5	

TGCATGTG	90	482	0.19	9.2	
TGCATGAT	44	235	0.19	6.5	
>cluster_27	TCGCATGA				
TCGCATGA	6	21	0.29	7.6	
TCGCATGG	6	24	0.25	7.0	MIR232
CTCGCATG	9	38	0.24	8.3	
TCGCATGC	6	27	0.22	6.5	
TAGCATGA	37	181	0.20	6.6	
TCCGCATG	8	43	0.19	6.7	
CGCATGCC	7	38	0.18	6.2	
>cluster_28	CTCAGGGA				
CTCAGGGA	129	451	0.29	16.9	miR-125b miR-125a
CTCAGGAA	118	434	0.27	15.5	MIR230
TCTCAGGG	96	382	0.25	13.0	
CTCAGGTA	39	156	0.25	8.2	
AACTCAGG	57	241	0.24	9.4	MIR94
TTCTCAGG	96	418	0.23	11.9	
ACTCAGGA	66	298	0.22	9.5	
ATCTCAGG	61	287	0.21	8.8	MIR94
ACTCAGGT	37	175	0.21	6.8	
ACTCAGGG	53	255	0.21	8.0	
TCTCAGGT	51	248	0.21	7.7	MIR107 MIR190
TCAGGGAA	90	436	0.21	10.3	
TTCAGGGA	81	395	0.21	9.7	
TCTCAGGA	79	402	0.20	9.2	
TTTCAGGG	80	417	0.19	9.0	MIR136 MIR138
TCAGGGAT	47	245	0.19	6.9	
CTCAGGTT	46	252	0.18	6.4	
>cluster_29	CAAGTGCC				
CAAGTGCC	80	285	0.28	13.1	MIR196
AAAGTGCC	69	278	0.25	10.9	
TAAGTGCC	45	192	0.23	8.3	
GAAGTGCC	45	210	0.21	7.6	MIR23
AAGTGCAT	56	267	0.21	8.3	
TCAAGTGC	36	184	0.20	6.2	
TAAAGTGC	53	293	0.18	6.8	
>cluster_30	ACTACTGA				
ACTACTGA	58	207	0.28	11.1	
TACTACTG	46	203	0.23	8.1	
AACTACTG	52	252	0.21	7.9	
>cluster_31	TGGACCAA				
TGGACCAA	67	240	0.28	11.9	MIR32
GTGACCAA	45	215	0.21	7.4	MIR97 MIR247
GGGACCAA	44	214	0.21	7.2	miR-133b miR-133a
>cluster_32	GTAAATAG				
GTAAATAG	88	317	0.28	13.6	
CTGTAAAT	180	671	0.27	18.9	
GTGTAAAT	118	491	0.24	13.8	MIR191
TGTAAATG	176	764	0.23	16.2	MIR191
GTAAATAC	78	348	0.22	10.5	
>cluster_33	TGTAGATA				
TGTAGATA	81	295	0.28	12.9	
>cluster_34	ACACTACA				
ACACTACA	54	199	0.27	10.4	miR-142-3p MIR95
AACACTAA	90	334	0.27	13.4	
AACACTAC	39	160	0.24	8.0	
TAACACTA	46	198	0.23	8.3	
ACACTAAT	39	205	0.19	6.2	MIR211 MIR256
>cluster_35	GTACAGTT				
GTACAGTT	81	313	0.26	12.3	
GTACAGAA	68	334	0.20	8.8	
GTACAGAT	44	223	0.20	6.9	
GTACAGTA	55	285	0.19	7.5	
GTACAGTG	43	229	0.19	6.4	
GTACAGAG	39	214	0.18	5.9	MIR65
>cluster_36	CACCAGCA				
CACCAGCA	104	405	0.26	13.8	miR-138(#1) MIR10 MIR25 MIR102 MIR213
ACCAGCAT	51	234	0.22	8.2	
TCACCAGC	62	318	0.20	8.1	MIR25 MIR102
>cluster_37	GGTACGAA				
GGTACGAA	7	28	0.25	7.6	
TGGTACGA	6	25	0.24	6.8	miR-126(#2)
>cluster_38	TGTATAGT				
TGTATAGT	77	315	0.24	11.3	
CTGTATAT	127	536	0.24	14.1	

