The use of covariance models to annotate RNAs in whole genomes

Paul P. Gardner

Abstract

In this review we discuss bioinformatic issues in non-coding RNA analysis, in particular the annotation of genome sequences using covariance models. Some recent innovations for improving the speed and accuracy of covariance models is discussed.

Keywords: *ncRNA; Rfam; Covariance Models*

INTRODUCTION

It is now well established that non-coding RNAs (ncRNAs) play a central role in important molecular processes across all kingdoms of life. Essential processes such as splicing, translation and gene regulation are dependent on the specialised functions and regulatory roles of RNA molecules. For example, the splicing and processing of eukaryotic messenger RNAs (mRNAs) is depends on the spliceosomal RNAs U1, U2, U4, U5 and U6 [1]. The translation of all mRNAs into proteins depends on transfer RNA (tRNA) and the ribosomal RNA–protein complexes (RNPs). In turn, maturation of functional tRNA molecules relies upon a splicing reaction that is carried out by the RNase P RNP [2]. Furthermore, the production of a mature ribosome is dependent on other ncRNAs, such as RNase MRP and the various Small nucleolar ribonucleic acids (snoRNAs) found in eukaryotes and archaea, which direct post-transcriptional modifications such as methylations, pseudouridylations and cleavages [3]. More exotic forms of splicing are carried out or regulated by ncRNAs such as Sm Y RNA found in nematodes [4] and the self-splicing introns found in most major lineages [5]. Finally, hordes of RNA are involved in the regulation of gene expression, these range from the cis-regulatory ncRNAs such as iron response element (IRE), Histone3, IRES and riboswitches to the trans-regulatory elements such as the eukaryotic miRNAs and bacterial 6S RNA and OxyS RNA [6].

There are many human diseases that have been linked to the aberrant production of ncRNAs and RNPs. A well-studied example is Prader–Willi syndrome (PWS), which is a genetic disorder resulting from a deletion of an imprinted locus on chromosome 15. PWS has been linked to the deletion of the C/D box snoRNA SNORD116 (also known as HBII-85) cluster [7–9]. This enigmatic RNA has no known function, yet has been linked to the regulation of alternative splicing [10]. Several RNAs have been implicated in the progression of human cancer. The Y RNA family is important for initiating DNA replication, while the telomerase RNAs are important for extending telomeres at chromosomal terminii. Both of these families are highly expressed in tumour tissues and possibly contribute to the disease [11–14]. The miRNAs, which have been shown to be central players in gene regulation, have been shown to undergo changes in expression in cancerous tissues [15]. Furthermore, some genetic variation within miRNA sequence have been linked to an increased susceptibility to cancer [16, 17]. MicroRNAs have also been linked to other diseases: it appears that variation within the seed region of miR-96 results in progressive hearing loss in both human and mouse models [18, 19]. Finally, microRNAs have been implicated in a variety of

Corresponding author. Paul P. Gardner, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK. Tel: +44 (0) 1223 494726; Fax: +44 (0) 1223 494 010; E-mail: pg5@sanger.ac.uk

