Review TRENDS in Genetics Vol.21 No.5 May 2005

Full text provided by www.sciencedirect.com

Non-coding RNAs: hope or hype?

Alexander Hüttenhofer[1], Peter Schattner[2] and Norbert Polacek[1]

[1]Division of Genomics and RNomics, Innsbruck Medical University-Biocenter, Fritz-Pregl-Strasse 3, 6020 Innsbruck, Austria
[2]Center for Biomolecular Science and Engineering, University of California, Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA

The past four years have seen an explosion in the number of detected RNA transcripts with no apparent protein-coding potential. This has led to speculation that non-protein-coding RNAs (ncRNAs) might be as important as proteins in the regulation of vital cellular functions. However, there has been significantly less progress in actually demonstrating the functions of these transcripts. In this article, we review the results of recent experiments that show that transcription of non-protein-coding RNA is far more widespread than was previously anticipated. Although some ncRNAs act as molecular switches that regulate gene expression, the function of many ncRNAs is unknown. New experimental and computational approaches are emerging that will help determine whether these newly identified transcription products are evidence of important new biochemical pathways or are merely 'junk' RNA generated by the cell as a by-product of its functional activities.

Introduction

RNAs are split into two distinct classes: messenger RNAs (mRNAs), which are translated into proteins, and the non-protein-coding RNAs (ncRNAs), which function at the RNA level. For many years it was believed that there were only a few ncRNAs, and they (e.g. tRNAs, rRNAs and spliceosomal RNAs) were considered accessory components to aid protein functioning. To some degree, these beliefs were fostered by the time-consuming and laborious techniques required to identify these RNAs experimentally, and by the lack of sequenced genomes and appropriate bioinformatics approaches needed to detect them computationally. Thus, identification of novel ncRNA species and elucidation of their function occurred rather by chance than by systematic screens. Hence, even large RNA classes, such as snoRNAs and microRNAs, remained undetected for many years.

Nevertheless, over time it became apparent that there are numerous ncRNAs, and that their cellular functions – on their own or in protein complexes – are varied and important (for reviews, see Refs [1–5]). In the past few years, new experimental strategies, termed 'experimental RNomics', were developed that demonstrated that the number of ncRNAs in genomes of model organisms is much greater than was previously anticipated (Box 1). The application of experimental RNomics from *Escherichia coli* to *Homo sapiens* resulted in the identification of numerous novel ncRNA candidates. However, the function of approximately half of these ncRNA candidates could not be deduced because they lacked the sequence or structural motifs that would have enabled their assignment to an existing ncRNA class [6–13]. Meanwhile, new computational approaches (Box 2) have also detected experimentally verified ncRNAs, particularly in the compact genomes of species of Bacteria [14–16] and Archaea [17].

These results have fuelled speculation that ncRNAs might be important to understanding the increased complexity observed in mammals, because mammalian genomes have only slightly more protein-coding genes than 'lower organisms' such as flies or worms [4]. It remains to be seen whether the current hope and excitement surrounding the discovery of novel ncRNAs is well deserved or whether all of the hype will soon vanish.

ncRNAs: the present

Much of the recent research on ncRNAs has focused on improving our understanding of the functions of two large classes of ncRNAs: (i) small nucleolar RNAs (snoRNAs); and (ii) the microRNA (miRNA) and small interfering RNA (siRNA) family, in addition to the identification of new ncRNA candidates that apparently do not belong to any known ncRNA family.

Small nucleolar RNAs: new targets and cellular locations are emerging

The snoRNA family is divided into two subclasses: box C/D and H/ACA snoRNAs, which direct site-specific 2′-O-ribose methylation and pseudouridylation of target RNAs, respectively [18,19]. snoRNAs have been identified by experimental and computational means. For organisms with compact genomes, the detection of novel snoRNAs by computational methods has been impressive. For example, in the yeast *Saccharomyces cerevisiae*, virtually all snoRNAs that are involved in rRNA modification have now been identified computationally [20,21]. By contrast, in mammals most C/D and nearly all H/ACA snoRNAs have been identified by experimental RNomics approaches on nucleolar or total cellular RNA from human or mouse, respectively [7,19].