TTGTATAG	70	316	0.22	9.8	
TGTATAGA	72	326	0.22	9.9	
TCTGTATA	91	465	0.20	9.8	
GTGTATAT	113	598	0.19	10.5	
GTTGTATA	52	276	0.19	7.1	
TGTGTATA	144	772	0.19	11.7	
CTTGTATA	62	336	0.19	7.6	miR-381
GTATAGTT	39	213	0.18	5.9	
>cluster_39	AAGGGCTA				
AAGGGCTA	39	164	0.24	7.9	MIR42
>cluster_40	AGCTTTAA				
AGCTTTAA	97	412	0.24	12.3	MIR63
AAGCTTTA	74	378	0.20	8.8	
GCTTTAAT	58	308	0.19	7.5	
>cluster_41	ATTTATCG				
ATTTATCG	9	39	0.23	8.2	
>cluster_42	GGCAGCTA				
GGCAGCTA	39	172	0.23	7.5	miR-22(#1) MIR164
>cluster_43	GCTGTAAA				
GCTGTAAA	68	300	0.23	9.9	
TGCTGTAA	80	389	0.21	9.7	MIR58 MIR197
CTTGTAAA	105	538	0.20	10.5	MIR177
GTTGTAAA	84	440	0.19	9.1	
TCTGTAAA	131	707	0.19	11.0	MIR173
TTGTAAAG	99	542	0.18	9.4	
>cluster_44	GCACTAAT				
GCACTAAT	26	120	0.22	5.8	
>cluster_45	AAAGGTGC				
AAAGGTGC	46	212	0.22	7.8	
>cluster_46	ATGTAGCA				
ATGTAGCA	57	265	0.22	8.6	miR-221(#1) miR-222(#1)
>cluster_47	ACACTGGA				
ACACTGGA	78	365	0.21	10.0	miR-199b(#1) miR-199a(#1) MIR227
AACTGGAA	101	501	0.20	10.7	miR-145(#1) MIR220
TAACTGGA	45	247	0.18	6.3	
>cluster_48	GTATATAG				
GTATATAG	54	253	0.21	8.3	
>cluster_49	TTTGATAA				
TTTGATAA	116	551	0.21	12.0	miR-361(#2)
TTTGATAC	56	299	0.19	7.3	
>cluster_50	AAGCATGC				
AAGCATGC	35	166	0.21	6.6	
TTAGCATG	43	211	0.20	7.0	
AAAGCATG	76	380	0.20	9.2	
GCATGCTT	45	229	0.20	6.9	MIR105
>cluster_51	TGCACGAT				
TGCACGAT	7	34	0.21	6.7	
GCACGATG	7	35	0.20	6.6	
TGCACGTT	16	81	0.20	9.9	
>cluster_52	TCAGGTAA				
TCAGGTAA	32	155	0.21	6.2	MIR222
>cluster_53	TTTCTATG				
TTTCTATG	109	538	0.20	11.1	
TTTTCTAC	115	592	0.19	10.9	
>cluster_54	TAATGTGA				
TAATGTGA	75	370	0.20	9.2	miR-323(#1)
AAATGTGA	172	945	0.18	12.3	MIR61
>cluster_55	TTTATTGC				
TTTATTGC	99	496	0.20	10.4	
>cluster_56	AAGCGCTT				
AAGCGCTT	10	50	0.20	7.9	

>cluster_57	GGGCATTA				
GGGCATTA	24	122	0.20	5.1	MIR138 MIR179
AGCATTAA	55	303	0.18	7.0	miR-155
>cluster_58	ATAGTGTA				
ATAGTGTA	36	183	0.20	6.2	
>cluster_59	GTATTGTA				
GTATTGTA	56	285	0.20	7.7	
>cluster_60	CACTGCCA				
CACTGCCA	98	502	0.20	10.1	miR-34a MIR141 MIR144 MIR199
TCACTGCC	83	441	0.19	8.9	MIR199
>cluster_61	ATAAGCTA				
ATAAGCTA	40	206	0.19	6.4	miR-21 miR-154(#3)
>cluster_62	TAAAGCTT				
TAAAGCTT	78	409	0.19	8.8	
ATAAAGCA	119	635	0.19	10.7	
ATAAAGCT	79	432	0.18	8.4	
>cluster_63	GTATTTTG				
GTATTTTG	141	737	0.19	11.9	
CTGTATTT	181	965	0.19	13.2	
>cluster_64	GTGGCCTT				
GTGGCCTT	75	394	0.19	8.6	MIR128
>cluster_65	GACTGTTA				
GACTGTTA	36	194	0.19	5.8	miR-212 miR-132
>cluster_66	CAGTGTTA				
CAGTGTTA	75	404	0.19	8.4	miR-200a miR-141
>cluster_67	AAAGGCTC				
AAAGGCTC	46	247	0.19	6.6	
>cluster_68	TTTGTGCA				
TTTGTGCA	86	468	0.18	8.9	
>cluster_69	TTTGGTAC				
TTTGGTAC	38	206	0.18	5.9	
>cluster_70	TTTTGCTA				
TTTTGCTA	92	510	0.18	9.0	
>cluster_71	GTCTTCCA				
GTCTTCCA	53	295	0.18	6.8	miR-7
>cluster_72	GATTAAAG				
GATTAAAG	45	250	0.18	6.2	