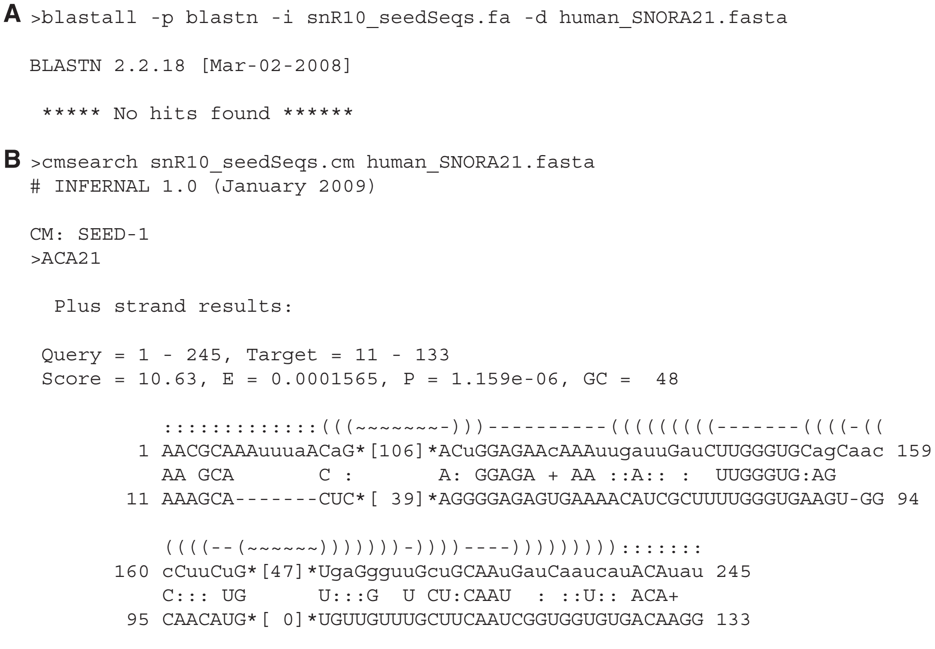
Paul P. Gardner is a senior scientist at the Wellcome Trust Sanger Institute where he maintains the Rfam database. He has worked as a post-doc at the Universities of Copenhagen and Bielefeld. He has a PhD in Biomathematics from Massey University in New Zealand.

virus–host interactions. For example, the expression of human miR-122 is required for infection by hepatitis C virus and the human immunodeficiency virus miRNA, TAR, is required for infection by HIV [20, 21].

In this review we discuss bioinformatic issues in ncRNA analysis, in particular the annotation of genome sequences. At present the options are rather limited in this field. Analysis tools for this purpose generally fall into one of the two categories: there are a few algorithms that are specialised for a minority of RNA families such as tRNA, C/D box snoRNAs and rRNA, or one can use the general purpose option which is to use profile stochastic context-free grammars, otherwise known as covariance models (CMs), and a large library of alignments and secondary structures of known ncRNAs, such as those provided by the Rfam database.

RNA INFORMATICS

Bioinformatic resources for ncRNAs must overcome a number of issues. The primary sequence encoding ncRNAs is frequently poorly conserved, therefore classical bioinformatic tools such as the sequence-based homology search tools BLAST, FASTA and SSEARCH do not perform well [22] (see Figure 1 for an example). The number of possible secondary structures for a given RNA sequence grows exponentially (1.8^N) with the length of the sequence; therefore, selecting a biologically active secondary structure from this large ensemble can be challenging [23]. There has been some progress in this field using evolutionary information to prune the structures that are not supported by sequence variation [24–26]. To date, there are relatively few known RNA tertiary structures and the ones that have been determined to cover just 20 RNA families in the Rfam database. These include several riboswitches, a handful of cis-regulatory elements such as IRE and a few of the classical ncRNA genes such as tRNA, rRNA, RNase P and signal recognition particle and the self-splicing group I intron. Some attempts at *de novo* ncRNA gene prediction have been made [27–29], but this is an extraordinarily difficult task, given that the statistical signals from sequence analysis are generally heterogeneous across both species and RNA families. There are no start and stop codons for



```
A >blastall -p blastn -i snR10_seedSeqs.fa -d human_SNORA21.fasta

  BLASTN 2.2.18 [Mar-02-2008]

   ***** No hits found ******

B >cmsearch snR10_seedSeqs.cm human_SNORA21.fasta
  # INFERNAL 1.0 (January 2009)

  CM: SEED-1
  >ACA21

    Plus strand results:

   Query = 1 - 245, Target = 11 - 133
   Score = 10.63, E = 0.0001565, P = 1.159e-06, GC =  48

             :::::::::::::(((~~~~~~~~-)))----------(((((((((-------((((-((
           1 AACGCAAAuuuaACaG*[106]*ACuGGAGAAcAAAuugauuGauCUUGGGUGCagCaac 159
             AA GCA       C :       A: GGAGA + AA ::A:: :  UUGGGUG:AG
          11 AAAGCA-------CUC*[ 39]*AGGGGAGAGUGAAAACAUCGCUUUUGGGUGAAGU-GG 94

             ((((--(~~~~~~)))))))-))))----)))))))))):::::::
         160 cCuuCuG*[47]*UgaGgguuGcuGCAAuGauCaaucauACAuau 245
             C::: UG      U:::G  U CU:CAAU  : ::U:: ACA+
          95 CAACAUG*[ 0]*UGUUGUUUGCUUCAAUCGGUGGUGUGACAAGG 133
```

