RNAcentral: A vision for an international database of RNA sequences

ALEX BATEMAN,[1,22] SHIPRA AGRAWAL,[2,3] EWAN BIRNEY,[4] ELSPETH A. BRUFORD,[4] JANUSZ M. BUJNICKI,[5,6] GUY COCHRANE,[4] JAMES R. COLE,[7] MARCEL E. DINGER,[8] ANTON J. ENRIGHT,[4] PAUL P. GARDNER,[1] DANIEL GAUTHERET,[9] SAM GRIFFITHS-JONES,[10] JEN HARROW,[1] JAVIER HERRERO,[4] IAN H. HOLMES,[11] HSIEN-DA HUANG,[12] KRYSTYNA A. KELLY,[13] PAUL KERSEY,[4] ANA KOZOMARA,[10] TODD M. LOWE,[14] MANJA MARZ,[15] SIMON MOXON,[16] KIM D. PRUITT,[17] TORE SAMUELSSON,[18] PETER F. STADLER,[19] ALBERT J. VILELLA,[4] JAN-HINNERK VOGEL,[1] KELLY P. WILLIAMS,[20] MATHEW W. WRIGHT,[4] and CHRISTIAN ZWIEB[21]

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, United Kingdom
[2]Institute of Bioinformatics and Applied Biotechnology (IBAB), Bangalore 560 100, India
[3]BioCOS Life Sciences Private Limited, Bangalore 560 100, India
[4]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SD, United Kingdom
[5]Laboratory of Bioinformatics and Protein Engineering, International Institute of Molecular and Cell Biology in Warsaw, Trojdena 4, 02-109 Warsaw, Poland
[6]Laboratory of Bioinformatics, Institute of Molecular Biology and Biotechnology, Faculty of Biology, Umultowska 89, 61-614 Poznan, Poland
[7]Microbial Ecology Center, Michigan State University, East Lansing, Michigan 48824-1319, USA
[8]Institute for Molecular Bioscience, The University of Queensland, St Lucia QLD 4072, Australia
[9]Institut de Génétique et Microbiologie–UMR CNRS 8621, Université Paris-Sud–Bâtiment 400, 91405 Orsay Cedex, France
[10]Faculty of Life Sciences, University of Manchester, Michael Smith Building, Manchester, M13 9PT, United Kingdom
[11]Department of Bioengineering, University of California, Berkeley, California 94720-1762, USA
[12]Institute of Bioinformatics and Systems Biology, National Chiao Tung University, HsinChu, 30050, Taiwan
[13]Department of Plant Sciences, University of Cambridge, Cambridge CB2 3EA, United Kingdom
[14]Department of Biomolecular Engineering, University of California, Santa Cruz, California 95064, USA
[15]RNA Bioinformatics Group, Institute of Pharmaceutical Chemistry, Marbacher Weg 6, 35037 Marburg, Germany
[16]University of East Anglia, Norwich, NR4 7TJ, United Kingdom
[17]National Center for Biotechnology Information, National Library of Medicine, Bethesda, Maryland 20894, USA
[18]Department of Medical Biochemistry, University of Goteborg, Medicinareg. 9A, S-405 30 Goteborg, Sweden
[19]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, 04009 Leipzig, Germany
[20]Sandia National Laboratories, MS 9291, Livermore, California 94551-0969, USA
[21]Department of Biochemistry, University of Texas Health Science Center at San Antonio, San Antonio, Texas 78229-3901, USA

ABSTRACT

During the last decade there has been a great increase in the number of noncoding RNA genes identified, including new classes such as microRNAs and piRNAs. There is also a large growth in the amount of experimental characterization of these RNA components. Despite this growth in information, it is still difficult for researchers to access RNA data, because key data resources for noncoding RNAs have not yet been created. The most pressing omission is the lack of a comprehensive RNA sequence database, much like UniProt, which provides a comprehensive set of protein knowledge. In this article we propose the creation of a new open public resource that we term RNAcentral, which will contain a comprehensive collection of RNA sequences and fill an important gap in the provision of biomedical databases. We envision RNA researchers from all over the world joining a federated RNAcentral network, contributing specialized knowledge and databases. RNAcentral would centralize key data that are currently held across a variety of databases, allowing researchers instant access to a single, unified resource. This resource would facilitate the next generation of RNA research and help drive further discoveries, including those that improve food production and human and animal health. We encourage additional RNA database resources and research groups to join this effort. We aim to obtain international network funding to further this endeavor.