Supplementary Table S6 **Pairing of conserved 8-mer motifs to miRNA sequences**

Perfect Waston-Crick pairing

miRNA	Sequence (reverse strand)	matched motifs	C	N	pC	MCS
hsa-miR-92	CAGGCCGGGACAAgtgcaata	GTGCAATA	160	291	0.55	30.6
hsa-miR-32	GCAACTTAGTAATgtgcaata	GTGCAATA	160	291	0.55	30.6
hsa-miR-124a	TGGCATTCACCGCgtgccttaA	GTGCCTTA	147	271	0.54	29.1
hsa-miR-98	AACAATACAACTTActacctca	CTACCTCA	139	263	0.53	27.8
hsa-let-7i	ACAGCACAAACTActacctca	CTACCTCA	139	263	0.53	27.8
hsa-let-7g	ACTGTACAAACTActacctca	CTACCTCA	139	263	0.53	27.8
hsa-let-7f	AACTATACAATCTActacctca	CTACCTCA	139	263	0.53	27.8
hsa-let-7e	ACTATACAACCTCctacctca	CTACCTCA	139	263	0.53	27.8
hsa-let-7c	AACCATACAACCTActacctca	CTACCTCA	139	263	0.53	27.8
hsa-let-7b	AACCACACAACCTActacctca	CTACCTCA	139	263	0.53	27.8
hsa-let-7a	AACTATACAACCTActacctca	CTACCTCA	139	263	0.53	27.8
hsa-miR-9	TCATACAGCTAGATAaccaaaga	ACCAAAGA	178	366	0.49	29.7
hsa-miR-30e-5p	TCCAGTCAAGGAtgtttaca	TGTTTACA	369	771	0.48	42.3
hsa-miR-30d	CTTCCAGTCGGGGAtgtttaca	TGTTTACA	369	771	0.48	42.3
hsa-miR-30c	GCTGAGAGTGTAGGAtgtttaca	TGTTTACA	369	771	0.48	42.3
hsa-miR-30b	AGCTGAGTGTAGGAtgtttaca	TGTTTACA	369	771	0.48	42.3
hsa-miR-30a-5p	CTTCCAGTCGAGGAtgtttaca	TGTTTACA	369	771	0.48	42.3
hsa-miR-20	CTACCTGCACTATAAgcacttta	GCACTTTA	193	405	0.48	30.5
hsa-miR-106b	ATCTGCACTGTCAgcacttta	GCACTTTA	193	405	0.48	30.5
hsa-miR-29c	ACCGATTTCAAAtggtgcta	TGGTGCTA	116	272	0.43	21.9
hsa-miR-29b	AACACTGATTTCAAAtggtgcta	TGGTGCTA	116	272	0.43	21.9
hsa-miR-29a	AACCGATTTCAGAtggtgcta	TGGTGCTA	116	272	0.43	21.9
hsa-miR-153	TCACTTTTGTGActatgcaa	CTATGCAA	97	231	0.42	19.8
hsa-miR-26b	AACCTATCCTGAATtacttgaa	TACTTGAA	178	425	0.42	26.8
hsa-miR-26a	GCCTATCCTGGATtacttgaa	TACTTGAA	178	425	0.42	26.8
hsa-miR-96	GCAAAAATGTGCTAgtgccaaa	GTGCCAAA	118	291	0.41	21.3
hsa-miR-101	CTTCAGTTATCACAgtactgta	GTACTGTA	136	338	0.40	22.7
hsa-miR-218	ACATGGTTAGATCaagcacaa	AAGCACAA	157	393	0.40	24.3