Figure 1: The H/ACA box snoRNA snR10 is important for guiding a psuedouridylation on rRNA in yeasts. It has been identified as an orthologue of SNORA21, a vertebrate H/ACA box [63, 64]. (**A**) Using NCBI-BLAST with default parameters, we find no areas of homology between the Rfam snR10 sequences used in the seed alignment and the human SNORA21 sequence. (**B**) Using a covariance model built from snR-10 sequence, the orthologous human sequence is readily identified with a significant score (*E*-value = 0.0001565). Since an *E*-value of 0.00015 means that a homolog as good as the one detected is expected one time in ten thousand in a random database [65], we can be fairly confident that the CM prediction represents a true homologue. The alignment is shown in a similar format to that used by BLAST, the chief difference being that this alignment is augmented with secondary structure information where matching parentheses indicate an RNA base pair.

open reading frame prediction, nor is there the biased codon usage pattern that is commonly used for *de novo* protein-coding gene prediction. In some limited cases, such as A + T-rich hyperthermophiles, there is a statistically significant signal from G + C content and that can be used [30, 31]. However, generally, there is not enough sequence or structure signal within single gene sequences for RNA gene prediction [32, 33]. There have been attempts at RNA gene prediction based upon evolutionary signals from sequence and structure conservation in genome sequence alignments [27–29], but the success of these has been varied, chiefly hampered by the high false-positive rate of these methods [34, 35].

COVARIANCE MODELS

A success story in the RNA field is the use of CMs for ncRNA homology search [36, 37]. CMs are a natural extension of profile hidden Markov models (pHMMs), which have been successfully applied to the protein world [38]. Profiles are generated from 'seed' alignments, which are alignments of representative members of a family of homologous sequences. These profiles can be used to automatically annotate other sequences as being either related or unrelated to the members of the seed. In the following discussion we will broadly outline the procedure.

The profile is generated from the seed alignment by reducing each column in the seed alignment to a vector of frequencies (probabilities) for each possible residue. The probabilities for each residue x_i corresponding to a given sequence X are multiplied together to calculate the likelihood that the same processes that produced the seed alignment would have produced X. This scoring, of course, assumes that the sequence is aligned to the profile in a reasonable way.

There are some standard computational algorithms for doing this: one that maximises the probability is called the Viterbi algorithm [39], while another, which sums the probabilities of all possible ways of aligning the sequence to the profile, is called the Forward algorithm [40]. The pHMM approach can be further empowered by using the insertion/deletion ('indel') information in the seed alignment to model explicitly insert and delete events, allowing for an increase in the probability of entering insert or delete states near the boundaries of 'gappy' columns in the seed alignment.

A limited sample of sequences in the seed may lead to a too narrow view of the possible nucleotides at each position, Dirichlet-mixture priors and entropy-weighting [41] adjust probabilities to reflect the possibility of homologues that have features not observed in the training set. Over-represented seed sequences could skew probabilities towards a biased sq of sequences, however, these can be down-weighted using tree weighting schemes [42].