Keywords: sequence database; federation; noncoding RNA

INTRODUCTION

In recent years there has been a fundamental shift in our understanding of the role of RNA molecules in cellular biology. The growth of the RNA field has been extraordinary: More than 7700 papers mentioning noncoding-related RNA keywords were published in 2009 alone (Fig. 1). The majority of noncoding DNA sequence has now been shown to be transcribed into so-called noncoding RNA transcripts (often abbreviated ncRNA) using techniques such as large-scale cDNA sequencing (Maeda et al. 2006), tiling arrays (Kapranov et al. 2005), and more recently by harnessing

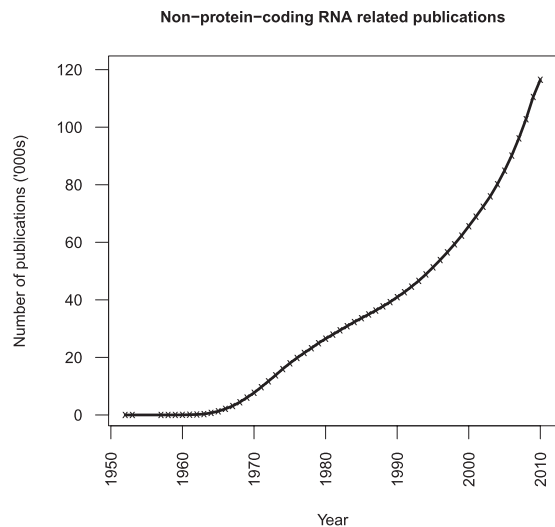RNA (2011), 17:1941–1946. Published by Cold Spring Harbor Laboratory Press.

1941

FIGURE 1. The cumulative number of papers in PubMed that contain noncoding RNA-related keywords in the title or abstract.

next-generation sequencing for RNA-seq (Mortazavi et al. 2008). During the past decade many new classes of functional RNA molecules have been discovered and characterized, such as microRNAs (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001), riboswitches (Lai 2003), plant *trans*-acting small RNAs (Peragine et al. 2004; Vazquez et al. 2004), and piwi-associated RNAs (Lau et al. 2006). The central importance of ncRNA was underlined by the discovery that the ribosome, which synthesizes proteins, is an RNA enzyme (Rodnina et al. 2007). Similarly, RNA molecules in the eukaryotic spliceosome responsible for the removal of introns are likely to be RNA enzymes (Valadkhan et al. 2009). It is probable that many other classes of RNAs still await discovery. Hundreds of thousands of RNAs with unknown function have been identified, including a large number of vertebrate long noncoding RNAs (Guttman et al. 2009), and the biological roles of noncoding transcription are far from understood. Because of the recency of these major discoveries, resources for RNA bioinformatics lag far behind those for proteins, and many researchers remain unaware of or are unable to access the latest RNA research output.

Current state of the field of RNA sequence databases

Presently, there are many specialized databases that collect information for specific RNA classes. However, many classes of RNAs are not represented in databases, and there is no centralized location that stores and organizes RNA sequences and annotations. The Rfam database of RNA families is widely used as a source of RNA sequences, but is restricted to RNA molecules or domains for which an expert multiple alignment is available that represents a limited subset of the potential collection of full-length noncoding transcripts. Any genome-wide application using RNA data

requires researchers to compile data sets from DDBJ/EMBL/GenBank databases, from specialist and general family databases, and from model organism databases. Such resources are not synchronized with respect to genome and annotation versions, and there is a wide range in terms of quality, data formats, and coverage. The task of compiling these data is too onerous for any single lab, and so the majority of researchers remain ignorant of ncRNAs that are relevant and informative to their studies.

