Published online 9 November 2010 *Nucleic Acids Research, 2011, Vol. 39, Database issue* ***D141–D145***
doi:10.1093/nar/gkq1129

Rfam: Wikipedia, clans and the "decimal" release

Paul P. Gardner[1,*], Jennifer Daub[1], John Tate[1], Benjamin L. Moore[1], Isabelle H. Osuch[1], Sam Griffiths-Jones[2], Robert D. Finn[3], Eric P. Nawrocki[3], Diana L. Kolbe[3], Sean R. Eddy[3] and Alex Bateman[1,*]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA0, [2]Faculty of Life Sciences, The University of Manchester, Manchester M13 9PL, UK and [3]Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn, VA 20147, USA

Received September 15, 2010; Revised October 20, 2010; Accepted October 21, 2010

ABSTRACT

The Rfam database aims to catalogue non-coding RNAs through the use of sequence alignments and statistical profile models known as covariance models. In this contribution, we discuss the pros and cons of using the online encyclopedia, Wikipedia, as a source of community-derived annotation. We discuss the addition of groupings of related RNA families into clans and new developments to the website. Rfam is available on the Web at http://rfam.sanger.ac.uk.

INTRODUCTION

The Rfam database maintains alignments, consensus secondary structures, covariance models (CMs) and corresponding annotation for RNA families. Each family represents a set of RNA sequences that function at the RNA level and share a clear common ancestor. Some examples are tRNA, microRNAs, spliceosomal RNAs, riboswitches, CRISPR elements and thermosensors. The primary purpose of the Rfam database is the automated, accurate annotation of non-coding RNAs (ncRNAs) in genomic sequences. Rfam is also frequently used as a source of high-quality alignments for training and benchmarking RNA sequence analysis software tools (1–5). Additionally, in the absence of a well-curated and up-to-date general RNA sequence database, equivalent to UniProt in the protein coding world, Rfam is also often used as a source of individual ncRNA sequences.

As described in previous Rfam publications, the database is built upon well-curated seed alignments of representative members of an RNA family (6–8). These are used to build CMs, statistical models of a family's conserved sequence and secondary structure, using the Infernal suite of analysis tools (9). The resultant covariance models are used to scan a large database of nucleotide sequences that is derived from the EMBL nucleotide archive (10). The searches return a list of putative homologs, or hits, ranked by bit-scores derived from the CMs. A hit's bit-score is the log odds ratio of the probability the hit was generated by the CM versus a random model of background sequence. An expert curator provides a threshold that in their opinion best discriminates between *bona fide* homologs to the seed sequences and the background distribution of false hits. Subsequently, all sequences with a bit-score above the threshold are included in an automatically generated alignment to the CM.

NEW DEVELOPMENTS

The Rfam 10.0 "decimal" release

In order to keep Rfam as up-to-date as possible we aim to make regular releases of the database. These releases are snap-shots of the live, internal version of the database that are made publicly available via the websites and ftp. We have two types of release. A major release (indicated by an integer and a '.0' in the version number e.g. '10.0') usually involves updating the underlying sequence database, Rfamseq, to the latest version of EMBL and remapping all the seed sequences to the new databases. All the families are subsequently searched against the new database and, if necessary, re-thresholded. Minor releases are indicated by '.1', '.2', etc. in the version number e.g. '10.1'. These are usually made after adding many new families to the database built on the same underlying sequence database.

Rfam 10.0 was released in early 2010. This release included a major update to the underlying search algorithm, switching to a new version of Infernal, v1.0 (9). This required individually re-thresholding each Rfam family due to an important change in Infernal's underlying

*To whom correspondence should be addressed. Tel: +44 1223 494 726; Fax: +44 1223 494 919; Email: pg5@sanger.ac.uk
Correspondence may be also be addressed to Alex Bateman. Tel: +44 1223 494950; Fax: +44 1223 494919; Email: agb@sanger.ac.uk

scoring scheme from maximum likelihood alignment scores to summed scores over all possible alignments [i.e. switching from using the CYK algorithm to the Inside algorithm (11)]. Additionally, the new version of Infernal reports estimates of the statistical significance of hits (*E*-values) returned from database searches using Rfam 10.0 CM files. We also mapped all the families and searched a new version of Rfamseq based on EMBL 100 (10). The result of these and other internal improvements to our pipeline resulted in a 178% increase in the number of regions that Rfam covers, which contrasts with the rather modest increase in the size of Rfamseq by 40%. This has caused some of our alignments to become very large. For example, the tRNA full alignment now contains more than 1 million sequences. The amount of compute required for this release was roughly 5 CPU months to calibrate the models, 1 CPU year to run blast, 3 CPU years to run CM-searches (cmsearch) and 15 CPU days to produce CM-derived multiple sequence alignments (cmalign).