hsa-let-7d	ACTATGCAACCTactacctcT	ACTACCTC	63	160	0.39	15.2
hsa-miR-99a	CACAAGATCGGATCtacgggtt	TACGGGTT	8	21	0.38	10.4
hsa-miR-100	CACAAGTTCGGATCtacgggtt	TACGGGTT	8	21	0.38	10.4
hsa-miR-137	CTACGCGTATTCTTaagcaata	AAGCAATA	145	386	0.38	22.3
hsa-miR-19b	TCAGTTTTGCATGGAtttgcaca	TTTGCACA	207	583	0.36	25.5
hsa-miR-19a	TCAGTTTTGCATAGAtttgcaca	TTTGCACA	207	583	0.36	25.5
hsa-miR-135b	CACATAGGAATGAAaagccata	AAGCCATA	92	261	0.35	16.9
hsa-miR-135a	TCACATAGGAATAAAaagccata	AAGCCATA	92	261	0.35	16.9
hsa-miR-27b	GCAGAACTTAGCCactgtgaa	ACTGTGAA	192	551	0.35	24.2
hsa-miR-27a	GCGGAACTTAGCCactgtgaa	ACTGTGAA	192	551	0.35	24.2
hsa-miR-182	TGTGAGTTCTACCAttgccaaa	TTGCCAAA	172	521	0.33	22.0
hsa-miR-367	TCACCATTGCTAAagtgcaatT	AGTGCAAT	90	273	0.33	15.9
hsa-miR-25	TCAGACCGAGACAagtgcaatG	AGTGCAAT	90	273	0.33	15.9
hsa-miR-195	GCCAATATTTCTGtgctgcta	TGCTGCTA	124	380	0.33	18.5
hsa-miR-16	CGCCAATATTTACGtgctgcta	TGCTGCTA	124	380	0.33	18.5
hsa-miR-15b	TGTAAACCATGATGtgctgcta	TGCTGCTA	124	380	0.33	18.5
hsa-miR-15a	CACAAACCATTATGtgctgcta	TGCTGCTA	124	380	0.33	18.5
hsa-miR-128b	GAAAGAGACCGGTTcactgtga	CACTGTGA	132	414	0.32	18.7
hsa-miR-128a	AAAAGAGACCGGTTcactgtga	CACTGTGA	132	414	0.32	18.7
hsa-miR-206	CCACACACTTCCTTacattcca	ACATTCCA	127	402	0.32	18.2
hsa-miR-1	TACATACTTCTTTacattcca	ACATTCCA	127	402	0.32	18.2
hsa-miR-93	CTACCTGCACGAACagcacttt	AGCACTTT	190	601	0.32	22.3
hsa-miR-372	ACGCTCAAATGTCGCagcacttt	AGCACTTT	190	601	0.32	22.3
hsa-miR-17-5p	ACTACCTGCACTGTAagcactttG	AGCACTTT	190	601	0.32	22.3
hsa-miR-106a	GCTACCTGCACTGTAagcactttT	AGCACTTT	190	601	0.32	22.3
hsa-miR-200c	CCATCATTACCCGGcagtatta	CAGTATTA	121	400	0.30	17.2
hsa-miR-200b	GTCATCATTACCAGGcagtatta	CAGTATTA	121	400	0.30	17.2
hsa-miR-219	AGAATTGCGTTTGgacaatca	GACAATCA	46	155	0.30	10.4
hsa-miR-125b	TCACAAGTTAGGGTctcaggga	CTCAGGGA	129	451	0.29	16.9
hsa-miR-125a	CACAGGTTAAAGGGTctcaggga	CTCAGGGA	129	451	0.29	16.9
hsa-miR-301	GCTTTGACAATACTAttgcactg	TTGCACTG	110	399	0.28	15.1