All of these pHMM concepts can be translated into the ncRNA domain by explicitly adding information about RNA secondary structure that then allows for modelling the structural constraints on RNA nucleotides. However, the algorithms are now much more computationally intensive, since the sequence database is now being searched for matches to a tree-like data-structure, which represents secondary structure, as opposed to a linear data-structure that represents a pHMM [40]. The class of algorithm used by CMs is no longer a PHMM; a profile stochastic context-free grammar takes its place. CMs have been shown to be accurate models for the ncRNA homology search problem, even when predicted alignments and secondary structures are used as input [22]. Recent improvements in one of the main software package, Infernal, have resulted in dramatic improvements in both the speed and accuracy of CMs [43–47]. These include using pHMMs as a 'fast' pre-filter that allows the method to skip low-scoring sequences before running the more intensive CM calculation [44, 48, 49]. Further, speed improvements can be made by using an approach called 'query-dependent banding' that allows a large proportion of the dynamic programming matrix to be pruned away [45]. The scoring scheme has also been improved by incorporating the Dirichlet-mixture prior approach originally developed for pHMMs that I mentioned earlier. This has resulted in a great improvement in the ability of CM searches to detect remote homologues [45]. The CM concept has been improved to allow for truncated sequences such as those one might expect from meta-genomic projects or processed transcripts, e.g. mature miRNAs scored with a model of the pre-miRNA [47]. This useful method came about by modifying the CYK algorithm (the CM equivalent of the Viterbi algorithm for pHMMs) and has resulted in a further improvement in sensitivity [47]. These improvements and more are outlined in further detail in the recent Infernal 1.0 publication [46].

The CM approach has been successfully used by the Rfam database for RNA sequence annotation for many years [50–52]. This database curates large numbers (1372 for the January 2009 Rfam 9.1 release) of trusted seed alignments, which are hand-curated ncRNA sequence alignments and secondary structures. In addition to the alignments, there are various value-added data available for each family: the families link prominently to external sources of data, such as the source alignments and structures from the literature or other databases; a short text describing the family is provided using Wikipedia [53]; the families are systematically given unique names; and literature references are curated. A 'full' alignment is automatically generated by searching a huge sequence database, derived from the EMBL nucleotide database. For Rfam 9.0 and 9.1, this sequence database contained more than 121 gigabases derived from more than 29.5 million sequences. More than 180 million nucleotides derived from 1 million regions are contained in the full alignments. On occasion an iterative refinement process is used to improve the coverage of each family; sequences that score above a curated threshold are automatically aligned to the seed CM, in order to form the full alignment (Figure 2).

In addition to facilitating the building of RNA families, the Rfam approach can be used to annotate genomic sequences. These annotations may be provided directly by Rfam, which currently curates 1140 viral, archaeal, bacterial and eukaryotic genome annotations. Alternatively, users can annotate their own genome sequences by downloading all the sequences and CMs and running an Infernal-based annotation pipeline themselves. An option for smaller projects is to use the website batch search facility. A number of independent genome annotation groups use Rfam models for annotating their sequences. Genome Reviews, for example, annotates 883 completed genomes from prokaryotes and selected eukaryotes such as *Saccharomyces cerevisiae* (Release 108.0, 7 July 2009) [54], and ENSEMBL uses Rfam models (excluding the cis-regulatory elements) for annotating vertebrate genome sequences [55].

A few examples of past successful genome annotation projects that used Rfam were chicken [56], 12 *Drosophila* [57], mouse [58], human chromosome 1 [59] and *Aspergillus* [60]. The chicken genome is a particularly interesting example as it has a surprising paucity of ncRNA-derived repeat elements compared with other sequenced vertebrate genomes. This has raised the possibility that genomes such as this could be used to discriminate ncRNA-derived pseudogenes from functional ncRNAs using synteny information, however, this hypothesis has yet to be proven. The study of 12 *Drosophila* genomes was extremely comprehensive, combining Rfam annotations, *de novo* predictions and expression data, showing overlaps between all these datasets. The Rfam annotations for *Aspergillus*, mouse and human chromosome 1 results were integrated into the gene-count summaries for these species. The mouse ncRNA genome analysis also discovered a number of ncRNA-derived repeat elements.

