Naming 'junk': Human non-protein coding RNA (ncRNA) gene nomenclature

*Mathew W. Wright** *and Elspeth A. Bruford*

HUGO Gene Nomenclature Committee (HGNC), EMBL-EBI, Wellcome Trust Genome Campus, Cambridge, CB10 1SD, UK
**Correspondence to:* Tel: +44 (0)1223 494 444; Fax: +44 (0)1223 494 468; E-mail: hgnc@genenames.org

Date received (in revised form): 4th October 2010

Abstract

Previously, the majority of the human genome was thought to be 'junk' DNA with no functional purpose. Over the past decade, the field of RNA research has rapidly expanded, with a concomitant increase in the number of non-protein coding RNA (ncRNA) genes identified in this 'junk'. Many of the encoded ncRNAs have already been shown to be essential for a variety of vital functions, and this wealth of annotated human ncRNAs requires standardised naming in order to aid effective communication. The HUGO Gene Nomenclature Committee (HGNC) is the only organisation authorised to assign standardised nomenclature to human genes. Of the 30,000 approved gene symbols currently listed in the HGNC database (http://www.genenames.org/search), the majority represent protein-coding genes; however, they also include pseudogenes, phenotypic loci and some genomic features. In recent years the list has also increased to include almost 3,000 named human ncRNA genes. HGNC is actively engaging with the RNA research community in order to provide unique symbols and names for each sequence that encodes an ncRNA. Most of the classical small ncRNA genes have now been provided with a unique nomenclature, and work on naming the long (>200 nucleotides) non-coding RNAs (lncRNAs) is ongoing.

Keywords: *ncRNA, RNA, nomenclature, non-protein coding*

Introduction

At the beginning of this century, many geneticists were predicting that the human genome contained around 100,000 protein-coding genes, partly based on the assumption that more complex organisms would have a greater number of genes. Ten years later, with far more genomic data from a wide variety of organisms and a much better-quality, well-annotated human genome, this original expectation has been downsized to around 20,000 protein-coding genes. This means that highly complex organisms like the human have about the same number of protein-coding genes as much simpler life forms such as the roundworm, *Caenorhabditis elegans*. If we look to the human's closest living relative, the chimpanzee, we see that the equivalent proteins in human and chimpanzee typically differ by only two amino acids, and approximately 29 per cent of all the orthologous proteins encoded in human and chimpanzee are identical.[1] Why, then, when the protein-coding components of our genomes are so similar, are humans and chimpanzees so strikingly different? Since protein-coding genes comprise only two per cent of the human genome, the answer may lie in the large swathes of the genome previously regarded as 'junk DNA'. Indeed, the ENCyclopedia Of DNA Elements (ENCODE) Consortium,[2] which is aiming to identify all the functional elements in the human genome, suggests that the vast majority of the genome is transcribed as non-protein-coding RNA (ncRNA). These RNAs could be responsible for some of the complex differences between humans and other primates, especially since the expression of many genes is now thought to be regulated by ncRNAs. They are also known to be involved in

Table 1. A summary of the current nomenclature for human non-protein coding RNA genes

Type of RNA	HGNC stem symbol	Name format
miRNA	*MIR#*	microRNA #
tRNA — genomic	*TRNA#*	transfer RNA 'single letter amino acid code' # (anticodon XXX)
— mitochondrial	*MT-T#*	mitochondrially encoded tRNA 'single letter amino acid code' #
rRNA — 5S	*RN5S#*	RNA, 5S ribosomal #
— 5.8S	*RN5-8S#*	RNA, 5.8S ribosomal #
— 18S	*RN18S#*	RNA, 18S ribosomal #
— 28S	*RN28S#*	RNA, 28S ribosomal #
Spliceosomal — U1	*RNU1-#*	RNA, U1 small nuclear #
— U2	*RNU2-#*	RNA, U2 small nuclear #
— U4	*RNU4-#*	RNA, U4 small nuclear #
— U5	*RNU5-#*	RNA, U5 small nuclear #
— U6	*RNU6-#*	RNA, U6 small nuclear #
— U4atac	*RNU4ATAC#*	RNA, U4atac small nuclear #
— U6atac	*RNU6ATAC#*	RNA, U6atac small nuclear #
— U11	*RNU11*	RNA, U11 small nuclear
— U12	*RNU12-#*	RNA, U12 small nuclear #
snoRNA — H/ACA box	*SNORA#*	small nucleolar RNA, H/ACA box #
— C/D box	*SNORD#*	small nucleolar RNA, C/D box #
— Cajal body specific	*SCARNA#*	small cajal body-specific RNA #
piRNA — cluster	*PIRC#*	piwi-interacting RNA cluster #
— individual	*PIRNA#*	piwi-interacting RNA #
RNase — MRP	*RMRP*	RNA component of mitochondrial RNA processing endoribonuclease
— P	*RPPH1*	ribonuclease P RNA component H1
U7	*RNU7-#*	RNA, U7 small nuclear #