Initially, the targets for snoRNA-mediated modifications appeared to be restricted to rRNA, and their only known subcellular location, in Eukarya, was the nucleolus. However, recently the range of targets has been extended to snRNAs and tRNAs. Moreover, those eukaryal

Corresponding authors: Hüttenhofer, A. (alexander.huettenhofer@uibk.ac.at), Schattner, P. (schattner@cse.ucsc.edu).

Available online 23 March 2005

www.sciencedirect.com 0168-9525/$ - see front matter © 2005 Elsevier Ltd. All rights reserved. doi:10.1016/j.tig.2005.03.007

Box 1. Experimental screens for identification of non-coding RNAs

'RNomics' is an area of research that seeks to identify ncRNA genes by experimental methods. This can be achieved by either generating specialized cDNA libraries encoding ncRNAs (Figure I) or micro-array techniques. For the generation of specialized cDNA libraries, a size selection of total, phenol-extracted-RNA derived from cells or tissues of a model organism is achieved by denaturing gel electrophoresis. For the majority of these approaches, a size selection of 50–500 nt is performed because small ncRNA species usually exhibit sizes <500 nt [3,84]. For identification of miRNAs and siRNAs, RNAs that are ~18–25 nt are selected specifically by gel electrophoresis. Alternatively, ribonucleo-protein particles (RNPs) can be purified from cells and immuno-precipitated with specific antibodies directed against known RNA-binding proteins, followed by phenol extraction (Figure I). This enables identification of a specific class of ncRNAs that associate with specific RNA-binding protein(s) [19].

Different strategies can be employed to convert isolated ncRNAs into cDNAs. First, reverse transcription of small ncRNAs (which usually lack a polyA tail) requires the addition of a linker fragment of known sequence to the 3′-end of the ncRNA. This can be achieved by an oligonucleotide (RNA or DNA), which is ligated to the 3′-end of ncRNAs by RNA T4 Ligase. Alternatively, polyA polymerase can be used to add a polyC tract to the 3′-ends of the ncRNA fraction. Subsequently, a second linker fragment of known sequence is ligated to the 5′-end of ncRNAs.

Linker-ligated ncRNAs or polyC tailed and linker ligated ncRNAs are reverse transcribed into cDNA by RT–PCR. The resulting cDNAs are cloned into a vector (e.g. pGEMT) and sequenced. To avoid redundant expression of already known ncRNA species (e.g. tRNAs, small rRNAs or snRNAs), cDNA clones can be spotted on filters in high-density arrays and screened with oligonucleotide probes directed against the most abundant known ncRNA species [84].

Initially, cDNA sequences are analysed by bioinformatical methods and database searches (e.g. BlastN). Second, the genomic localization of potential novel ncRNA genes (i.e. novel ncRNA 'candidates') is followed by analysis of their temporal, developmental and tissue-specific expression by northern-blot analysis. Third, sub-cellular localization of ncRNAs can be analysed by using sub-fractionated total RNA (cytoplasmic or nuclear fraction) in northern-blot analysis. Finally, in selected cases, affinity chromatography, using the *in vitro* transcribed RNA as 'bait', can identify proteins that bind to the ncRNA 'candidate'. By loading protein fractions onto the purification column, ncRNA-binding proteins can be isolated and micro-sequenced.

Although none of these steps, by itself, is likely to determine the function of a novel ncRNA, taken together they might hint at its cellular function. Ultimately, only the genomic deletion analysis of an ncRNA gene, which is currently a time-consuming method, can identify the function of an ncRNA in a cell *in vivo*.

'snoRNAs' that target snRNAs have been found to localize to the Cajal bodies (intra-nuclear structures involved in snRNA processing), belying their designation as nucleolar RNAs [19,22]. In addition, a growing number of so-called 'orphan' snoRNAs, which do not appear to target any ncRNA species, have been isolated independently by several screens [7,18,19]. Orphan snoRNA specimens might function distinctly, for example, as RNA chaperones or by guiding modification of other cellular RNAs. In this context, it was proposed that a brain-specific, human snoRNA might target a protein encoding mRNA (the serotonin receptor 5-HT2c mRNA); however, this still awaits experimental confirmation [6].

