

Post-transcriptional regulation in the human genome: motif discovery, microRNA gene identification and miRNA targets.

Xiaohui Xie¹, Eric S. Lander^{1,2}, Manolis Kellis^{1,3}

xhx@broad.mit.edu, lander@broad.mit.edu, manoli@mit.edu

(1) Broad Institute of MIT and Harvard, Cambridge MA 02139

(2) Whitehead Institute for Biomedical Research, Cambridge MA 02139

(3) MIT Computer Science and Artificial Intelligence Laboratory, Cambridge MA 02139

1. Introduction

The study of gene regulation has largely focused on transcriptional control and the events leading to transcription initiation, which is perhaps the single most important determinant of gene expression levels. However, many mechanisms of post-transcriptional control govern cellular control in higher eukaryotes, and their computational study can lead to major insights in the understanding of cellular circuitry. In order to discover the regulatory motifs involved in post-transcriptional regulation in the human genome, we studied the conservation properties of human 3'-untranslated regions aligned across the four sequenced mammal species, human, mouse, rat, and dog¹. The motifs discovered fall in two classes, the first apparently targeted by RNA-binding proteins, and the second involved in microRNA (miRNA) regulation. The results suggest a new method for miRNA identification, starting from well-conserved 8-mers in 3' untranslated regions of human, and leading to the discovery of miRNAs that target them. This method led to the discovery of hundreds of new miRNA genes, several of which were experimentally confirmed.

2. Motif discovery in 3'-untranslated regions

We searched for highly conserved sequence profiles in 3'-UTR regions. By enumerating a large number of motifs possibly with 2-fold degeneracies and unspecified bases, we searched for highly conserved and frequently occurring sequence profiles (for details, please see our poster on motif discovery in human promoter regions). After clustering, we discovered 106 highly conserved motifs in 3'-UTRs with conservation scores greater than 6 standard deviations above what is expected by chance under a binomial model.

The discovered motifs showed several unusual properties. First, they show a strong directional bias with respect to DNA strand; whereas promoter motifs tend to have similar conservation rates on the coding and non-coding strands, the 3'-UTR motifs are preferentially conserved in only one strand. This strand specificity suggests that 3'-UTR motifs act at the level of RNA rather than DNA, and thus are likely to play a role in post-transcriptional regulation.

Several 3'-UTR motifs match known post-transcriptional regulatory signals. The list includes the known poly-adenylation signal and various AT-rich elements², which are believed to be involved in controlling mRNA stability and degradation. It also contains 11 motifs with a TGTA core sequence, flanked by different variants of an AT-rich (the most conserved being as TGTANATA). Recent work in yeast has identified such motifs as binding sites for the Puf family of RNA-binding proteins; they may represent binding sites for the homologous mammalian proteins (such as PUM1 and PUM2 in human³).

The remaining 3'-UTR motifs showed an unusual length distribution, with a strong peak at length 8. By contrast, promoter motifs do not show any such bias in their length distribution. Moreover, the motifs of length 8 have a strong tendency to end with the nucleotide 'A' (whereas other 3'-UTR motifs do not show this tendency). This suggested that the conserved 8-mers constitute a separate class of motifs, which we investigated further, finding these to be related to micro RNA genes.

3. Relationship with microRNAs

MicroRNA genes (miRNA)⁴ are small endogenous RNA genes, transcribed from long mRNA messages, processed into a 100-bp precursor, folded into a 50-bp stem loop. This stem is cleaved into a 20-bp double-stranded mature miRNA, which targets genes by loose complementarity with one of the two RNA strands, leading to degradation or translational repression of the target mRNA. Two properties of miRNAs suggest that the discovered 8-mer motifs may be miRNA targets: many mature miRNAs start with a 'U' base followed by a 7-base 'seed' complementary to a site in the 3'-UTR of target mRNAs⁵⁻⁷. These properties explain the preponderance of 8-mer motifs, and the large proportion of 8-mers ending in 'A' (complementary to 'U').

We thus reasoned that many of the highly conserved 8-mer motifs discovered by our unbiased procedure might be binding sites for conserved miRNAs. To investigate the relationship with miRNAs, we repeated the motif discovery procedure using only contiguous non-degenerate 8-mers. We identified the subset of 8-mers with conservation rate > 18% (vs. rate for a random 8-mer of 7.6%) and clustered these 8-mers into motifs defined as sets of similar 8-mers, resulting in 72 highly conserved 8-mer motifs.

We found that the 8-mer motifs discovered have complementary matches to the 207 distinct human miRNAs listed in the miRNA registry. Roughly 43.5% of the known miRNAs can match through Watson-Crick pairing to the highly conserved 8-mer motifs (versus 2% for an equal number of random 8-mers). Moreover, the matches begin at nucleotide 1 or 2 of the miRNA gene in more than 95% of cases. Thus, we confirm that the 8-mer motifs are likely targets of miRNA genes.