Continued

Table 1. Continued

Type of RNA	HGNC stem symbol	Name format
Vault	*VTRNA#*	vault RNA #
7SK	*RN7SK*	RNA, 7SK small nuclear
Y	*RNY#*	RNA, Ro-associated Y #
SRP/7SL	*RN7SL#*	RNA, 7SL, cytoplasmic #
Telomerase	*TERC*	telomerase RNA component
lncRNA — known function	Function-based name (eg *XIST*)	eg X (inactive)-specific transcript (non-protein coding)
— antisense	'-AS' suffix* (eg *BOK-AS1*)	eg BOK antisense RNA #1 (non-protein coding)
— intronic	'- IT' suffix* (eg *MAGI2-IT1*)	eg MAGI2-IT1 intronic transcript #1 (non-protein coding)
— host gene of small ncRNA	'HG' suffix (eg *SNHG1*)	eg small nucleolar RNA host gene 1 (non-protein coding)
— intergenic	*LINC#*	long intergenic non-protein coding RNA #

*to a protein-coding gene symbol

many other diverse and vital roles within our bodies, including protein biosynthesis and the splicing of messenger RNAs (mRNAs), and have been implicated in many diseases, such as Prader–Willi syndrome and various cancers (for review, see Taft *et al.*[3]). It is not surprising, then, that interest in ncRNAs has accelerated every year and that several thousand papers were published on them in 2010. As more ncRNAs are identified and characterised, the wealth of information concerning these genes and their encoded RNAs in the public domain increases greatly. When the human genome was initially sequenced and annotated, there was a vast increase in the data available on protein-coding genes. Fortunately, the HUGO Gene Nomenclature Committee (HGNC)[4] was on hand to provide unique names for these genes, thus ensuring that everyone was able to retrieve and discuss relevant information concerning specific loci. The proper characterisation and naming of human ncRNA genes and transcripts is equally vital, and so the HGNC has been actively engaging with the RNA research community in order to provide unique names for all the sequences encoding ncRNAs; the progress so far is shown in Table 1 and discussed below.

MicroRNAs

The first microRNA (miRNA), lin-4, was identified in *C. elegans* in 1993,[5] but the term 'microRNA' was not introduced until 2001.[6] MicroRNA genes encode primary transcripts (pri-miRNAs), which are processed into short stem-loop structures called pre-miRNAs, and these in turn are modified into mature miRNAs. Varying in length from 19 to 25 nucleotides (nt), these single-stranded mature miRNAs bind to the 3′-untranslated regions (3′-UTRs) of target mRNAs to destabilise or inhibit their translation.[7] This regulation of gene expression by miRNAs has been shown to affect diverse cellular functions and has been implicated in the molecular mechanisms of many diseases; for instance, mutations in *MIR96* have been linked to progressive hearing loss.[8] Since 2002, miRBase[9] (http://www.mirbase.org — formerly the miRNA Registry) has catalogued all