The rapidly expanding world of miRNAs and siRNAs

miRNAs and siRNAs are the second relatively new class of ncRNAs [23–29]. They have a similar size (~21 nucleotides) and a similar mechanism of generation. Both miRNAs and siRNAs are excised from longer, double-stranded-RNA-precursor molecules by an RNase III-like enzyme, Dicer [30–32]). Until recently, miRNAs and siRNAs had only been identified in Eukarya [26]. However, now, viral encoded miRNAs have also been detected, including one that can target viral DNA polymerase [33].

Both miRNAs and siRNAs function by an antisense-based mechanism, similar to that observed for snoRNAs. In contrast to snoRNAs, which only target other ncRNA species, miRNA and siRNAs target protein-coding mRNAs. In vertebrates, for example, miRNAs are thought to inhibit the translation of target mRNAs through partial complementarity to their 3′-untranslated region (UTR) by a currently unknown mechanism [34,35], whereas siRNAs that exhibit full complementarity to target mRNAs direct

Box 2. Computational screens for identification of non-coding RNAs

Computational identification of ncRNAs has been attempted for at least 25 years – almost as long as protein-coding genefinders have been used. However, computational detection of ncRNAs is far more difficult than detection of protein-coding genes. ncRNAs often lack the characteristic features used by genefinders for protein-coding genes (e.g. start and stop codons, synonymous codon degeneracies, open reading frames, splice sites, polyadenylation sites and proximity to CpG islands). In contrast to protein-coding genes, ncRNA genes are typically short, have widely varying motifs and are often characterized more by their secondary structure than by their primary sequence.

Consequently, most successful ncRNA genefinders have been custom-designed programs that search for a single class of ncRNAs. Examples include tRNAScan-SE for tRNAs (http://lowelab.ucsc.edu/tRNAscan-SE/), snoscan for C/D Box snoRNAs (http://lowelab.ucsc.edu/snoscan/), snoGPS for H/ACA snoRNAs (http://lowelab.ucsc.edu/snoGPS/) and mirscan (http://genes.mit.edu/mirscan/) and other programs for microRNAs. By contrast, some ncRNA genefinding programs are designed to be reconfigurable – so that the user can specify a set of ncRNA motifs to search for. Programs of this class include RNAmotif (http://www.scripps.edu/mb/case/casegr-sh-3.5.html), Erpin (http://tagc.univ-mrs.fr/erpin/) and Patsearch (http://www.ba.itb.cnr.it/BIG/PatSearch/). However, such reconfigurable programs typically have less sensitivity and specificity than customized ncRNA genefinders. In practice, few experimentally verified ncRNAs have been found with reconfigurable ncRNA genefinders [85].

Programs that attempt to identify ncRNAs for which no *a priori* model of sequence or secondary structure is available have also been attempted. One example is QRNA a program that searches for pairs of secondary-structure-conserving mutations between homologous sequences (www.genetics.wustl.edu/eddy/software/) [14]. QRNA has detected ncRNAs in *Escherichia coli*, *Saccharomyces cerevisiae* and *Pyrococcus furiosus* that were subsequently verified experimentally. Other genefinder programs exploit specific base-composition variations characteristic of ncRNAs (e.g. GC%). Such programs have had modest success in detecting novel ncRNAs, particularly in hyperthermophiles. However, computational detection without sequence or secondary-structure models is difficult, and such programs often miss known ncRNAs. For original references on these programs and additional details on the development of computational genefinders for ncRNAs, see Ref [85].