Evaluating the success of the Wikipedia community annotation model

One of the fundamental problems facing any biocuration effort is keeping the annotation of the entities stored in a database up to date with the current literature. Typically, the annotation of existing entries changes less quickly than new data are added, so entries become rapidly out-of-date.

In mid-2007, Rfam began experimenting with using Wikipedia as a means for storing and curating the textual annotation of RNA families. Three years on, the RNA family pages have received more than 9000 edits from more than 1000 unique users. Slightly over 1% of these edits have been recognized as possible vandalism (Figure 1). The resulting marked-up annotation and curated references has dramatically improved the content of the Rfam database compared with the pre-2007 static text. The Wikipedia entries also help drive users to the Rfam website. Approximately 15% of all the web-traffic to http://rfam.sanger.ac.uk now comes via Wikipedia. As has been observed by others, a typical Google search for a biological term returns a Wikipedia entry among the top hits (12,13). From a curator's viewpoint, Wikipedia is an excellent model to take advantage of as it includes a large community of contributors and comes with a number of user-friendly tools that help with basic editing, maintaining references and automated updates to pages with programs called bots. The large community also has other benefits, such as the well documented long-tail effect, where the majority of new content is added by a large number of editors, each of whom makes just a few edits (12,13). There are also dedicated editors who are obsessed with small but important details that an average curator may not have time to attend to, such as consistency of style, grammar and spelling. There are also editors who are dedicated to reverting obvious non-constructive edits, commonly referred to as 'vandalism', which are usually recognized and reverted within seconds. It is important to note that all edits are reviewed before appearing on the Rfam website, so the amount of overt vandalism reaching Rfam is 0. Given our positive experiences, we can highly recommend other curation efforts turning to Wikipedia for their annotation. However, it must be borne in mind that Wikipedia is built by consensus and to gain its benefits you will lose the tight control of the data allowed by in-house curation.

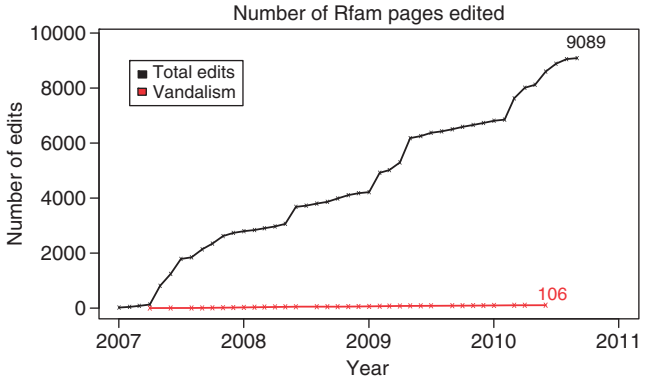

Figure 1. Edits for Wikipedia articles on RNA families. The cumulative number of edits since 1st January 2007 for the 733 Wikipedia articles that are associated with Rfam entries is shown in black. The total number of edits that were reverted or labeled as vandalism is shown in red. To mid-2010, there were just 106 of these. However, some reverted edits may have been well-intentioned but were deemed inappropriate for Wikipedia.

Rfam clans

One of the fundamental quality control steps that Rfam employs is that no two families can annotate the same nucleotide. This rule prevents us building two or more families for essentially the same entity. When building new Rfam families or extending an existing family, we sometimes find ourselves artificially increasing the threshold to avoid overlaps with another family or trimming the ends of families that have incorrect boundaries. We also find that a single alignment may not capture all the diversity of a group of homologous RNAs. To resolve some of these issues, we have borrowed the concept of a clan from the MEROPS and Pfam databases (14,15).