identified human miRNAs and provided each sequence with a stable accession number, precise genomic location and a name in the format 'mir-#' for each stem-loop sequence and 'miR-#' for each resultant mature miRNA. For example, the *MIR100* gene encodes the mir-100 stem-loop, which is modified to create the miR-100 mature transcript. Working in collaboration with miRBase, HGNC has provided approved gene symbols for all of the ~1,000 currently identified human miRNAs. These are named using the miRBase names with the stem symbol MIR#. The MIR# symbols are assigned as sequential numerical identifiers to each novel miRNA, with those miRNAs that encode homologous mature transcripts sharing the same MIR number but with differing suffixes. If the mature miRNAs differ by only or two nucleotides, they are allocated letter suffixes (eg *MIR10A* and *MIR10B*), whereas if the mature miRNAs are identical, the genes are given hyphenated numerical suffixes (eg *MIR1-1* and *MIR1-2*).

Transfer RNAs

Transfer RNAs (tRNAs) are small RNA molecules, about 80 nucleotides in length, that play an essential role in protein translation by transporting specific amino acids to the ribosomes to be added to an elongating peptide chain. There are three essential regions on each tRNA: the anticodon, which comprises three nucleotides that can base-pair to one or more specific triplet codons on the mRNA being translated; the attachment site that covalently binds the particular amino acid specified by the codon; and a second attachment site that recognises a specific aminoacyl tRNA synthetase, an enzyme that catalyses the binding of an amino acid to its compatible tRNA. Due to their high degree of structural conservation, such as the distinctive cloverleaf secondary folding and L-shaped tertiary structures, tRNAs can be accurately predicted within a genome using a sequence comparison algorithm. The Genomic tRNA database,[10] GtRNAdb (http://lowelab.ucsc.edu/GtRNAdb/), contains 506 tRNA genes predicted in the human genome using the tRNAscan-SE[11] program. Working with this dataset, the HGNC has provided approved symbols for these genomic tRNA genes in the format 'TRNA + single letter code for amino acid isotype + incremental number', with the specific anticodon type also specified in the name. For example, the tRNA gene '*TRNAA1*' has the name 'transfer RNA alanine 1 (anticodon UGC)'. GtRNAdb also identifies 110 putative tRNA pseudogenes that are named in the same format as coding tRNA genes, except that the symbol is appended with a 'P' for 'pseudogene' (eg *TRNAA44P*, 'transfer RNA alanine 44 [anticodon AGC] pseudogene'). The 'P' suffix is also used for pseudogenes in other ncRNA classes and for pseudogenes of protein-coding genes. Any ncRNAs that have been derived by computational prediction — as is the case for most of the genomic tRNAs — will require experimental validation to determine whether they do encode functional transcripts. There are 22 tRNAs encoded in the human mitochondrial genome; these are thought to have a classical cloverleaf-like secondary structure but differ in the length of their loops and are missing some conserved residues.[12] There is much interest in mitochondrial tRNAs, since many pathological point mutations have been identified in these genes. Mamit-tRNAdb (http://mamit-trna.u-strasbg.fr/),[13] a database of mammalian mitochondrial tRNAs, contains the cloverleaf structures of the 22 human mitochondrial tRNAs, highlighting the positions of each of the mutations reported in the literature. The HGNC names for the mitochondrial tRNA genes are in the format 'MT-T + single letter code for amino acid isotype + incremental number', again with the specific anticodon type included in the name. For example, the tRNA gene '*MT-TS1*' has the name 'mitochondrially encoded tRNA serine 1' (UCN).

Ribosomal RNAs

Ribosomes, the sites of protein translation, comprise both ribosomal proteins and ribosomal RNAs (rRNAs). Eukaryotic ribosomes are 80S in size but are formed by two subunits, one large 60S subunit and one small 40S subunit. The passage of the tRNAs along the mRNA during translation is

facilitated by the rRNAs within the ribosome. The ribosome brings about the interaction between the anticodon of each aminoacyl tRNA and the equivalent codon of the mRNA at its aminoacyl (A) site, and then aids formation of the peptide bond between the amino acids at the peptidyl (P) site before the tRNA exits the ribosome at the exit (E) site. Each single mRNA can be translated at multiple ribosomes at the same time. For a recent review on the structural dynamics of the ribosome, see Korostelev *et al.*[14]