hsa-miR-130b	ATGCCCTTTCATCAttgcactg	TTGCACTG	110	399	0.28	15.1
hsa-miR-130a	ATGCCCTTTTAACAttgcactg	TTGCACTG	110	399	0.28	15.1
hsa-miR-142-3p	TCCATAAAGTAGGAAacactaca	ACACTACA	54	199	0.27	10.4
hsa-miR-152	CCCAAGTTCTGTCAtgcactga	TGCACTGA	103	403	0.26	13.7
hsa-miR-148b	ACAAAGTTCTGTGAtgcactga	TGCACTGA	103	403	0.26	13.7
hsa-miR-148a	ACAAAGTTCTGTAGtgcactga	TGCACTGA	103	403	0.26	13.7
hsa-miR-302d	ACACTCAAACATGGaagcacttA	AAGCACTT	123	483	0.26	14.9
hsa-miR-302c	CCACTGAAACATGGaagcacttA	AAGCACTT	123	483	0.26	14.9
hsa-miR-302b	CTACTAAAACATGGaagcacttA	AAGCACTT	123	483	0.26	14.9
hsa-miR-302a	TCACCAAAACATGGaagcacttA	AAGCACTT	123	483	0.26	14.9
hsa-miR-373	ACACCCCAAAATCGaagcacttC	AAGCACTT	123	483	0.26	14.9
hsa-miR-199a*	AACCAATGTGCAGActactgta	CTACTGTA	69	283	0.24	10.7
hsa-miR-183	CAGTGAATTCTACCAgtgccata	GTGCCATA	41	170	0.24	8.2
hsa-miR-196b	CCAACAACAGGAaactacctA	AACTACCT	43	194	0.22	7.7
hsa-miR-196a	CCAACAACATGAaactacctA	AACTACCT	43	194	0.22	7.7
hsa-miR-424	TTCAAAACATGAAttgctgctG	TTGCTGCT	145	664	0.22	13.9
hsa-miR-139	AGACACGtgcactgtAGA	TGCACTGT	90	418	0.22	10.8
hsa-miR-181c	ACTCACCGACAGGttgaatgtT	TTGAATGT	123	586	0.21	12.3
hsa-miR-181a	ACTCACCGACAGCGttgaatgtT	TTGAATGT	123	586	0.21	12.3
hsa-miR-133b	TAGCTGGTTGAAGgggaccaa	GGGACCAA	44	214	0.21	7.2
hsa-miR-133a	ACAGCTGGTTGAAGgggaccaa	GGGACCAA	44	214	0.21	7.2
hsa-miR-34a	AACAACCAGCTAAGAcactgcca	CACTGCCA	98	502	0.20	10.1
hsa-miR-21	TCAACATCAGTCTGataagcta	ATAAGCTA	40	206	0.19	6.4
hsa-miR-144	CTAGTACATCATCTatactgta	ATACTGTA	80	420	0.19	8.9
hsa-miR-212	GGCCGTGACTGGAgactgtta	GACTGTTA	36	194	0.19	5.8
hsa-miR-200a	ACATCGTTACCAGAcagtgtta	CAGTGTTA	75	404	0.19	8.4
hsa-miR-141	CCATCTTTACCAGAcagtgtta	CAGTGTTA	75	404	0.19	8.4
hsa-miR-132	CGACCATGGCTGTAgactgtta	GACTGTTA	36	194	0.19	5.8
hsa-miR-381	ACAGAGAGCTTGCCcttgtata	CTTGTATA	62	336	0.19	7.6
hsa-miR-155	CCCCTATCACGATTagcattaa	AGCATTAA	55	303	0.18	7.0
hsa-miR-7	CAACAAAATCACTAgtcttcca	GTCTTCCA	53	295	0.18	6.8