There is an important and timely need to organize information about RNAs to facilitate research in medicine, clinical diagnosis, molecular biology, biotechnology, agriculture, ecology, and many related fields. The ability of these communities to access this new wealth of information is greatly inhibited by the lack of a single resource of RNA sequences and their annotation. In July 2010 a meeting was held at the Wellcome Trust Genome Campus with numerous members of the RNA community to discuss how to address these issues. Participants included representatives from the following databases: EMBL (Leinonen et al. 2011), Ensembl genomes (Kersey et al. 2010), gtRNAdb (Chan and Lowe 2009), HGNC (Seal et al. 2011), lncRNAdb (Amaral et al. 2011), miRBase (Kozomara and Griffiths-Jones 2011), Modomics (Czerwoniec et al. 2009), piRNAbank (Sai Lakshmi and Agrawal 2008), Pombase, Refseq (Pruitt et al. 2009), Rfam (Gardner et al. 2011), the Ribosomal Database Project (Cole et al. 2009), RNAdb (Pang et al. 2007), sRNAmap (Huang et al. 2009), SRPDB (Andersen et al. 2006), tmRDB (Andersen et al. 2006), the tmRNA website (Gueneau de Novoa and Williams 2004), and VEGA (Wilming et al. 2008). In this work we propose the creation of a new open public resource, which we term RNAcentral, that will provide a comprehensive collection of RNA sequences and fill an important gap in the provision of biomedical databases.

RELEVANCE OF RNA INFORMATION

A centralized RNA database is urgently needed to facilitate the full annotation of the existing and rapidly emerging new genome sequences. Furthermore, the research community is looking forward to a single authoritative resource that will enable searches for RNA sequence similarities, the prediction of RNA structure, and discovery of new functional classes and interactions. The RNA community will be important users of an RNA sequence database, but there is wider relevance and a much larger potential set of users. The database will be used by investigators spanning diverse life-science research communities, ranging from bioinformaticians, to experimental biologists, to academic clinicians. For example, a typical functional genomics experiment might study a transcription factor by knocking out its gene in mouse, then monitoring gene expression with methods such as RNAseq or tiling arrays that are not biased to protein genes. Typically, half of the hits in such studies are to noncoding regions. Further study of such

loci, especially when repeatedly identified, often leads to the discovery of novel RNAs for which no repository yet exists.

In this section we outline the importance of information about RNAs for a variety of scientific areas.

Medicine

New discoveries of the roles of RNA in human disease are increasing. There have been many discoveries implicating a variety of RNAs in human health. For example, mutations in the microRNA miR-96 have been linked to progressive hearing loss (Lewis et al. 2009) and the disease cartilage hair hypoplasia has been associated with mutations in the RNA component of RNase MRP (Ridanpaa et al. 2001). Similarly, variation in hyperferritinemia cataract syndrome is the result of a mutation in a noncoding RNA element called the iron-responsive element (Perez de Nanclares et al. 2001). There is growing evidence that the deletion of a locus in the human genome containing multiple copies of the snoRNA SNORD116 causes the major Prader-Willi phenotypes (Buiting 2010). Additionally, multiple different types of RNAs have been linked to a wide variety of cancer types. MicroRNAs have been shown to be important regulators of growth and differentiation and are strongly implicated in cancer as oncogenes or tumor suppressors (He et al. 2005; Lu et al. 2005). Y RNAs are massively overexpressed in tumors relative to normal tissue types (Christov et al. 2008), and several groups have linked long ncRNAs to carcinogenesis (Braconi et al. 2011).

The pathogenicity of several infectious agents is dependent on RNA elements. Small RNAs and RNA switches are involved in the virulence and antibiotic resistance of pathogenic bacteria. For example, infection by hepatitis C virus is dependent on the expression of miR-122 (Jopling et al. 2005). This has led to promising new treatments of this viral disease (Elmen et al. 2008). In the bacterium *Listeria monocytogenes* the expression of virulence genes depends on an RNA element called the prfA thermosensor (Johansson et al. 2002). RNAs and RNA processes unique to pathogen groups, such as tmRNA in bacteria and mRNA *trans*-splicing in trypanosomes, may serve as targets for novel drugs.