4. Identification of new microRNA genes

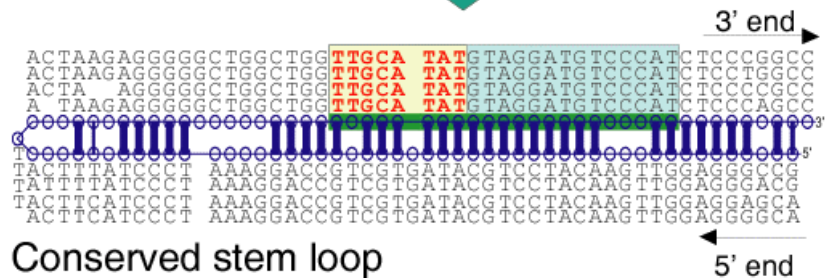
We then used the highly conserved 8-mer motifs to identify new miRNA genes that may target them by complementarity (see Figure). We first identified conserved occurrences of the 8-mer motifs in the entire human genome, searching both strands for motifs reverse complementary to each 8-mer, and excluding 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. 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, 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, alignment, conservation and location of the core 22-mer sequence containing the original motif at 5' end.

Motif enumeration and testing of 8mers



Conserved occurrence in the genome



Conserved stem loop

We developed a program to classify whether each of the windows passing the folding free energy thresholds could be a miRNA gene. For this purpose, we studied the folded structures and conservation properties of the 222 known miRNA genes, and used nine features to build a Bayesian classifier. The nine features we used were: (1) folding free energy; (2) conservation of the core 22-mer sequence; (3) conservation of the 8-mer sequence complementary to the core 22-mer; (4) number of paired bases in the core 22-mer sequence; (5) number of asymmetric bulges contained the 22-mer region; (6) distance of the core 22-mer to the loop; (7) size of the extended stem outside the core 22-mer; (8) conservation of the sequence in the loop region; and (9) conservation of the sequence surrounding the entire 110 bp window. For each feature f_i , we estimated its frequency $P(f_i)$ in the known miRNA genes and the background sequences $Q(f_i)$; to estimate the background rate, we used all returned windows above the specified folding energy threshold. Based on these nine features, we assigned a score S_{mir} to each window, as the sum of the log ratios: $S_{\text{mir}} = \sum_i \log_2 [P(f_i)/Q(f_i)]$ for all nine features. We report all candidate intervals whose S_{mir} score is above 5.0. This results in 242 candidate miRNA genes, including 113 known miRNA genes (51% of the total 222), and 129 candidate new miRNA genes.

5. Targets of miRNAs

The properties of the conserved motifs also allow inferences about the prevalence of regulation by miRNAs. Roughly 40% of human 3'-UTRs contain a conserved occurrence of one of the miRNA-associated 8-mer motifs, whereas only ~25% contain conserved occurrences of a comparable control set. This suggests that at least 20% of 3'-UTRs may be targets for conserved miRNA-based regulation at the 8-mer motifs. This greatly expands previous estimates of the number of targets of miRNA genes⁶. With sequence from more mammalian genomes, it should be possible to distinguish the conserved target sites with high specificity and sensitivity⁸. There are likely to be additional miRNA target sites in the human genome that are not conserved across all mammals; it should be possible to find most of these by genomic comparison with closer relatives such as primates.

6. Conclusion

The results thus provide an unbiased assessment of the relative importance of miRNA-based regulation in the human genome, consistent with recent independent studies⁹⁻¹¹. Overall, ~45% of the 108 highly conserved motifs in 3'-UTRs appear to be related to miRNA regulation and ~5000 human genes (~20% of the genome) are likely to be regulated by miRNAs through these conserved motifs.

1. Xie, X. et al. Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* (2005).
2. Chen, C. Y. & Shyu, A. B. AU-rich elements: characterization and importance in mRNA degradation. *Trends Biochem Sci* **20**, 465-70 (1995).
3. Spassov, D. S. & Jurecic, R. The PUF family of RNA-binding proteins: does evolutionarily conserved structure equal conserved function? *IUBMB Life* **55**, 359-66 (2003).
4. 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).
5. Lai, E. C. Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet* **30**, 363-4 (2002).
6. Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of mammalian microRNA targets. *Cell* **115**, 787-98 (2003).
7. Lim, L. P., Glasner, M. E., Yekta, S., Burge, C. B. & Bartel, D. P. Vertebrate microRNA genes. *Science* **299**, 1540 (2003).
8. Eddy, S. R. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol* **3**, e10 (2005).
9. Lewis, B. P., Burge, C. B. & Bartel, D. P. in *Cell* 15-20 (2005).
10. Lim, L. P. et al. in *Nature* (2005).
11. Berezikov, E. et al. in *Cell* 21-4 (2005).