CONCLUDING REMARKS

The Rfam approach to genome annotation currently provides the only available comprehensive resource for detecting known ncRNAs on a large scale. The three main options are to run algorithms specialised for a limited number of families, run a generic *de novo* prediction tool or run Rfam. The specialist tools such as tRNAscan-SE, RNAMMER, snoscan, RNAmicro infer sequences belonging to the RNA families tRNA, rRNA, C/D box snoRNA and microRNA, respectively, these methods are generally very accurate. The *de novo* prediction tools QRNA, RNAz and EvoFold attempt to find conserved and structured regions in genomic alignments; whilst all these methods have made useful predictions in the past, they do seem to have rather high false-positive rates. Therefore, for current genome annotation projects a combination of all these approaches will provide the most useful information. The UCSC genome browser [61] hosted by the Functional RNA Database 3.0 is a useful example of pooling RNA-centric annotations for a small number of genomes [62].

Key Points

- How ncRNAs are important for biological processes and some recent discoveries with implications for human health are discussed.
- The current state of bioinformatic resources for ncRNA research is outlined.
- pHMMs that are frequently used for protein homology searches are outlined.
- The pHMM discussion is used as a grounding for discussing the more complex class of methods called CMs.
- Some recent improvements for CM methodologies are discussed.

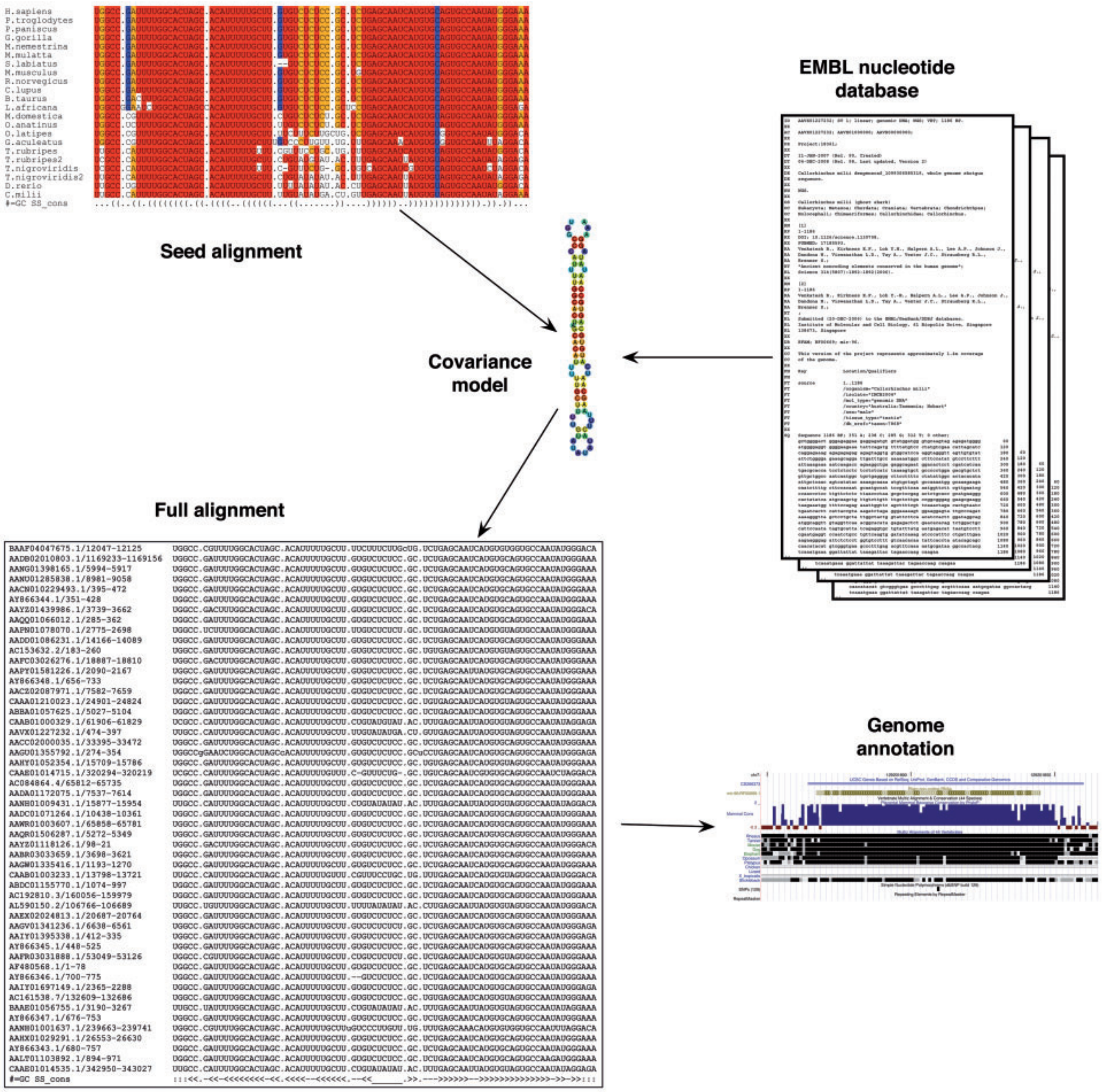


Figure 2: An illustration of the Rfam annotation pipeline using the miR-96 family as an example. A hand-curated seed alignment is used to search the EMBL nucleotide database using an Infernal covariance model, the resulting hits are used to automatically generate a full alignment that can potentially be used for genome annotation, illustrated in this case using the UCSC genome browser.