In total, >80% of the loci associated with disease discovered by genome-wide studies map to noncoding regions of the human genome (Manolio et al. 2009). Given the pervasive transcription of mammalian genomes, it is highly likely that many of the causal variants will turn out to be in ncRNA genes and RNA regulatory elements.

Biotechnology and therapeutics

Small RNAs are increasingly being tested as therapeutic agents. Currently, a number of efforts are underway to determine the viability of both siRNAs and microRNAs for therapeutics. Delivery of these molecules and prediction of secondary toxic effects are a major challenge. There are current clinical trials in cancer, autoimmune disorders, and heart disease to assess the performance of small RNAs as therapeutics. Also, microRNAs are increasingly being used in both cancer and heart disease as diagnostic and prognostic indicators.

The ribosome is one of the major antibiotic targets of bacteria. Many established antibiotics are now known to interact directly with the RNA component of the ribosome, thereby inhibiting protein synthesis and, therefore, bacterial growth (Tenson and Mankin 2006).

Another direction that is being explored for developing novel RNA-based therapeutic agents is the use of ribozymes or RNAs that can catalyze reactions. In particular, RNA-cleaving ribozymes designed to target and cleave specific sites on specific target RNAs are undergoing clinical trial (Citti and Rainaldi 2005).

Agriculture

Food security is one of this century's key global challenges. The challenge to produce sufficient food for the increasing global population must be met in the face of changing consumption patterns, demand for biofuels, the impact of climate change, and the growing scarcity of water and land. In plants, noncoding RNAs are involved in many physiological processes determining growth and development and, ultimately, crop yield (including leaf morphogenesis, floral differentiation and development, root initiation and development, vascular development, the transition from vegetative growth to reproductive growth, and fruit ripening). Plant small RNAs are involved in responses to environmental factors such as water (Trindade et al. 2010), salt (Borsani et al. 2005), and metals (Sunkar et al. 2006). There is emerging evidence that plant small RNAs play a role in hybrid necrosis, a phenomenon that is a barrier to conventional plant breeding. Small RNAs are also important in defense against viral pathogens that have a devastating impact on important food crops such as rice and cereals (Mlotshwa et al. 2008). A class of self-replicating RNA-based pathogen, known as viroids, are a major economic threat to horticultural crops (Tsagris et al. 2008). Perhaps most notable of these is the *Potato spindle tuber viroid*, which can cause stunting and distortion of leaves and fruit, necrosis, and even death of the host plant.

Ecological relevance

Ribosomal RNA sequences are widely used in molecular phylogeny and evolutionary biology, microbial ecology, bacterial identification, characterizing microbial populations, and in understanding the diversity of life. In *Bacteria* and *Archaea*, CRISPR RNAs function as an immune system against phage infection, greatly influencing bacterial and phage population dynamics (Grissa et al. 2007). Small RNAs and RNA switches allow microbial gene expression to

respond to changes in levels of specific small molecules and in environmental factors such as temperature. Microbial processes in turn play an important role in carbon cycling, carbon sequestration, and reducing organic matter to carbon dioxide. A better understanding of factors controlling microbial population dynamics, and the response of these factors to elevated temperature and carbon dioxide are important for modeling climate change.

A VISION FOR RNACENTRAL

RNAcentral will provide a central entry point for those seeking to exploit noncoding RNA data. A standardized set of reference records, representing noncoding RNA mature transcripts and precursor molecules, will form the core content. Key information, such as sequence, biological source, function, and supporting evidence will be attached to each record. Supplementary information, such as mappings to source genomes, tissue and developmental stage patterns of expression, secondary structure, and links to the literature will also be made available. This rich information resource will be made possible through the provision of data from specialist databases, called RNAcentral expert databases, to the central hub (see Fig. 2). In this model, we

take advantage of the wealth of expertise that already exists in many independent data resources, but whose full potential has not yet been realized as a result of their isolation. We strongly encourage other expert databases to join the RNAcentral network.