We have added 99 clans for the Rfam 10.0 release. These clans describe explicit relationships between families that either clearly share a common ancestor but are too divergent to be reasonably aligned or groups of families that could be aligned, but have clearly distinct functions and therefore should be kept as separate families. For example, the RNase P clan contains five homologous families RNase MRP, archeal RNase P, nuclear RNase P and the bacterial RNase P, types a and b. These RNAs are ribozymes involved in processing of pre-tRNA and pre-rRNA sequences. The RNase Ps are, however, notoriously difficult to align to each other. Furthermore, RNase P and RNase MRP are functionally distinct molecules (16). Another clan of interest is Glm; this clan contains two homologous but functionally distinct bacterial small RNAs, GlmY and GlmZ, which act in a hierarchical fashion to regulate the translation of the *glmS* coding gene. GlmY activates expression of GlmZ which in turn de-sequesters the GlmS Shine-Dalgarno

sequence via an anti-antisense interaction (17). The new clans mean that some of the internal quality control measures that Rfam uses can be relaxed for the clanned families. Primarily this means we can ignore our no-overlap rule, which has meant that in the past some of these families have had artificially high thresholds to avoid overlapping a related but distinct family.

In order to help assess the likelihood of a relationship between two or more families, we used a number of independent lines of evidence. These included sequence analysis based upon a SCOOP-like analysis for comparing overlapping hits from both profile hidden Markov model (HMM) and covariance model searches (18), the profile-profile comparison tool PRC (19) and literature searches for functional and evolutionary relationships. For the snoRNA and miRNA families, we were able to utilize some additional sources of information in order to establish homology. For the snoRNAs, we used some of the specialized snoRNA databases to confirm whether families targeted orthologous regions of rRNA, for many snoRNAs this helped to confirm a relationship between the families (20–23). For the miRNAs, we used the annotated seed region of the mature miRNA (24). If two or more miRNA families shared a significant amount of similarity in the seed region, and if they had further similarities identified by the sequence analysis tools, then these too were added to clans.

Species labels

The new set of seed and full alignments available via the website use descriptive species labels for sequence names rather than the more cryptic EMBL accessions and coordinates that were previously provided. The provenance of the sequence data is maintained by using '#=GS' tags from Stockholm format (25) to provide a mapping back to EMBL accessions (Figure 2). Stockholm is a versatile markup format for biological sequence alignments. It allows the markup of general file information, including references, comments and cross-links. It also allows the mark-up of regions of an alignment that cannot be aligned with tildes in the '#=GC RF' lines.

Ontologies

An important feature for any biocuration effort is linking to related resources, for example, primary sequence resources databases, genomes and to specialized resources such as miRBase and the snoRNA databases. Recently, a number of groups have started developing controlled vocabularies for describing biological entities. Two efforts of particular relevance to Rfam are the sequence ontology (SO) and the gene ontology (GO) (26,27). For the majority of Rfam families, we have now added cross-links to both the SO and the GO. Many of these were provided by researchers at the functional RNA database (28). In the near future, we plan to introduce more ncRNA terms back into the ontologies. Until then the mapping will remain rather coarse-grained and closely related to the existing types Rfam uses as annotation (6). This mapping groups the RNAs into three main groups: 'cis-reg', 'gene' and 'intron' with subtypes such as 'riboswitch', 'miRNA' and 'snoRNA'.

```
# STOCKHOLM 1.0
#=GF ID    UPSK
#=GF AC    RF00390
#=GF DE    UPSK RNA
#=GF AU    Moxon SJ
#=GF GA    27.0
#=GF SE    PMID:9223489, INFERNAL
#=GF SS    Published; PMID:9223489
#=GF TP    Cis-reg;
#=GF RN    [1]
#=GF RM    9223489
#=GF RT    The role of the pseudoknot at the 3' end of
#=GF RT    turnip yellow mosaic virus RNA in minus-strand
#=GF RT    synthesis by the viral RNA-dependent
#=GF RT    RNA polymerase.
#=GF RA    Deiman BA, Kortlever RM, Pleij CW;
#=GF RL    J Virol 1997;71:5990-5996.
#=GF WK    http://en.wikipedia.org/wiki/UPSK_RNA
#=GF SQ    6

#=GS TYMV.1 AC    M24801.1/79-101
#=GS TYMV.2 AC    AF035635.1/619-641
#=GS TYMV.3 AC    M24804.1/82-104
#=GS TYMV.4 AC    M24805.1/67-89
#=GS TYMV.5 AC    J04373.1/6212-6234
#=GS TYMV.6 AC    M24803.1/1-23

TYMV.1             UAAGUUCUCGAUCUUUAAAAUCG
TYMV.2             UGAGUUCUCGAUCUCUAAAAUCG
TYMV.3             UGAGUUCUCUAUCUCUAAAAUCG
TYMV.4             UGAGUUCUCGAUCUCUAAAAUCG
TYMV.5             UAAGUUCUCGAUCUUUAAAAUCG
TYMV.6             UAAGUUCUCGAUCUCUAAAAUCG
#=GC SS_cons       .AAA....<<<<aaa....>>>>
#=GC RF            UaAGUUCUCGauCUcUAAAauCG
//
```

Figure 2. An example Stockholm alignment for the UPSK pseudoknot from turnip yellow mosaic virus. The Stockholm alignment format is flexible enough to allow generic mark-up of file information with '#=GF' lines, sequence information with '#=GS' lines and column information with '#=GC' lines. Each is followed by at least a two-letter code giving an indication for what follows e.g. 'ID' implies 'identifier', 'AC' implies 'accession', 'AU' implies 'author', etc. All the commonly used tags are documented in the Wikipedia article for Stockholm alignment (25).