Acknowledgements

The authors heartfelt thanks go to Jennifer Daub and John Tate for assisting with translating the manuscript.

References

1. Dávila López M, Rosenblad MA, Samuelsson T. Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components. *Nucleic Acids Res* 2008;**36**(9):3001–10.
2. López MD, Rosenblad MA, Samuelsson T. Conserved and variable domains of RNase MRP RNA. *RNA Biol* 2009;**6**(3):208–20.
3. Decatur WA, Liang XH, Piekna-Przybylska D, Fournier MJ. Identifying effects of snoRNA-guided modifications on the synthesis and function of the yeast ribosome. *Methods Enzymol* 2007;**425**:283–316.
4. Jones TA, Otto W, Marz M, *et al*. A survey of nematode SmY RNAs. *RNA Biol* 2009;**6**:5–8.
5. Mattick JS. Introns: evolution and function. *Curr Opin Genet Dev* 1994;**4**(6):823–31.

6. Bompfünewerer AF, Flamm C, Fried C, *et al*. Evolutionary patterns of non-coding RNAs. *Theory Biosci* 2005; **123**(4):301–69.

7. Skryabin BV, Gubar LV, Seeger B, *et al*. Deletion of the MBII-85 snoRNA gene cluster in mice results in postnatal growth retardation. *PLoS Genet* 2007;**3**(12):e235.

8. Sahoo T, del Gaudio D, German JR, *et al*. Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. *Nat Genet* 2008; **40**(6):719–21.

9. Ding F, Li HH, Zhang S, *et al*. SnoRNA Snord116 (Pwcr1/MBII-85) deletion causes growth deficiency and hyperphagia in mice. *PLoS One* 2008;**3**(3):e1709.

10. Bazeley PS, Shepelev V, Talebizadeh Z, *et al*. snoTARGET shows that human orphan snoRNA targets locate close to alternative splice junctions. *Gene* 2008;**408**(1–2):172–9.

11. Avilion AA, Piatyszek MA, Gupta J, *et al*. Human telomerase RNA and telomerase activity in immortal cell lines and tumor tissues. *Cancer Res* 1996;**56**(3):645–50.