Supplementary Table S6, Continued

Allow one-base mismatch in Watson-Crick pairing

miRNA	Sequence (reverse strand)	matched motifs	C	N	pC	MCS	last 8-mer in miRNA	matched motifs (allow one mismatch)	C	N	pC	MCS	Mismatched pairing
T-G pairing													
hsa-miR-126	GCATTATTACTCAcggtacga	CGGTACGA	0	7	0.00	-0.4	CGGTACGA	TGGTACGA	6	25	0.24	6.8	C->T
hsa-miR-18	TATCTGCACTAGATgcacctta	GCACCTTA	24	137	0.18	4.4	GCACCTTA	GCACTTTA	193	405	0.48	30.5	C->T
hsa-miR-361	GTACCCCTGGAGATtctgataa	TCTGATAA	36	284	0.13	3.3	TCTGATAA	TTTGATAA	116	551	0.21	12.0	C->T
hsa-miR-122a	ACAAACACCATTGTCacactcca	ACACTCCA	23	228	0.10	1.4	ACACTCCA	ACATTCCA	127	402	0.32	18.2	C->T
Mismatch between first base of miRNA and last letter 'A' of the conserved motifs													
hsa-miR-181b	CCCACCGACAGCAatgaatgtT	ATGAATGT	108	626	0.17	9.2	TGAATGTT	TGAATGTA	183	594	0.31	21.4	T->A
hsa-miR-323	AGAGGTCGACCGtgtaatgtGC	TGTAATGT	76	451	0.17	7.5	TAATGTGC	TAATGTGA	75	370	0.20	9.2	C->A
hsa-miR-221	GAAACCCAGCAGACaatgtagcT	AATGTAGC	41	244	0.17	5.5	ATGTAGCT	ATGTAGCA	57	265	0.22	8.6	T->A
hsa-miR-145	AAGGGATTCCTGGGAaaactggaC	AAACTGGA	78	466	0.17	7.5	AACTGGAC	AACTGGAA	101	501	0.20	10.7	C->A
hsa-miR-199b	GAACAGATAGTCTAaacactggG	AACACTGG	47	316	0.15	4.9	ACACTGGG	ACACTGGA	78	365	0.21	10.0	G->A
hsa-miR-199a	GAACAGGTAGTCTGaacactggG	AACACTGG	47	316	0.15	4.9	ACACTGGG	ACACTGGA	78	365	0.21	10.0	G->A
hsa-miR-23b	GGTAATCCCTGGcaatgtgaT	CAATGTGA	45	310	0.15	4.6	AATGTGAT	AATGTGAA	207	872	0.24	18.1	T->A
hsa-miR-23a	GGAAATCCCTGGcaatgtgaT	CAATGTGA	45	310	0.15	4.6	AATGTGAT	AATGTGAA	207	872	0.24	18.1	T->A
hsa-miR-138	GATTCACAacaccagcT	ACACCAGC	35	271	0.13	3.3	CACCAGCT	CACCAGCA	104	405	0.26	13.8	T->A
hsa-miR-22	ACAGTTCTTCAACtggcagctT	TGGCAGCT	54	460	0.12	3.4	GGCAGCTT	GGCAGCTA	39	172	0.23	7.5	T->A
hsa-miR-222	GAGACCCAGTAGCCAGatgtagct	ATGTAGCT	30	268	0.11	2.3	ATGTAGCT	ATGTAGCA	57	265	0.22	8.6	T->A
Other mismatches													
hsa-miR-34c	GCAATCAGCTAACTacactgccT	ACACTGCC	55	323	0.17	6.4	CACTGCCT	CAGTGCCT	102	512	0.20	10.6	C->G
hsa-miR-107	TGATAGCCCTGTACAatgctgct	ATGCTGCT	80	501	0.16	7.1	ATGCTGCT	ATGGTGCT	108	343	0.32	16.8	C->G
hsa-miR-103	TCATAGCCCTGTACAatgctgct	ATGCTGCT	80	501	0.16	7.1	ATGCTGCT	ATGGTGCT	108	343	0.32	16.8	C->G
hsa-miR-330	TCTCTGCAGGCCgtgtgcttTGC	GTGTGCTT	46	327	0.14	4.5	TGCTTTGC	TGCCTTGC	62	336	0.19	7.6	T->C
hsa-miR-208	ACAAGCTTTTTGCtcgtcttaT	TCGTCTTA	4	33	0.12	3.5	CGTCTTAT	CGCCTTAT	6	24	0.25	7.0	T->C
hsa-miR-224	TAAACGGAACCACtagtgactTG	TAGTGACT	19	185	0.10	1.4	GTGACTTG	GTGCCTTG	133	355	0.38	21.3	A->C
hsa-miR-99b	CGCAAGGTCGGttctacggGTG	TTCTACGG	2	32	0.06	1.3	TACGGGTG	TACGGGTT	8	21	0.38	10.4	G->T
hsa-miR-9*	ACTTTCGGTTATCtagcttta	TAGCTTTA	52	326	0.16	5.7	TAGCTTTA	TAGCCTTA	33	162	0.20	6.2	T->C
hsa-miR-34b	CAATCAGCTAATGacactgccTA	ACACTGCC	55	323	0.17	6.4	ACTGCCTA	AGTGCCTA	41	185	0.22	7.5	C->G
hsa-miR-217	ATCCAATCAGTTCCTGatgcagta	ATGCAGTA	32	253	0.13	3.1	ATGCAGTA	ATGCAATA	85	294	0.29	13.8	G->A
hsa-miR-338	TCAACAAAATCACTgatgctggA	GATGCTGG	25	318	0.08	0.2	ATGCTGGA	ATGCTGCA	75	332	0.23	10.4	G->C
hsa-miR-154	CGAAGGCAACACGGataaccta	ATAACCTA	10	141	0.07	-0.2	ATAACCTA	ATAAGCTA	40	206	0.19	6.4	C->G

