The following software and databases were used in this work.

Automated: explains how software/database was used in the automated pipeline.
Analysis: explains how software/database was used in subsequent non-automated analysis.

Name	Description	Reference
Conserved Domain Database (CDD) version 2.08	Database of conserved domains, including descriptions of the domains, and profile HMMs. Automated and Analysis: used profile HMMs to classify domains within ORFs	(1)
RefSeq versions 14 and 19	Database of genome sequences, and nucleotide positions of ORFs. ORF positions are used with the CDD to classify their domains. Automated: used version 14 to collect potential UTRs, and search for additional homologs in automated refinement of motifs. Analysis: used version 19.	(2)
Rfam database version 7.0	Database of known structured RNAs. Automated: Used to mark known RNAs in motif predictions.	(3)
KEGG	Database of information on metabolism and associated genes. Analysis: informed analysis of motifs.	(4)
Acid mine drainage shotgun sequences	Analysis: searched these sequences for homologs.	(5)
Sargasso Sea shotgun sequences	Analysis: searched these sequences for homologs.	(6)
MicroFootPrinter	Automated: used to extract potential UTRs of genes, and, in some cases, to highlight promising motifs based on sequence conservation.	(7)
CMfinder version 0.2	Automated: predicted motifs using potential homologous UTRs, and refined motifs using additional homologs. Analysis: assisted in manual improvement of alignments.	(8)
RAVENNA version 0.2f	Automated and Analysis: homology searches for RNA motifs.	(9-11)
Infernal version 0.7	Software integrated into RAVENNA. Implements covariance models, and GSC algorithm (used to establish levels of conservation/covariation in motif diagrams).	(12)
RSEARCH	Software integrated into RAVENNA. Implements E-value statistics for covariance models.	(13)
Additional scripts	Some scripts were created specifically to support this pipeline.	(14)

RALEE	Analysis: used to edit motif alignments.	(15)
NCBI BLAST	Analysis: tblastn was used to find homologs of protein-coding genes, in order to search their potential UTRs. rpsblast was used to classify ORFs in RefSeq into profile HMMs from the Conserved Domain Database.	(16)
Mfold	Analysis: used to predict potential structures, particularly for variable-length stems, whose identity might not be conserved among homologs.	(17)
Rnall	Analysis: used to predict rho-independent transcription terminator hairpins.	(18)

1. Marchler-Bauer, A., Anderson, J.B., Cherukuri, P.F., DeWeese-Scott, C., Geer, L.Y., Gwadz, M., He, S., Hurwitz, D.I., Jackson, J.D., Ke, Z. *et al.* (2005) CDD: a Conserved Domain Database for protein classification. *Nucleic Acids Research*, **33,** 192-196.
2. Pruitt, K., Tatusova, T. and Maglott, D. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, **33,** 501-504.
3. Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R. and Bateman, A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Research*, **33,** 121-124.
4. Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. and Hirakawa, M. (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res*, **34,** D354-357.
5. Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S. and Banfield, J.F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428,** 37-43.
6. Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W. *et al.* (2004) Environmental Genome Shotgun Sequencing of the Sargasso Sea. *Science*, **304,** 66-74.
7. Neph, S. and Tompa, M. (2006) MicroFootPrinter: a Tool for Phylogenetic Footprinting in Prokaryotic Genomes. *Nucleic Acids Research*, **34**.
8. Yao, Z., Weinberg, Z. and Ruzzo, W.L. (2006) CMfinder–a covariance model based RNA motif finding algorithm. *Bioinformatics*, **22,** 445-452.
9. Weinberg, Z. and Ruzzo, W.L. (2004), *RECOMB04: Proceedings of the Eighth Annual International Conference on Computational Molecular Biology*, San Diego, CA, pp. 243-251.
10. Weinberg, Z. and Ruzzo, W.L. (2004) Exploiting conserved structure for faster annotation of non-coding RNAs without loss of accuracy. *Bioinformatics*, **20,** i334-i341.
11. Weinberg, Z. and Ruzzo, W.L. (2006) Sequence-based heuristics for faster annotation of non-coding RNA families. *Bioinformatics*, **22,** 35-39.
12. Eddy, S.R. (2005) Infernal User's Guide.

13. Klein, R.J. and Eddy, S.R. (2003) RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics*, **4,** 44.
14. Yao, Z., Barrick, J.E., Weinberg, Z., Neph, S., Breaker, R.R., Tompa, M. and Ruzzo, W.L. (2007) A computational pipeline for high throughput discovery of *cis*-regulatory noncoding RNA in prokaryotes. In press.
15. Griffiths-Jones, S. (2005) RALEE-RNA ALignment Editor in Emacs. *Bioinformatics*, **21,** 257-259.
16. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25,** 3389-3402.
17. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, **31,** 3406-3415.
18. Wan, X.-F. and Xu, D. (2004) Intrinsic Terminator Prediction and Its Application in *Synechococcus* sp. WH8102. *J Comp Sci & Tech,* **20,** 465-482.