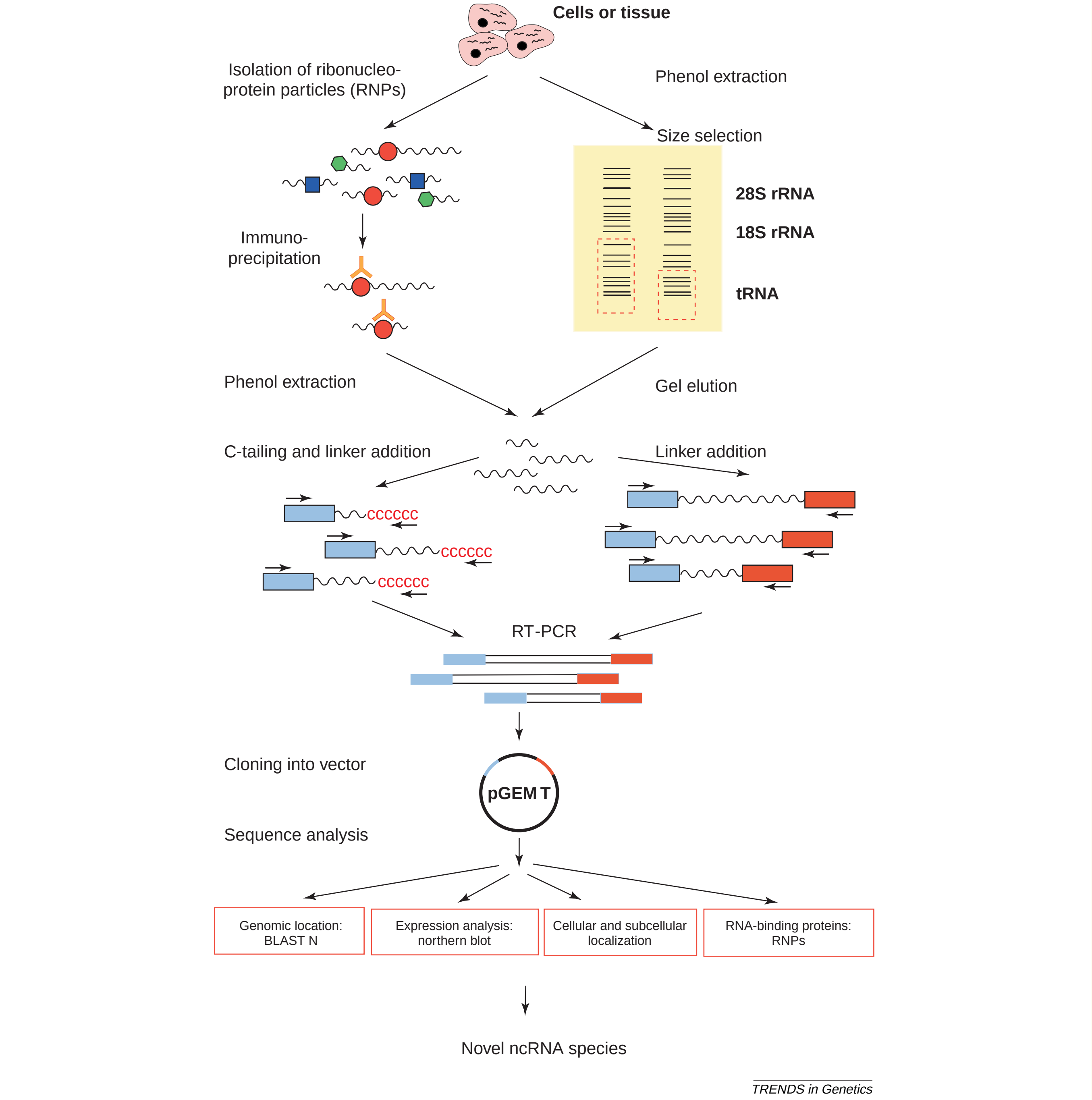


Figure I. The construction of an unbiased, general purpose cDNA library (upper right) of ncRNAs or a more specialized library encoding RNAs form a specific ncRNA subclass, for example, snoRNAs (upper left). Sub-class enrichment is achieved by immuno-precipitation of a cell lysate with an antibody directed against an RNA-binding protein that is specific to the subclass (indicated by red circles; green hexagons and blue squares represent other RNA-binding proteins). Two different methods of generating a cDNA library enriched for ncRNAs are illustrated in the middle part of the figure. The C-tailing and linker-ligation approach is shown on the left and a linker-ligation protocol is shown on the right [84].

mRNA cleavage; this mechanism is designated as RNA interference (RNAi). Thus, both subclasses of siRNAs and miRNAs can be considered as molecular switches, which can regulate gene expression. Because siRNAs can be administered to cells by lipotransfection [36], the therapeutic application of siRNAs as a tool to treat human diseases by knocking down the expression of certain disease related genes is currently being investigated, with promising results [37,38].

siRNAs and miRNAs do not exert their function as naked RNA; they work in combination with RNA-binding proteins, forming a ribonucleo-protein complex known as RNA induced silencing complex (RISC). RISC contains a single strand of the miRNA–siRNA duplex, which targets mRNAs by base complementarity [32]. Recently, the enzyme required for mRNA cleavage by siRNAs has been identified as one of the integral components of the RISC complex (argonaute or Ago2) [39].