C: **Number of conserved instances**
N: **Number of total instances in human sequences**
PC: **Conservation rate**
MCS: **Conservation Score**

Supplementary Table S7 **Discovered 3' UTR motifs not related to miRNA**

No.	Motif	Conserved Num	Total Num	Pc	MCS
1	AATAAA	6617	14266	0.46	135.7
2	TATTTAT	1758	3706	0.47	79.4
3	TGTAnATA	1528	2968	0.51	70.4
4	TATTTTT	2068	6861	0.30	58.9
5	TTTGTA	2777	8873	0.31	53.4
6	TTTTATA	1185	3861	0.31	45.4
7	TTTTGT	3185	13094	0.24	41.7
8	TGTRnnTTT	1575	6878	0.23	36.9
9	TAATTTAT	331	868	0.38	33.3
10	TGTACAKW	712	2048	0.35	33.0
11	TGTRnnnnTGT	1018	4650	0.22	31.1
12	AGCMWTAA	348	1064	0.33	29.0
13	TATTAAA	665	2800	0.24	26.0
14	TGTRnnATA	678	2727	0.25	25.0
15	TATTTATTG	152	344	0.44	24.1
16	WnTATWTTG	707	3180	0.22	23.4
17	WGTAWWTATT	229	742	0.31	22.8
18	TTTnnnnYGTA	685	3322	0.21	22.0
19	TAATATAT	192	641	0.30	20.9
20	TATWTTnnTAC	145	440	0.33	20.7
21	TTTKnnTAC	686	3330	0.21	19.9
22	WRTAAATG	550	2736	0.20	19.4
23	TGTAnnnTAT	421	1786	0.24	18.6
24	TGTAnnWWnTGTA	113	313	0.36	18.6
25	TTCnnWATAAA	127	511	0.25	17.0
26	CTTWRTAA	325	1584	0.21	16.4
27	CTATKYATT	130	491	0.26	16.1
28	YGTAnAKRnTTT	112	353	0.32	15.7
29	CTCRnTAAA	117	525	0.22	15.1
30	TnTATnTGTAnR	139	598	0.23	14.9
31	TGCnnWRTAAA	122	513	0.24	14.8
32	TGTRCCAW	220	988	0.22	14.7
33	TGTnnnAWTAAA	128	608	0.21	14.6
34	AATAWAnnTTG	110	489	0.22	14.5
35	WRTAAnnnnYGTAnW	108	437	0.25	14.3
36	GTTWTnTAT	240	1170	0.21	14.3
37	AWTAAAnnCTT	109	530	0.21	13.7
38	TATTTWnATG	142	610	0.23	13.7
39	ATAnTGTAnW	230	989	0.23	13.6
40	TTCnAnTAAA	117	553	0.21	13.3
41	TTTnnnRYCAAA	128	616	0.21	12.8
42	TTGKAWTTAW	117	480	0.24	12.8
43	AATRMAnTGT	165	823	0.20	12.8
44	TCTRTRnATA	124	531	0.23	11.9
45	AATMWAGTT	117	577	0.20	11.9
46	TGTRYMAATR	113	482	0.23	11.8
47	YAATRWAGC	106	488	0.22	11.7
48	AGAnTATTWW	127	632	0.20	11.5
49	AGAKnTnTATW	120	582	0.21	11.2
50	WKTACWnKAAA	116	580	0.20	10.6
51	TGTWnAnAGC	115	572	0.20	10.3
52	YRAAGYnTTA	123	606	0.20	9.9
53	YYGTAnnnnKATT	108	514	0.21	9.7
54	GTTGTAnA	191	927	0.21	9.5
55	GGTACGAA	8	25	0.32	9.3
56	TATTKnnnnGTAnW	110	545	0.20	9.2
57	CTTRYRnATA	111	513	0.22	8.4
58	GTCAATAA	49	214	0.23	8.2
59	TAACGGGT	5	14	0.36	7.8
60	TRTAAnTAC	116	574	0.20	7.5