In eukaryotes, there are four types of rRNA: 18S rRNA is found in the small subunit of the ribosome and 28S, 5.8S and 5S rRNAs in the large subunit. The 18S, 5.8S and 28S rRNA genes are arranged in tandem repeats, with the genes separated by transcribed spacers known as externally and internally transcribed sequences (abbreviated to ETS and ITS). Each repeat found in the arrangement 5′ETS-18S-ITS1-5.8S-ITS2-28S-3′ETS produces one precursor transcript, which is then cleaved to produce the three types of rRNAs.[15] 5S rRNA genes are also found between spacer sequences in tandem repeats scattered throughout the genome, but with several main clusters located on the q arm of chromosome 1. Large amounts of RNA are required to make ribosomes, so there are hundreds of copies of both types of rRNA repeats in the human genome. Repetitive sequences, such as these rRNA tandem repeat genes, prove to be problematic when sequencing and assembling genomes and so some rRNA genes are likely to be missing from the genome builds. A simple gene nomenclature for rRNAs has been proposed in which the 18S, 28S, 5.8S and 5S types use the stem symbols *RN18S#*, *RN28S#*, *RN5-8S#* and *RN5S#*, respectively (eg '*RN5S1*' for 'RNA, 5S ribosomal 1'). If agreed, this scheme will be implemented by the rRNA research community. A comprehensive set of all rRNAs can be found in the SILVA ribosomal rRNA database (http://www.arb-silva.de/).[16]

Spliceosomal RNAs

After transcription, most primary transcripts (pre-mRNAs) undergo a series of modifications — termed splicing — in which introns are removed so that the exons can be joined to form a mature mRNA (for a review, see Ritchie *et al.*[17]). Some introns are self-splicing but most require the intervention of the spliceosome, a large ribonucleoprotein (RNP) made up of over 200 different proteins and five small nuclear RNAs (snRNAs), known as U1, U2, U4, U5 and U6. These snRNAs are highly conserved across genomes, since they play a key role in the multiple splicing events catalysed by the spliceosome. The U1, U2, U4, U5 and U6 snRNAs are assembled around the newly transcribed pre-mRNA following a precise pattern that produces a common canonical structure, known as the major or U2-dependent spliceosome. The pre-mRNA contains specific sequences that guide the formation of the major spliceosome, and also specific GT/AG splice sites that flank the introns (known as major or U2-type introns) to identify where the excisions should be made. The nomenclature of these snRNA genes follows the format 'RN + snRNA species + numerical identifier' (eg '*RNU1-1*' for 'RNA, U1 small nuclear 1'). A much less common type of intron, the U12-type, is spliced by the minor spliceosome, which, in addition to U5, contains four different snRNAs, known as U11, U12, U4atac and U6atac.[18] These four snRNAs each has a functional counterpart in the major spliceosome, with U11 being analogous to U1, U12 to U2, U4atac to U4 and U6atac to U6. The 'atac' suffix on U4atac and U6atac denotes the unusual AT/AC splice sites for U12-type introns. The gene names for the U12-type snRNAs follow the same format as the U2-type snRNAs; for example, '*RNU6ATAC5*' is 'RNA, U6atac small nuclear 5'.

Small nucleolar RNAs

Small nucleolar RNAs (snoRNAs) are responsible for guiding a series of site-specific post-transcriptional modifications to rRNAs, tRNAs and snRNAs. There are two main types of snoRNAs: H/ACA box snoRNAs direct the conversion of the nucleoside uridine to pseudouridine by the pseudouridine synthase protein dyskerin,