Whole-genome screens for miRNAs and siRNAs were initially based on experimental RNomics approaches: by cloning the 21-nt RNA fraction from total RNA in model organisms, thus generating specialized cDNA libraries (Box 1). Later, it became evident that miRNAs were processed from distinct stem-loop structures of pre-miRNA precursors. This observation enabled the development of new miRNA search algorithms that could be applied to genomic sequences [40–42]. Using these algorithms, the number of miRNAs in *H. sapiens* has now reached ~250 (i.e. ~1% of the number of human protein-coding genes) [26]. In addition to finding novel miRNAs, bioinformatical analysis has led to the identification of potential target sites within the 3′-UTRs of mRNAs [42–46]. Interestingly, in vertebrates single miRNAs might have as many as 100 different mRNA targets; therefore, it is possible that ~10 000 protein-coding genes in humans are regulated by miRNAs [47]

Detecting other classes of ncRNAs

Numerous RNAs have been identified that do not belong to any of the ncRNA families of known function (tRNAs, rRNAs, snRNAs, snoRNAs, miRNAs and siRNAs). Most of these – particularly those identified in experimental RNomics screens – are short (i.e. <500 nt). Although some have known functions [48], for the majority – especially for the more recently discovered ncRNAs – no function has been elucidated [5–11,13]. Indeed, it is not even known whether these RNA transcripts have any function, and consequently they are more appropriately considered as ncRNA 'candidates'.

In addition to the short ncRNAs, numerous long ncRNAs have been detected that consist of several thousand nucleotides or more. In Eukarya, these RNAs are believed to be transcribed from Pol II promoters, and sometimes are alternatively spliced and/or polyadenylated and, therefore, might have evolved from genes that have lost their protein-coding capacity. Some of these long ncRNAs function in regulatory mechanisms such as dosage-compensation of the sex chromosomes [e.g. X-inactive specific transcript (*Xist*), RNA antisense to Xist (*Tsix*), RNA on X1 (*roX1*) and RNA on X2 (*roX2*)]. Several others [e.g. antisense to Igf2r (*Air*), *H19,* long QT intronic transcript 1 (*LIT1*) and E6-AP unbiquintin protein ligase antisense transcript (*UBE3A-ATS*)] have been implicated in the regulation of imprinted genes in mammals.

An intriguing example of an imprinted ncRNA with no known function is *UBE3A-ATS*. This ncRNA is a 460-kb long transcript containing >100 exons and its introns include ~80 snoRNA genes [6,49]. Notably, most of the known antisense RNAs found in imprinted domains are expressed on the paternal chromosome, whereas all methylation-based imprints are observed on the maternal allele, leading to the speculation that ncRNAs could have evolved as a paternal alternative to the DNA-methylation imprint [50].

Other long ncRNAs that are thought to be of functional importance for cell viability are expressed from both chromosomes. The exact roles of these RNAs remain largely elusive but potential functions for some of them have been suggested [51,52]. Those ncRNAs that are involved in human diseases are particularly interesting. NcRNAs have been linked to human neurodegenerative disease (spinocerebellar ataxia type 8 (SCA8); [53]), hereditary hemochromatosis [54], schizophrenia [55] and lung cancer [56]. It is not yet clear whether there are additional large families of functional ncRNAs that have not yet been identified. This might be the case for ncRNA classes that are only present in a limited number of organisms. One example of such a species-specific family is the guide RNAs (gRNAs) of *Trypanosome* mitochondria [57]. Another example was found recently in the slime mold *Dictyostelium discoideum.* In this case, 14 representatives of a novel class of RNAs exhibiting conserved sequence and secondary-structure motifs were identified by an experimental RNomics approach [58]. Interestingly, all members of this RNA class are developmentally regulated, analogous to some miRNAs.

Other potentially new classes of ncRNAs include the antisense RNAs, transcribed pseudogenes and riboswitch-related ncRNAs. Recent evidence indicates that ~12% of mammalian ncRNA transcription is antisense to some other known gene [59]. It is not known to what extent these antisense transcripts are functional. However, many antisense transcripts are subject to additional RNA processing such as splicing or RNA editing. Moreover, some antisense transcripts have been implicated in gene regulation at the level of imprinting, transcriptional interference, RNAi or methylation modification [60]. In addition, variations in antisense-transcription levels appear to be correlated with certain disease states in humans [60]. This evidence combined with the overall level of antisense transcription suggests that many of these ncRNAs might be important in regulating gene expression.