Supplementary Table S10 **List of 3' primers used for tested miRNAs**

miRNA ID	Predicted miRNA mature sequence	Gene-specific 3’ primer
MIR1	TATTGCACTCGTCCCGGCCTCC	TGGAGGCCGGGACGA
MIR21	CACAGTGTGGTTTGGACGTGGC	TGGCCACGTCCAAACC
MIR41	TTGCATATGTAGGATGTCCCAT	GAGATGGGACATCCTACA
MIR57	AAGGCAACTTTTGTTTGAGTAT	TTTGATACTCAAACAAAAGT
MIR115	TAATACTGTCTGGTAAAACCGT	GGACGGTTTTACCAGAC
MIR134	TTTGGTACTTGGAGAGTGGTTA	GATAACCACTCTCCAAG
MIR136	CCCTGAAAATTTCTCATTTAGG	CTGGCCTAAATGAGAAATTT
MIR138/MIR179	TAATGCCCCTAAAAATCCTTAT	ACAATAAGGATTTTTAGGG
MIR144	TGGCAGTGTATTGTTAGCTGGT	CAACCAGCTAACAATACA
MIR156	TACAAAAGCTTATTTGAACATG	CCCATGTTCAAATAAGCT
MIR178	TTTTTGCGATGTGTTCCTAATA	TGCATATTAGGAACACATC
MIR211	ATTAGTGTGGGATGATCATGAC	AATGTCATGATCATCCCA

Supplementary Table S11 **List of predicted miRNAs that share high sequence similarity to known miRNAs**

ID	**Predicted miRNA sequence**	**Predicted miRNA location**	**Known miRNA**	**known miRNA sequence**	**Known miRNA location**	**Number of similar bases in ungapped alignment**
MIR258	TTAAGGTGCATCTAGTGCAGTT	chrX:133029570-133029680	hsa-mir-18	TAAGGTGCATCTAGTGCAGATA	chr13:90801006-90801076	20
MIR103	CCAAAGTGCTCATAGTGCAGGT	chrX:133029336-133029446	hsa-mir-20	TAAAGTGCTTATAGTGCAGGTAG	chr13:90801320-90801390	19
MIR1	TATTGCACTCGTCCCGGCCTCC	chr1:151978032-151978142	hsa-mir-92	TATTGCACTTGTCCCGGCCTG	chrX:133029088-133029162	19
MIR115	CTAATACTGTCTGGTAAAACCG	chr1:1144291-1144401	hsa-mir-141	TAACACTGTCTGGTAAAGATGG	chr12:6943521-6943615	17
MIR199	TTGGCAGTGATTATGCGGGTTG	chr15:35094715-35094825	hsa-mir-34a	TGGCAGTGTCTTAGCTGGTTGTT	chr1:9145993-9146102	16
MIR150	TTATTGCCACAACTTTGGGGTG	chrX:39268491-39268601	hsa-mir-373	GAAGTGCTTCGATTTTGGGGTGT	chr19:58983771-58983839	15
MIR157	GAAGGCACAGTTAAAGGGTCAT	chr19:19461191-19461301	hsa-mir-130a	CAGTGCAATGTTAAAAGGGCAT	chr11:57165247-57165335	15
MIR192	TTCACAGTGGGAGAAATATGCT	chr1:117276990-117277100	hsa-mir-27b	TTCACAGTGGCTAAGTTCTGC	chr9:94927282-94927378	15
MIR221	ATCTTTGGCTGTATATCTTTCT	chr10:77069708-77069818	hsa-mir-9	TCTTTGGTTATCTAGCTGTATGA	chr15:87712252-87712341	15
MIR238	TTCAGTGCAGAACTAAAATATG	chr21:15392473-15392583	hsa-mir-148a	TCAGTGCACTACAGAACTTTGT	chr7:25762779-25762846	15
MIR243	TTCAAGTAAATCACTTTTTGTC	chr9:16793571-16793681	hsa-mir-26b	TTCAAGTAATTCAGGATAGGTT	chr2:219092874-219092950	15
MIR252	CATTGCACTGTATGAATCTGGA	chr1:107675785-107675895	hsa-mir-367	AATTGCACTTTAGCAATGGTGA	chr4:113926634-113926701	15
MIR257	GTGTAAACACCATAAAGCAAGC	chr8:65341824-65341934	hsa-mir-30b	TGTAAACATCCTACACTCAGCT	chr8:135881945-135882032	15
MIR87	TTATGGCACCCATGGCTGCCTC	chrX:138895637-138895747	hsa-mir-346	TGTCTGCCCGCATGCCTGCCTCT	chr10:88014424-88014509	15