and C/D box snoRNAs guide the addition of a methyl group by the methyltransferase protein fibrillarin.[19] Each snoRNA guides one, or sometimes two, such modification(s) by binding to complementary regions of the pre-RNA. H/ACA box snoRNAs are named after their intrinsic H box (ANANNA) and ACA box sequences and are found in RNPs containing the same four proteins: DKC1 (dyskerin), GAR1, NOP10 and NHP2. Similarly, the C/D box snoRNAs are named after the conserved C (RUGAUGA) and D (CUGA) boxes and form RNPs with four different proteins: FBL (fibrillarin), NOP56, NOP58 and NHP2L1. The H/ACA box snoRNAs were previously referred to in the literature and databases by the stem symbols ACA# or HBI-# (eg ACA1 and HBI-6); and the C/D box snoRNAs were named using either 14q(I-#), 14q(II-#) or HBII-# symbols (eg 14q(0), 14q(II-1) and HBII-99). These names were confusing and so, in collaboration with snoRNABase (http://www-snorna.biotoul.fr)[20] and experts in the field, HGNC devised an approved nomenclature for the snoRNA genes using a common stem symbol of *SNORA#* for the H/ACA box genes and *SNORD#* for the C/D boxes. HBI-6 is now approved as '*SNORA26*' for 'small nucleolar RNA, H/ACA box 26' and 14q(0) is now '*SNORD112*' for 'small nucleolar RNA, C/D box 112'. Another class of snoRNAs is the small Cajal body-specific RNAs (scaRNAs), named after the sub-organelles within the nucleus where they are located.[21] These RNAs often contain both H/ACA box and C/D box domains, but sometimes have only one of these domains. Previously, scaRNA genes were grouped in with the H/ACA box and C/D box snoRNA genes but they now have their own approved *SCARNA#* gene nomenclature; for example, HBII-382 is now *SCARNA3* for 'small Cajal body-specific RNA 3'.

Piwi-interacting RNAs

Piwi-interacting RNAs (piRNAs) are the largest class of small ncRNAs expressed in vertebrates. Generally ranging from 25 to 33 nucleotides in length, piRNAs play a key role during spermatogenesis in defending germline cells against transposons by selectively silencing them.[22] They are found in positionally conserved clusters throughout mammalian genomes, although the piRNAs within these clusters are not conserved. A cluster can encode from tens up to thousands of individual piRNAs. piRNABank (http://pirnabank.ibab.ac.in/)[23] — a web resource that classifies piRNAs and groups them into their genomic clusters — has so far identified 114 clusters in the human genome and HGNC has provided each of these with a *PIRC#* symbol for 'piwi-interacting RNA cluster #'. A vast number of individual human piRNA sequences has been identified, but piRNABank has removed repetitive and overlapping sequences in order to produce a non-redundant set of around 23,000 piRNAs that map to the human genome. HGNC and piRNABank are currently working together to develop a nomenclature for each piRNA gene, possibly with the stem symbol *PIRNA#*.

RNase P/MRP genes

The RNA component of ribonuclease (RNase) P plays a role in the processing of tRNA precursors (pre-tRNAs) into their mature products by cleaving sequence from the 5′ end. It is also thought to be involved in RNA polymerase (Pol) III transcription.[24] This RNA is encoded by the RNase P RNA component H1 (*RPPH1*) gene in the human genome. The evolutionarily related RNase mitochondrial RNA-processing (MRP) enzyme is involved in the maturation of precursor rRNAs (pre-rRNAs), by splicing out internally transcribed sequences, and also in mitochondrial DNA replication.[25] Although the RNA component is now known to be mostly localised in the nucleus, it was first identified in the mitochondria and this is reflected in the nomenclature of the gene, '*RMRP*' for 'RNA component of mitochondrial RNA processing endoribonuclease'.

Other small ncRNAs

U7 snRNA has a role in histone pre-mRNA processing by specifically binding to the histone

downstream element (HDE).[26] A computational study[27] has revealed only one functional copy (encoded by the gene *RNU7-1*) and 85 non-functional pseudogenes to be present in the human genome. A further computational analysis by the same group identified the four human vault RNA genes.[28] These genes encode the RNA components of the vault RNP. The function of this RNP is still unknown, but a role in drug resistance has been suggested.[29] To avoid confusion with viral RNAs (vRNAs), the genes encoding these ncRNAs have been named using the stem symbol *VTRNA#* (eg '*VTRNA2*' for 'vault RNA 2'). Other small classes of ncRNAs in the human include: 7SK RNA, which is involved in the regulation of Pol II transcription[30] and is encoded by the 'RNA, 7SK small nuclear' (*RN7SK*) gene; the Y RNAs that form part of the Ro RNP are encoded by the *RNY#* genes (eg '*RNY1*' for 'RNA, Ro-associated Y1'); the genes for the RNAs that form part of the signal recognition particle (commonly known as SRP or 7SL), which targets proteins and translocates them across membranes, have the stem symbol *RN7SL* (eg '*RN7SL1*' for 'RNA, 7SL, cytoplasmic 1'); and the RNA component of the enzyme telomerase, which adds TTAGGG DNA sequence repeats to chromosome ends (telomeres) in order to prevent their continual erosion in cell division,[31] is encoded by the telomerase RNA component (*TERC*) gene.