From a user perspective, RNAcentral will provide a web portal that allows querying by name or accession number as well as searches by sequence similarity. A researcher who identified a potential novel RNA transcript could use RNAcentral to search for homologs in the complete collection of known RNAs. For example, they may identify that their transcript is similar to a known microRNA. From the RNAcentral web portal, the user could see information about the homologous microRNA, including its sequence and genomic location. The user would be directed to miRBase, an RNAcentral expert database, for further information such as target sites and expression patterns.

A model for a federated database

Federated and collaborative models have frequently proved successful in the establishment of biomedical informatics resources. For example, the InterPro database of protein domains (Hunter et al. 2009) includes a number of independent resources, each scientifically active and with their own funding streams, but collaborating to contribute to the integrated database through their use of common data types. Federated models are of particular relevance for noncoding RNA, since the discovery of new classes of RNA genes has resulted in complementary, but distinct activities across the globe centered upon individual RNA classes. These activities comprise collation of dispersed data, annotation of ncRNA genes in complete genomes, functional annotation, and data presentation. In addition, new high-throughput technologies are enabling new approaches to ncRNA research and are generating increasingly large quantities of complex sequence data.
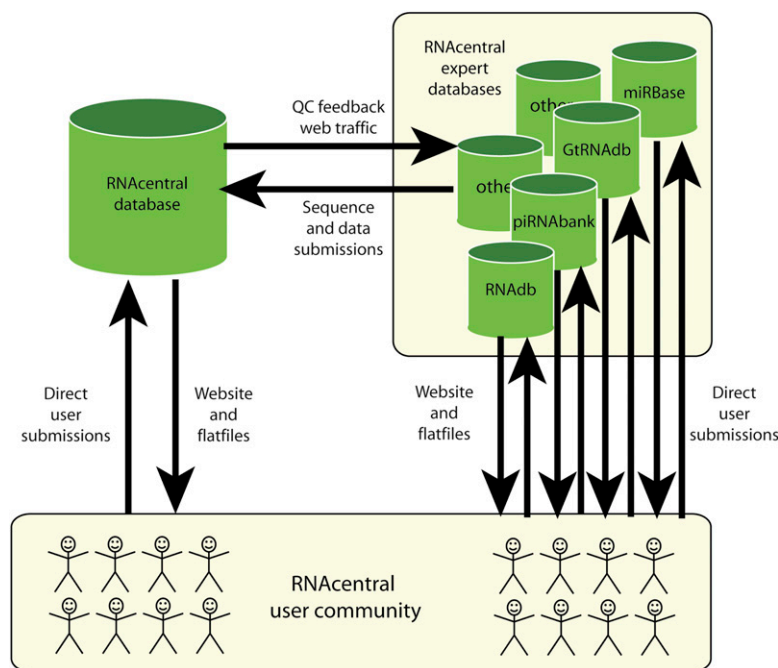
Functions of the RNAcentral expert databases

The RNAcentral expert databases are run by RNA biologists with many years of expertise on specific types of noncoding RNAs. These databases already have excellent links with their user communities from whom they often accept submissions and corrections. They are able to collate new data and refine it

FIGURE 2. Organization of RNAcentral and the RNAcentral expert databases. Many RNAcentral expert databases exist already (such as Rfam, RNAdb, piRNAbank, etc.), but the scheme can flexibly include new databases that are created over time. The RNAcentral expert databases provide their content via standard exchange formats to the RNAcentral database, which holds information about all ncRNAs and links special features (such as alignments or predicted 3D-structure) back to RNAcentral expert databases. All information is freely available for the RNA community both through the RNAcentral database as well as the expert databases. In the case of novel ncRNA sequences that do not fit into a class covered by the RNAcentral expert databases, the RNA community will be able to communicate with and submit their data directly to the RNAcentral database.