12. González-Suárez E, Samper E, Flores JM, Blasco MA. Telomerase-deficient mice with short telomeres are resistant to skin tumorigenesis. *Nat Genet* 2000;**26**:114–17.

13. Li Y, Li H, Yao G, *et al*. Inhibition of telomerase RNA (hTR) in cervical cancer by adenovirus-delivered siRNA. *Cancer Gene Ther* 2007;**14**(8):748–55.

14. Christov CP, Trivier E, Krude T. Noncoding human Y RNAs are overexpressed in tumours and required for cell proliferation. *Br J Cancer* 2008;**98**(5):981–8.

15. Lu J, Getz G, Miska EA, *et al*. MicroRNA expression profiles classify human cancers. *Nature* 2005;**435**(7043):834–8.

16. Calin GA, Ferracin M, Cimmino A, *et al*. A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *N Engl J Med* 2005; **353**(17):1793–801.

17. Calin GA, Dumitru CD, Shimizu M, *et al*. Frequent deletions and down-regulation of micro- RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. *Proc Natl Acad Sci USA* 2002;**99**(24):15524–9.

18. Mencía A, Modamio-Høybjør S, Redshaw N, *et al*. Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. *Nat Genet* 2009;**41**(5):609–13.

19. Lewis MA, Quint E, Glazier AM, *et al*. An ENU-induced mutation of miR-96 associated with progressive hearing loss in mice. *Nat Genet* 2009;**41**(5):614–18.

20. Jopling CL, Yi M, Lancaster AM, *et al*. Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. *Science* 2005;**309**(5740):1577–81.

21. Ouellet DL, Plante I, Landry P, *et al*. Identification of functional microRNAs released through asymmetrical processing of HIV-1 TAR element. *Nucleic Acids Res* 2008; **36**(7):2353–65.

22. Freyhult EK, Bollback JP, Gardner PP. Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res* 2007;**17**:117–25.

23. Hofacker IL, Schuster P, Stadler PF. Combinatorics of RNA secondary structures. *Discr Appl Math* 1996;**89**:207–37.

24. Hofacker IL, Fekete M, Stadler PF. Secondary structure prediction for aligned RNA sequences. *J Mol Biol* 2002; **319**(5):1059–66.

25. 25. Yao Z, Weinberg Z, Ruzzo WL. CMfinder—a covariance model based RNA motif finding algorithm. *Bioinformatics* 2006;**22**(4):445–52.

26. Torarinsson E, Lindgreen S. WAR: webserver for aligning structural RNAs. *Nucleic Acids Res* 2008;**36**(Web Server issue):W79–84.

27. Rivas E, Eddy SR. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* 2001;**2**:8.

28. Washietl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci USA* 2005;**102**(7):2454–9.

29. Pedersen JS, Bejerano G, Siepel A, *et al*. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput Biol* 2006;**2**(4):e33.

30. Klein RJ, Misulovin Z, Eddy SR. Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc Natl Acad Sci USA* 2002;**99**(11):7542–7.

31. Schattner P. Searching for RNA genes using base-composition statistics. *Nucleic Acids Res* 2002;**30**(9):2076–82.

32. Workman C, Krogh A. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res* 1999; **27**(24):4816–22.

33. Rivas E, Eddy SR. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* 2000;**16**(7):583–605.

34. Washietl S, Pedersen JS, Korbel JO, *et al*. Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res* 2007;**17**(6):852–64.

35. Babak T, Blencowe BJ, Hughes TR. Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC Bioinformatics* 2007;**8**:33.

36. Eddy SR, Durbin R. RNA sequence analysis using covariance models. *Nucleic Acids Res* 1994;**22**(11):2079–88.

37. Sakakibara Y, Brown M, Hughey R, *et al*. Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res* 1994;**22**(23):5112–20.