Long non-coding RNAs

The current set of small ncRNAs, as shown above, is necessarily biased towards those with conserved sequence homology, since this feature is used for their computational prediction and classification. There is a further class of ncRNAs — known as long non-coding RNAs (lncRNAs) because they are over 200 nucleotides in length (sometimes even more than 15 kilobases[32]) — that, in the majority of cases, do not share sequence homology with each other. These longer transcripts are spliced, capped and polyadenylated, suggesting that they are expressed and potentially functional within cells. Indeed, some lncRNAs do now have proven functions, and where these are known they have been named accordingly; for instance, '*XIST*' 'X (inactive)-specific transcript (non-protein coding)' is involved in transcriptionally silencing one of the pair of X chromosomes.[33] There are, however, potentially thousands of lncRNAs, and for the vast majority their function remains unresolved. Where lncRNAs reside on the opposite strand to a protein-coding gene, it is thought that they could potentially regulate the expression of the coding gene.[34] These antisense transcripts are named using the approved HGNC symbol for the protein-coding gene with the suffix '-AS' for 'antisense'; the lncRNA gene on the opposite strand to the *BOK* gene is '*BOK-AS1*' for 'BOK antisense RNA 1 (non-protein coding)'. Likewise, those lncRNA genes that reside entirely within an intron of a protein-coding gene are symbolised by the suffix '-IT' for 'intronic transcript' (eg '*MAGI2-IT1*' for 'MAGI2 intronic transcript 1 (non-protein coding)'). There are also lncRNAs that are postulated to function only as transcriptional apparatus for the expression of small ncRNA genes nested within their introns. These 'host genes' are named with the suffix 'HG' (eg '*SNHG1*' for 'small nucleolar RNA host gene 1 (non-protein coding)'). A small number of lncRNAs share homology with each other and are named as paralogues (eg *TTTY1A* and *TTTY1B*). Many transcripts do not fit any of these scenarios, that is: the function of the mature transcript is unknown; they are not proximal to a protein-coding gene; and they are not a member of a homologous family. Such 'orphan' ncRNA genes were previously all named with the anonymous stem symbol *NCRNA#* (eg '*NCRNA00029*') but, in collaboration with the lncRNA database lncRNAdb (http://www.lncrnadb.org)[35] and the Vertebrate Genome Annotation (VEGA; http://vega.sanger.ac.uk/) team,[36] HGNC has recently decided to name these intergenic lncRNA genes with the symbol '*LINC#*' for 'long intergenic non-protein coding RNA #'.

RNA nomenclature across species

Where an equivalent orthologous ncRNA gene can be shown to exist in another species, the human RNA gene nomenclature could be transferred directly to the other species and, indeed, this is already happening for highly conserved classes of small RNAs, such as the microRNAs. For example, mouse *Mir100* is orthologous to human *MIR100*. The nomenclature for other ncRNA classes that have greatly diverged across genomes will need careful annotation and may require species–specific nomenclature.

Conclusion

Recent years have shown us that the human genome does not comprise transcriptional deserts of 'junk' DNA lying between protein-coding genes, but rather that these regions encode thousands of transcribed ncRNAs that may play crucial roles in vital biological processes. As a result, interest in these RNAs is growing quickly, and HGNC aims to keep apace with the discovery of new ncRNA classes to ensure it can provide a robust and systematic nomenclature for these intriguing genes. All human RNA genes named to date can be found at the HGNC RNA webpage (http://www.genenames.org/rna).