Future developments

New families in Rfam 10.1. For the forthcoming minor release of Rfam, we have added a number of new and notable families. Of particular note are the direct submissions of Stockholm formatted alignments and corresponding Wikipedia articles from the RNA community via the RNA families track at RNA Biology (8). This track has released much of the burden of building these new families from our curators, and the families produced have been built and annotated by experts and are therefore of high quality. Updated families from this route include RNase MRP, SRP, tmRNA and the U3 snoRNA (29–32). In addition, several families missing from past Rfam releases have been published, including the SmY RNA, the cyanobacterial RNA Yfr2, several Trypanosomatid snoRNAs, the self-splicing ribozyme GIR1, an influenza pseudoknot, the Staphylococcus small RNA RsaOG and a putative RNA antitoxin, ptaRNA1 (33–39). The ptaRNA1 article alerted us to the fact that Rfam contains none of the published and well-characterized RNA antitoxins such as sok and symE (40). These

omissions will be remedied in Rfam 10.1. A growing class of *cis*-regulatory elements are the environmental sensors. These are generally structured 5′ UTR elements that change conformation in response to environmental changes such as temperature or pH; this change subsequently influences the expression of the protein encoded in the host mRNA. We have added the first examples of a cold sensor and a pH sensor (41,42). Finally, we have received a dramatic number of submissions from a recent bioinformatic screen that was followed by a thorough analysis of the predictions largely based upon genomic context. This has resulted in more than 80 new additions to the database (43). Fortunately, the authors kindly provide both Stockholm formatted alignments and Wikipedia articles for these new families.

Covariance model pre-filters. A pressing issue for Rfam is the replacement of WU-BLAST as a pre-filter for searching the Rfamseq database. The legal rights to up-to-date versions of WU-BLAST were recently acquired by a commercial entity and the software can no longer be considered free in any meaningful sense. However, there have been several developments that should allow profile HMMs to be used as effective pre-filters for covariance model searches (44). Accelerated profile HMM searches are now available through the HMMER package (45–47). In the near future, Rfam will therefore be in a position to replace the current BLAST-based filters with accelerated profile HMMs.

Scale. Sequencing projects such as the Genome 10K (48) and other attempts to fill sequencing gaps in the tree of life (49) mean that most Rfam families will dramatically increase in depth in the near future. Large alignments already pose a considerable challenge when it comes to displaying or distributing the alignments themselves, or building and displaying related data such as species and phylogenetic trees. Novel techniques will need to be developed in order to deal with these and many other issues of scale. We look forward to working with the wider community to develop these new tools and techniques.

ACKNOWLEDGEMENTS

Many thanks to Guy Coates, James Beal and Peter Clapham for assistance with improving the performance of computational and software infrastructure. The authors received invaluable feedback at the 2009 Benasque RNA Workshop.

FUNDING

Wellcome Trust (grant number WT077044/Z/05/Z) (to P.P.G., J.D., J.T., I.H.O., B.M. and A.B.); Howard Hughes Medical Institute (R.D.F, E.P.N., D.L.K. and S.R.E); University of Manchester (S.G.J.). Funding for open access charge: The Wellcome Trust (grant number WT077044/Z/05/Z).

Conflict of interest statement. None declared.