38. Krogh A, Brown M, Mian IS, *et al*. Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol* 1994;**235**(5):1501–31.

39. Viterbi A. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans Inf Theory* 1967;**13**(2):260–9.

40. Durbin R, Eddy SR, Krogh A, Mitchison G. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge: Cambridge University Press, 1999.

41. Sjölander K, Karplus K, Brown M, *et al*. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput Appl Biosci* 1996; **12**(4):327–45.

42. Gerstein M, Sonnhammer EL, Chothia C. Volume changes in protein evolution. *J Mol Biol* 1994;**236**(4):1067–78.

43. Eddy SR. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics* 2002;**3**:18.

44. Weinberg Z, Ruzzo WL. Exploiting conserved structure for faster annotation of non-coding RNAs without loss of accuracy. *Bioinformatics* 2004;**20**(Suppl 1):i334–41.

45. Nawrocki EP, Eddy SR. Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS Comput Biol* 2007; **3**(3):e56.

46. Nawrocki EP, Kolbe DL, Eddy SR. Infernal 1.0: inference of RNA alignments. *Bioinformatics* 2009;**25**(10):1335–7.
47. Kolbe DL, Eddy SR. Local RNA structure alignment with incomplete sequence. *Bioinformatics* 2009;**25**(10): 1236–43.
48. Weinberg Z, Ruzzo WL. Sequence-based heuristics for faster annotation of non-coding RNA families. *Bioinformatics* 2006;**22**:35–9.
49. Zhang S, Borovok I, Aharonowitz Y, *et al*. A sequence-based filtering method for ncRNA identification and its application to searching for riboswitch elements. *Bioinformatics* 2006;**22**(14):e557–65.
50. Griffiths-Jones S, Bateman A, Marshall M, *et al*. Rfam: an RNA family database. *Nucleic Acids Res* 2003;**31**:439–41.
51. Griffiths-Jones S, Moxon S, Marshall M, *et al*. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* 2005;**33**(Database issue):D121–4.
52. Gardner PP, Daub J, Tate JG, *et al*. Rfam: updates to the RNA families database. *Nucleic Acids Res* 2009;**37**(Database issue):D136–40.
53. Daub J, Gardner PP, Tate J, *et al*. The RNA WikiProject: community annotation of RNA families. *RNA* 2008; **14**(12):2462–4.
54. Sterk P, Kersey PJ, Apweiler R. Genome Reviews: standardizing content and representation of information about complete genomes. *OMICS* 2006;**10**(2):114–18.
55. Hubbard TJ, Aken BL, Ayling S, *et al*. Ensembl 2009. *Nucleic Acids Res* 2009;**37**(Database issue):D690–7.
56. International Chicken Genome Sequencing Consortium. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* 2004;**432**(7018):695–716.
57. Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the Drosophila phylogeny. *Nature* 2007; **450**(7167):203–18.
58. Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature* 2002;**420**(6915):520–62.
59. Gregory SG, Barlow KF, McLay KE, *et al*. The DNA sequence and biological annotation of human chromosome 1. *Nature* 2006;**441**(7091):315–21.
60. Galagan JE, Calvo SE, Cuomo C, *et al*. Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. *Nature* 2005;**438**(7071):1105–15.
61. Kuhn RM, Karolchik D, Zweig AS, *et al*. The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res* 2009;**37**(Database issue):D755–61.
62. Mituyama T, Yamada K, Hattori E, *et al*. The Functional RNA Database 3.0: databases to support mining and annotation of functional RNAs. *Nucleic Acids Res* 2009; **37**(Database issue):D89–92.
63. Lestrade L, Weber MJ. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res* 2006;**34**(Database issue):D158–62.
64. Piekna-Przybylska D, Decatur WA, Fournier MJ. New bioinformatic tools for analysis of nucleotide modifications in eukaryotic rRNA. *RNA* 2007;**13**(3):305–12.
65. Eddy SR. Infernal user's guide 2009. http://infernal.janelia.org/ 12–14, Technical manual, HHMI Janelia Farm, VA, USA.