into high-quality curated information for automated submission to RNAcentral. The RNAcentral expert databases will commit to providing regular updates, which must be provided in complete form and made freely available to the public without restriction. The expert databases will play a key role in developing and maintaining domain-specific structured vocabularies that will allow the richness of the expert database to make its way into RNAcentral, while being consistent with the other expert databases. Rfam, as a special RNAcentral expert database, provides ncRNA classes not covered by existing RNAcentral expert databases. The smooth running of this federated endeavor will require the RNAcentral expert databases to provide their content via standard exchange formats. The RNAcentral expert databases will continue to provide their own domain-specific information, which is beyond the scope of the RNAcentral database. The expert databases are also well placed to provide systematic names for ncRNAs as well as nomenclature rules.

Function of the RNAcentral database

The RNAcentral database will provide a central repository of RNA sequences with stable identifiers. As well as taking submissions from the RNAcentral expert databases, it will also receive submissions from the RNA community directly. This is particularly important when the RNA in question is not covered by one of the RNAcentral expert databases. Inconsistencies are reported back to the expert databases by individual database reports. An important function of the RNAcentral database will be to provide a central portal for RNA sequence that will allow sequence searches as well as browsing of the data. As the RNAcentral database project grows, it is envisaged that more curated biological function data will be incorporated. However, users will be provided with links out to the expert databases for more specialized data wherever appropriate. The RNAcentral database will also identify inconsistencies, such as in naming, between the expert databases, and errors identified during quality control will be communicated back. Finally, the RNAcentral resource will provide code and tools that enable to RNAcentral expert databases to integrate their data smoothly into the central repository, while achieving format compliance.

We believe that this federated model will help to support the diversity of expertise within the RNA community, while also allowing scientists access to a unified resource. We will work with the journals to make submission of RNA sequences into RNAcentral mandatory, as it is for nucleic acid sequence deposition in ENA/GenBank/DDBJ and molecular structures in the wwPDB.

FUNDING

Initial funding for RNAcentral would allow the creation of the core activities of sequence collection and assignment of accessions as well as creating a simple web interface to the data. This initial resource would integrate the existing ncRNA databases and mirror their core content. Once this initial core was operational and a proof of concept established, then further funding would be sought to expand the remit of the database to include curation and the addition of functional information. The funding for the RNAcentral expert databases would continue to be a heterogeneous mix of institutional, national, and international funding. We also see opportunities for large-scale international network funding for this important community-led effort to integrate our knowledge of noncoding RNA biology.

REFERENCES

Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS. 2011. lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Res* **39:** D146–D151.

Andersen ES, Rosenblad MA, Larsen N, Westergaard JC, Burks J, Wower IK, Wower J, Gorodkin J, Samuelsson T, Zwieb C. 2006. The tmRDB and SRPDB resources. *Nucleic Acids Res* **34:** D163–D168.

Borsani O, Zhu J, Verslues PE, Sunkar R, Zhu JK. 2005. Endogenous siRNAs derived from a pair of natural *cis*-antisense transcripts regulate salt tolerance in *Arabidopsis*. *Cell* **123:** 1279–1291.

Braconi C, Valeri N, Kogure T, Gasparini P, Huang N, Nuovo GJ, Terracciano L, Croce CM, Patel T. 2011. Expression and functional role of a transcribed noncoding RNA with an ultraconserved element in hepatocellular carcinoma. *Proc Natl Acad Sci* **108:** 786–791.

Buiting K. 2010. Prader-Willi syndrome and Angelman syndrome. *Am J Med Genet C Semin Med Genet* **154C:** 366–376.

Chan PP, Lowe TM. 2009. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. *Nucleic Acids Res* **37:** D93–D97.

Christov CP, Trivier E, Krude T. 2008. Noncoding human Y RNAs are overexpressed in tumours and required for cell proliferation. *Br J Cancer* **98:** 981–988.

Citti L, Rainaldi G. 2005. Synthetic hammerhead ribozymes as therapeutic tools to control disease genes. *Curr Gene Ther* **5:** 11–24.