Acknowledgments

A number of researchers have provided invaluable support in naming various classes of RNAs: Sam Griffiths-Jones with miRNAs, Michel Weber with snoRNAs, Todd Lowe with tRNAs, Peter Stadler with U7 and vault RNAs, Shipra Agrawal with piRNAs and John Mattick with lncRNAs. Particular thanks go to former HGNC team member Kate Sneddon, and also to current colleagues Ruth Seal, Susan Gordon, Michael Lush and Louise Daugherty.

The work of HGNC is supported by National Human Genome Research Institute (NHGRI) grant P41 HG03345 and Wellcome Trust grant 081979/Z/07/Z.

References

1. Watanabe, H., Fujiyama, A., Hattori, M., Taylor, T.D. *et al.* (2004), 'DNA sequence and comparative analysis of chimpanzee chromosome 22', *Nature* Vol. 429, pp. 382–388.
2. Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R. *et al.* (2007), 'Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project', *Nature* Vol. 447, pp. 799–816.
3. Taft, R.J., Pang, K.C., Mercer, T.R., Dinger, M. *et al.* (2010), 'Non-coding RNAs: Regulators of disease', *J. Pathol.* Vol. 220, pp. 126–139.
4. Seal, R.L., Gordon, S.M., Lush, M.J., Wright, M.W. and Bruford, E.A. (2011), 'genenames.org: The HGNC resources in 2011', *Nucleic Acids Res.* Vol. 39 (Database issue), pp. D514–D519.
5. Lee, R.C., Feinbaum, R.L. and Ambros, V. (1993), 'The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14', *Cell* Vol. 75, pp. 843–854.
6. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. and Tuschl, T. (2001), 'Identification of novel genes coding for small expressed RNAs', *Science* Vol. 294, pp. 853–858.
7. Bartel, D.P. (2009), 'MicroRNAs: Target recognition and regulatory functions', *Cell* Vol. 136, pp. 215–233.
8. Mencia, A., Modamio-Hoybjor, S., Redshaw, N., Morin, M. *et al.* (2009), 'Mutations in the seed region of human miR-96 are responsible for non-syndromic progressive hearing loss', *Nat. Genet.* Vol. 41, pp. 609–613.
9. Kozomara, A. and Griffiths-Jones, S. (2011), 'miRBase: integrating microRNA annotation and deep-sequencing date', *Nucleic Acids Res.*. Vol. 39 (Database issue), pp. D152–D157
10. Chan, P.P. and Lowe, T.M. (2009), 'GtRNAdb: A database of transfer RNA genes detected in genomic sequence', *Nucleic Acids Res.* Vol. 37 (Database issue), pp. D93–D97, PMID: 18984615.
11. Lowe, T.M. and Eddy, S.R. (1997), 'tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence', *Nucleic Acids Res.* Vol. 25, pp. 955–964.
12. Helm, M., Brule, H., Friede, D. and Giege, R. (2000), 'Search for characteristic structural features of mammalian mitochondrial tRNAs', *RNA* Vol. 6, pp. 1356–1379.
13. Putz, J., Dupuis, B., Sissler, M. and Florentz, C. (2007), 'Mamit-tRNA, a database of mammalian mitochondrial tRNA primary and secondary structures', *RNA* Vol. 13, pp. 1184–1190.
14. Korostelev, A., Ermolenko, D.N. and Noller, H.F. (2008), 'Structural dynamics of the ribosome', *Curr. Opin. Chem. Biol.* Vol. 12, pp. 674–683.
15. Hillis, D.M. and Dixon, M.T. (1991), 'Ribosomal DNA: Molecular evolution and phylogenetic inference', *Q. Rev. Biol.* Vol. 66, pp. 411–453.
16. Pruesse, E., Quast, C., Knittel, K., Fuchs, B.M. *et al.*(2007), 'SILVA: A comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB', *Nucleic Acids Res.* Vol. 35, pp. 7188–7196.
17. Ritchie, D.B., Schellenberg, M.J. and MacMillan, A.M. (2009), 'Spliceosome structure: Piece by piece', *Biochim. Biophys. Acta* Vol. 1789, pp. 624–633.
18. Will, C.L. and Luhrmann, R. (2005), 'Splicing of a rare class of introns by the U12-dependent spliceosome', *Biol. Chem.* Vol. 386, pp. 713–724.
19. Decatur, W.A. and Fournier, M.J. (2003), 'RNA-guided nucleotide modification of ribosomal and other RNAs', *J. Biol. Chem.* Vol. 278, pp. 695–698.
20. 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.* Vol. 34 (Database issue), pp. D158–D162.
21. Darzacq, X., Jady, B.E., Verheggen, C., Kiss, A.M. *et al.* (2002), 'Cajal body-specific small nuclear RNAs: A novel class of 2′-O-methylation and pseudouridylation guide RNAs', *EMBO J.* Vol. 21, pp. 2746–2756.
22. Lin, H. (2007), 'piRNAs in the germ line', *Science* Vol. 316, p. 397.
23. Sai Lakshmi, S. and Agrawal, S. (2008), 'piRNABank: A web resource on classified and clustered Piwi-interacting RNAs', *Nucleic Acids Res.* Vol. 36 (Database issue), pp. D173–D177.
24. Reiner, R., Ben-Asouli, Y., Krilovetzky, I. and Jarrous, N. (2006), 'A role for the catalytic ribonucleoprotein RNase P in RNA polymerase III transcription', *Genes Dev.* Vol. 20, pp. 1621–1635.
25. Esakova, O. and Krasilnikov, A.S. (2010), 'Of proteins and RNA: The RNase P/MRP family', *RNA* Vol. 16, pp. 1725–1747.