REFERENCES

1. Holmes,I. (2004) A probabilistic model for the evolution of RNA structure. *BMC Bioinformatics*, **5**, 166.
2. Do,C.B., Woods,D.A. and Batzoglou,S. (2006) CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, **22**, e90–e98.
3. Yao,Z., Weinberg,Z. and Ruzzo,W.L. (2006) CMfinder–a covariance model based RNA motif finding algorithm. *Bioinformatics*, **22**, 445–452.
4. Sun,Y. and Buhler,J. (2008) Designing secondary structure profiles for fast ncRNA identification. *Comput. Syst. Bioinformatics Conf.*, **7**, 145–156.
5. Yusuf,D., Marz,M., Stadler,P.F. and Hofacker,I.L. (2010) Bcheck: a wrapper tool for detecting RNase P RNA genes. *BMC Genomics*, **11**, 432.
6. Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
7. 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 Res.*, **33(Database issue)**, D121–D124.
8. Gardner,P.P., Daub,J., Tate,J.G., Nawrocki,E.P., Kolbe,D.L., Lindgreen,S., Wilkinson,A.C., Finn,R.D., Griffiths-Jones,S., Eddy,S.R. *et al.* (2009) Rfam: updates to the RNA families database. *Nucleic Acids Res.*, **37(Database issue)**, D136–D1340.
9. Nawrocki,E.P., Kolbe,D.L. and Eddy,S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.
10. Leinonen,R., Akhtar,R., Birney,E., Bonfield,J., Bower,L., Corbett,M., Cheng,Y., Demiralp,F., Faruque,N., Goodgame,N. *et al.* (2010) Improvements to services at the European Nucleotide Archive. *Nucleic Acids Res.*, **38(Database issue)**, D39–D45.
11. Eddy,S.R. (2002) A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, **3**, 18.
12. Huss,J.W., Orozco,C., Goodale,J., Wu,C., Batalov,S., Vickers,T.J., Valafar,F. and Su,A.I. (2008) A gene wiki for community annotation of gene function. *PLoS Biol.*, **6**, e175.
13. Huss,J.W., Lindenbaum,P., Martone,M., Roberts,D., Pizarro,A., Valafar,F., Hogenesch,J.B. and Su,A.I. (2010) The Gene Wiki: community intelligence applied to human gene annotation. *Nucleic Acids Res.*, **38(Database issue)**, D633–D639.
14. Rawlings,N.D. and Barrett,A.J. (1993) Evolutionary families of peptidases. *Biochem. J.*, **290(Pt 1)**, 205–218.
15. Finn,R.D., Mistry,J., Schuster-Böckler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34(Database issue)**, D247–D251.
16. Ellis,J.C. and Brown,J.W. (2009) The RNase P family. *RNA Biol.*, **6**, 362–369.
17. Urban,J.H. and Vogel,J. (2008) Two seemingly homologous noncoding RNAs act hierarchically to activate glmS mRNA translation. *PLoS Biol.*, **6**, e64.
18. Bateman,A. and Finn,R.D. (2007) SCOOP: a simple method for identification of novel protein superfamily relationships. *Bioinformatics*, **23**, 809–814.
19. Madera,M. (2008) Profile Comparer: a program for scoring and aligning profile hidden Markov models. *Bioinformatics*, **24**, 2630–2631.
20. Samarsky,D.A. and Fournier,M.J. (1999) A comprehensive database for the small nucleolar RNAs from Saccharomyces cerevisiae. *Nucleic Acids Res.*, **27**, 161–164.
21. Brown,J.W., Echeverria,M., Qu,L.H., Lowe,T.M., Bachellerie,J.P., Hüttenhofer,A., Kastenmayer,J.P., Green,P.J., Shaw,P. and Marshall,D.F. (2003) Plant snoRNA database. *Nucleic Acids Res.*, **31**, 432–435.
22. Li,S.G., Zhou,H., Luo,Y.P., Zhang,P. and Qu,L.H. (2005) Identification and functional analysis of 20 Box H/ACA small nucleolar RNAs (snoRNAs) from Schizosaccharomyces pombe. *J. Biol. Chem.*, **280**, 16446–16455.
23. Lestrade,L. and Weber,M.J. (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res.*, **34(Database issue)**, D158–D162.