Cole JR, Wang Q, Cardenas E, Fish J, Chai B, Farris RJ, Kulam-Syed-Mohideen AS, McGarrell DM, Marsh T, Garrity GM, et al. 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res* **37:** D141–D145.

Czerwoniec A, Dunin-Horkawicz S, Purta E, Kaminska KH, Kasprzak JM, Bujnicki JM, Grosjean H, Rother K. 2009. MODOMICS: a database of RNA modification pathways. 2008 update. *Nucleic Acids Res* **37:** D118–D121.

Elmen J, Lindow M, Schutz S, Lawrence M, Petri A, Obad S, Lindholm M, Hedtjarn M, Hansen HF, Berger U, et al. 2008.

LNA-mediated microRNA silencing in non-human primates. *Nature* 452: 896–899.

Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe DL, Eddy SR, et al. 2011. Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res* 39: D141–D145.

Grissa I, Vergnaud G, Pourcel C. 2007. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics* 8: 172. doi: 10.1186/1471-2105-8-172.

Gueneau de Novoa P, Williams KP. 2004. The tmRNA website: reductive evolution of tmRNA in plastids and other endosymbionts. *Nucleic Acids Res* 32: D104–D108.

Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, et al. 2009. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458: 223–227.

He L, Thomson JM, Hemann MT, Hernando-Monge E, Mu D, Goodson S, Powers S, Cordon-Cardo C, Lowe SW, Hannon GJ, et al. 2005. A microRNA polycistron as a potential human oncogene. *Nature* 435: 828–833.

Huang HY, Chang HY, Chou CH, Tseng CP, Ho SY, Yang CD, Ju YW, Huang HD. 2009. sRNAMap: genomic maps for small non-coding RNAs, their regulators and their targets in microbial genomes. *Nucleic Acids Res* 37: D150–D154.

Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, et al. 2009. InterPro: the integrative protein signature database. *Nucleic Acids Res* 37: D211–D215.

Johansson J, Mandin P, Renzoni A, Chiaruttini C, Springer M, Cossart P. 2002. An RNA thermosensor controls expression of virulence genes in Listeria monocytogenes. *Cell* 110: 551–561.

Jopling CL, Yi M, Lancaster AM, Lemon SM, Sarnow P. 2005. Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. *Science* 309: 1577–1581.

Kapranov P, Drenkow J, Cheng J, Long J, Helt G, Dike S, Gingeras TR. 2005. Examples of the complex architecture of the human transcriptome revealed by RACE and high-density tiling arrays. *Genome Res* 15: 987–997.

Kersey PJ, Lawson D, Birney E, Derwent PS, Haimel M, Herrero J, Keenan S, Kerhornou A, Koscielny G, Kahari A, et al. 2010. Ensembl Genomes: extending Ensembl across the taxonomic space. *Nucleic Acids Res* 38: D563–D569.

Kozomara A, Griffiths-Jones S. 2011. miRBase: integrating microRNA annotation and deep-sequencing data. *Nucleic Acids Res* 39: D152–D157.

Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* 294: 853–858.

Lai EC. 2003. RNA sensors and riboswitches: self-regulating messages. *Curr Biol* 13: R285–R291.

Lau NC, Lim LP, Weinstein EG, Bartel DP. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* 294: 858–862.

Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, Kingston RE. 2006. Characterization of the piRNA complex from rat testes. *Science* 313: 363–367.

Lee RC, Ambros V. 2001. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* 294: 862–864.

Leinonen R, Akhtar R, Birney E, Bower L, Cerdeno-Tarraga A, Cheng Y, Cleland I, Faruque N, Goodgame N, Gibson R, et al. 2011. The European Nucleotide Archive. *Nucleic Acids Res* 39: D28–D31.

Lewis MA, Quint E, Glazier AM, Fuchs H, De Angelis MH, Langford C, van Dongen S, Abreu-Goodger C, Piipari M, Redshaw N, et al. 2009. An ENU-induced mutation of miR-96 associated with progressive hearing loss in mice. *Nat Genet* 41: 614–618.

Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, et al. 2005. MicroRNA expression profiles classify human cancers. *Nature* 435: 834–838.

Maeda N, Kasukawa T, Oyama R, Gough J, Frith M, Engstrom PG, Lenhard B, Aturaliya RN, Batalov S, Beisel KW, et al. 2006. Transcript annotation in FANTOM3: mouse gene catalog based on physical cDNAs. *PLoS Genet* 2: e62. doi: 10.1371/journal.pgen.0020062.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, et al. 2009. Finding the missing heritability of complex diseases. *Nature* 461: 747–753.

Mlotshwa S, Pruss GJ, Vance V. 2008. Small RNAs in viral infection and host defense. *Trends Plant Sci* 13: 375–382.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 5: 621–628.

Pang KC, Stephen S, Dinger ME, Engstrom PG, Lenhard B, Mattick JS. 2007. RNAdb 2.0–an expanded database of mammalian non-coding RNAs. *Nucleic Acids Res* 35: D178–D182.

Peragine A, Yoshikawa M, Wu G, Albrecht HL, Poethig RS. 2004. SGS3 and SGS2/SDE1/RDR6 are required for juvenile development and the production of *trans*-acting siRNAs in *Arabidopsis*. *Genes Dev* 18: 2368–2379.

Perez de Nanclares G, Castano L, Martul P, Rica I, Vela A, Sanjurjo P, Aldamiz-Echevarria K, Martinez R, Sarrionandia MJ. 2001. Molecular analysis of hereditary hyperferritinemia-cataract syndrome in a large Basque family. *J Pediatr Endocrinol Metab* 14: 295–300.

Pruitt KD, Tatusova T, Klimke W, Maglott DR. 2009. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res* 37: D32–D36.

Ridanpää M, van Eenennaam H, Pelin K, Chadwick R, Johnson C, Yuan B, vanVenrooij W, Pruijn G, Salmela R, Rockas S, et al. 2001. Mutations in the RNA component of RNase MRP cause a pleiotropic human disease, cartilage-hair hypoplasia. *Cell* 104: 195–203.

Rodnina MV, Beringer M, Wintermeyer W. 2007. How ribosomes make peptide bonds. *Trends Biochem Sci* 32: 20–26.

Sai Lakshmi S, Agrawal S. 2008. piRNABank: a web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Res* 36: D173–D177.

Seal RL, Gordon SM, Lush MJ, Wright MW, Bruford EA. 2011. genenames.org: the HGNC resources in 2011. *Nucleic Acids Res* 39: D514–D519.

Sunkar R, Kapoor A, Zhu JK. 2006. Posttranscriptional induction of two Cu/Zn superoxide dismutase genes in *Arabidopsis* is mediated by downregulation of miR398 and important for oxidative stress tolerance. *Plant Cell* 18: 2051–2065.

Tenson T, Mankin A. 2006. Antibiotics and the ribosome. *Mol Microbiol* 59: 1664–1677.

Trindade I, Capitao C, Dalmay T, Fevereiro MP, Santos DM. 2010. miR398 and miR408 are up-regulated in response to water deficit in *Medicago truncatula*. *Planta* 231: 705–716.

Tsagris EM, Martinez de Alba AE, Gozmanova M, Kalantidis K. 2008. Viroids. *Cell Microbiol* 10: 2168–2179.

Valadkhan S, Mohammadi A, Jaladat Y, Geisler S. 2009. Protein-free small nuclear RNAs catalyze a two-step splicing reaction. *Proc Natl Acad Sci* 106: 11901–11906.

Vazquez F, Vaucheret H, Rajagopalan R, Lepers C, Gasciolli V, Mallory AC, Hilbert JL, Bartel DP, Crete P. 2004. Endogenous *trans*-acting siRNAs regulate the accumulation of *Arabidopsis* mRNAs. *Mol Cell* 16: 69–79.

Wilming LG, Gilbert JG, Howe K, Trevanion S, Hubbard T, Harrow JL. 2008. The vertebrate genome annotation (Vega) database. *Nucleic Acids Res* 36: D753–D760.