26. Mowry, K.L. and Steitz, J.A. (1987), 'Identification of the human U7 snRNP as one of several factors involved in the 3' end maturation of histone premessenger RNA's', *Science* Vol. 238, pp. 1682–1687.
27. Marz, M., Mosig, A., Stadler, B.M. and Stadler, P.F. (2007), 'U7 snRNAs: A computational survey', *Genomics Proteomics Bioinformatics* Vol. 5, pp. 187–195.
28. Stadler, P.F., Chen, J.J., Hackermuller, J., Hoffmann, S. *et al*. (2009), 'Evolution of vault RNAs', *Mol. Biol. Evol.* Vol. 26, pp. 1975–1991.
29. Steiner, E., Holzmann, K., Elbling, L., Micksche, M. *et al.* (2006), 'Cellular functions of vaults and their involvement in multidrug resistance', *Curr. Drug Targets* Vol. 7, pp. 923–934.
30. Peterlin, B.M. and Price, D.H. (2006), 'Controlling the elongation phase of transcription with P-TEFb', *Mol. Cell* Vol. 23, pp. 297–305.
31. Feng, J., Funk, W.D., Wang, S.S., Weinrich, S.L. *et al.* (1995), 'The RNA component of human telomerase', *Science* Vol. 269, pp. 1236–1241.
32. Khachane, A.N. and Harrison, P.M. (2010), 'Mining mammalian transcript data for functional long non-coding RNAs', *PLoS One* Vol. 5, p. e10316.
33. Brown, C.J., Ballabio, A., Rupert, J.L., Lafreniere, R.G. *et al*. (1991), 'A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome', *Nature* Vol. 349, pp. 38–44.
34. Faghihi, M.A. and Wahlestedt, C. (2009), 'Regulatory roles of natural antisense transcripts', *Nat. Rev. Mol. Cell Biol*. Vol. 10, pp. 637–643.
35. Amaral, P.P., Clark, M.B., Gascoigne, D.K., Dinger, M.E. and Mattick, J.S. (2011), 'lncRNAdb: A reference database for long noncoding RNAs', *Nucleic Acids Res*. Vol. 39 (Database issue), pp. D146–D151.
36. Wilming, L.G., Gilbert, J.G., Howe, K., Trevanion, S. *et al*. (2008), 'The vertebrate genome annotation (Vega) database', *Nucleic Acids Res*. Vol. 36 (Database issue), pp. D753–D760.