24. Griffiths-Jones,S., Saini,H.K., van Dongen,S. and Enright,A.J. (2008) miRBase: tools for microRNA genomics. *Nucleic Acids Res.*, **36(Database issue)**, D154–D158.
25. Stockholm format. http://en.wikipedia.org/wiki/Stockholm_format Stockholm format (19 June 2010, date last accessed).
26. Eilbeck,K., Lewis,S.E., Mungall,C.J., Yandell,M., Stein,L., Durbin,R. and Ashburner,M. (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.*, **6**, R44.
27. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
28. Mituyama,T., Yamada,K., Hattori,E., Okida,H., Ono,Y., Terai,G., Yoshizawa,A., Komori,T. and Asai,K. (2009) The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs. *Nucleic Acids Res.*, **37(Database issue)**, D89–D92.
29. Dávila López,M., Rosenblad,M.A. and Samuelsson,T. (2009) Conserved and variable domains of RNase MRP RNA. *RNA Biol.*, **6**, 208–220.
30. Rosenblad,M.A., Larsen,N., Samuelsson,T. and Zwieb,C. (2009) Kinship in the SRP RNA family. *RNA Biol.*, **6**, 508–516.
31. Mao,C., Bhardwaj,K., Sharkady,S.M., Fish,R.I., Driscoll,T., Wower,J., Zwieb,C., Sobral,B.W. and Williams,K.P. (2009) Variations on the tmRNA gene. *RNA Biol.*, **6**, 355–361.
32. Marz,M. and Stadler,P.F. (2009) Comparative analysis of eukaryotic U3 snoRNA. *RNA Biol.*, **6**, 503–507.
33. Jones,T.A., Otto,W., Marz,M., Eddy,S.R. and Stadler,P.F. (2009) A survey of nematode SmY RNAs. *RNA Biol.*, **6**, 5–8.
34. Gierga,G., Voss,B. and Hess,W.R. (2009) The Yfr2 ncRNA family, a group of abundant RNA molecules widely conserved in cyanobacteria. *RNA Biol.*, **6**, 222–227.
35. Doniger,T., Michaeli,S. and Unger,R. (2009) Families of H/ACA ncRNA molecules in trypanosomatids. *RNA Biol.*, **6**, 370–374.
36. Nielsen,H. and Johansen,S.D. (2009) Group I introns: moving in new directions. *RNA Biol.*, **6**, 375–383.
37. Gultyaev,A.P. and Olsthoorn,R.C. (2010) A family of non-classical pseudoknots in influenza A and B viruses. *RNA Biol.*, **7**, 125–129.
38. Marchais,A., Bohn,C., Bouloc,P. and Gautheret,D. (2010) RsaOG, a new staphylococcal family of highly transcribed non-coding RNA. *RNA Biol.*, **7**, 116–119.
39. Findeiss,S., Schmidtke,C., Stadler,P.F. and Bonas,U. (2010) A novel family of plasmid-transferred anti-sense ncRNAs. *RNA Biol.*, **7**, 120–124.
40. Fozo,E.M., Hemm,M.R. and Storz,G. (2008) Small toxic proteins and the antisense RNAs that repress them. *Microbiol. Mol. Biol. Rev.*, **72**, 579–589.
41. Giuliodori,A.M., Di Pietro,F., Marzi,S., Masquida,B., Wagner,R., Romby,P., Gualerzi,C.O. and Pon,C.L. (2010) The cspA mRNA is a thermosensor that modulates translation of the cold-shock protein CspA. *Mol. Cell.*, **37**, 21–33.
42. Nechooshtan,G., Elgrably-Weiss,M., Sheaffer,A., Westhof,E. and Altuvia,S. (2009) A pH-responsive riboregulator. *Genes Dev.*, **23**, 2650–2662.
43. Weinberg,Z., Wang,J.X., Bogue,J., Yang,J., Corbino,K., Moy,R.H. and Breaker,R.R. (2010) Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome. Biol.*, **11**, R31.
44. Weinberg,Z. and Ruzzo,W.L. (2006) Sequence-based heuristics for faster annotation of non-coding RNA families. *Bioinformatics*, **22**, 35–39.
45. Eddy,S.R. (2008) A probabilistic model of local sequence alignment that simplifies statistical significance estimation. *PLoS Comput. Biol.*, **4**, e1000069.
46. Eddy,S.R. (2009) A new generation of homology search tools based on probabilistic inference. *Genome Inform.*, **23**, 205–211.
47. Johnson,L.S., Eddy,S.R. and Portugaly,E. (2010) Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics*, **11**, 431.
48. Genome 10K Community of Scientists, C. (2009) Genome 10K: a proposal to obtain whole-genome sequence for 10,000 vertebrate species. *J. Hered.*, **100**, 659–674.
49. Wu,D., Hugenholtz,P., Mavromatis,K., Pukall,R., Dalin,E., Ivanova,N.N., Kunin,V., Goodwin,L., Wu,M., Tindall,B.J. *et al.* (2009) A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature*, **462**, 1056–1060.
50. Brown,J.W., Birmingham,A., Griffiths,P.E., Jossinet,F., Kachouri-Lafond,R., Knight,R., Lang,B.F., Leontis,N., Steger,G., Stombaugh,J. *et al.* (2009) The RNA structure alignment ontology. *RNA*, **15**, 1623–1631.