An unexpected new family of ncRNA candidates is the expressed pseudogenes. The human genome is estimated to contain up to 20 000 pseudogenes and it has been predicted that ~3% of them are transcribed [61]. In the mouse, ~5000 processed pseudogenes have been identified and, to date, 48 have been detected in expressed cDNA libraries [62,63]. The first examples of functional, expressed pseudogenes were recently reported: these pseudogenes either regulate the mRNA stability or regulate the translation of their homologous coding genes [61,64]. It seems probable that these are not isolated examples and that other expressed pseudogenes will prove to be functional.

Finally, the 'riboswitches' are ligand-binding regulatory domains in the 5′-UTRs of certain eubacterial mRNAs. Following ligand-binding, a conformational change occurs that ultimately results in a change in the expression of the downstream gene(s) [65]. In a broader sense, riboswitches can be viewed as 'mRNA-enslaved' ncRNAs because they execute their cellular function entirely at the RNA level. In support of this view, a novel riboswitch class was recently identified that functions as a metabolite-induced ribozyme and acts independently of the downstream mRNA sequences to which it is attached [66]. It remains to be seen whether the concept of ligand-induced-ncRNA activation is a more common phenomenon and if it also works in eukaryal species to regulate cell metabolism.

ncRNAs: the future

The number of known ncRNAs and putative ncRNAs of unknown function has increased dramatically in the past few years (Figure 1). Moreover, particularly in higher eukaryotes, there is still room in the genome for the discovery of novel ncRNA genes. Only a fraction of the genome (i.e. ~1.4% in humans) is translated into proteins, whereas ~27% is transcribed as introns and UTRs but not translated [48,59,62,67–69]. In addition, ~25% of mammalian genomes are predicted to be transcribed but not translated [52], further increasing the space for potential novel ncRNA genes (Figure 2).

We still do not know whether we have identified all of the ncRNAs, even in well-studied organisms with small genomes such as *S. cerevisiae* or *Escherichia coli* let alone in far more complex and larger genomes such as *H. sapiens*. Recent data suggest that we might still have only seen the 'tip of the ncRNA iceberg'. For example, data from micro-array experiments with whole-chromosome tiling [59] and from genome-wide, full-length-cDNA libraries [62,70,71] suggest that the number of transcribed ncRNAs is far greater than previously thought. However, there is still considerable debate as to whether the recently detected transcripts are functional ncRNAs or merely some kind of 'transcriptional noise'.

This debate has occurred partly because the term 'transcriptional noise' is used in different, and not always clearly defined, ways in the literature. In some cases, transcriptional noise refers to experimental artifacts such as genomic contaminants, incomplete intron digests of protein-coding genes or non-specific hybridization that appear as cellular transcription. We will refer to these as transcriptional artifacts. By contrast, the term transcriptional noise is also sometimes used to indicate genuine cellular transcription that, however, results in transcripts with no biological function or activity; to avoid confusion, we will refer to this as non-functional transcription. Finally, there is the original definition of 'transcriptional noise' – variations in the production level of a functional transcript among identical cells in identical environments [72].

The case for the recently detected non-coding transcripts not being transcriptional artifacts is strong. These transcripts generally have polyA tails and nearby, experimentally verified, transcription-factor-binding sites [59,73]. Many of these transcripts appear to be spliced; some have been verified by northern-blot analysis or quantitative PCR, thereby demonstrating their relative high abundance and making it difficult to dismiss them as experimental artifacts or mRNA-degradation fragments [59]. Although some of these transcripts might have protein-coding regions that had not been identified previously [74], many do not appear to have any significant coding potential.

Are the newly identified ncRNA transcripts functional?

Whether the recent data demonstrate functional transcription is less clear. On the one hand, many of the transcripts are spliced or differentially expressed, suggesting that the cell is devoting resources not only to ensure their production but also for their regulation. Moreover, a few functional ncRNAs, involved in genetic regulation through sequence homology or complementarity to protein-coding genes, have been identified among large classes of detected transcripts (e.g. pseudogenes and

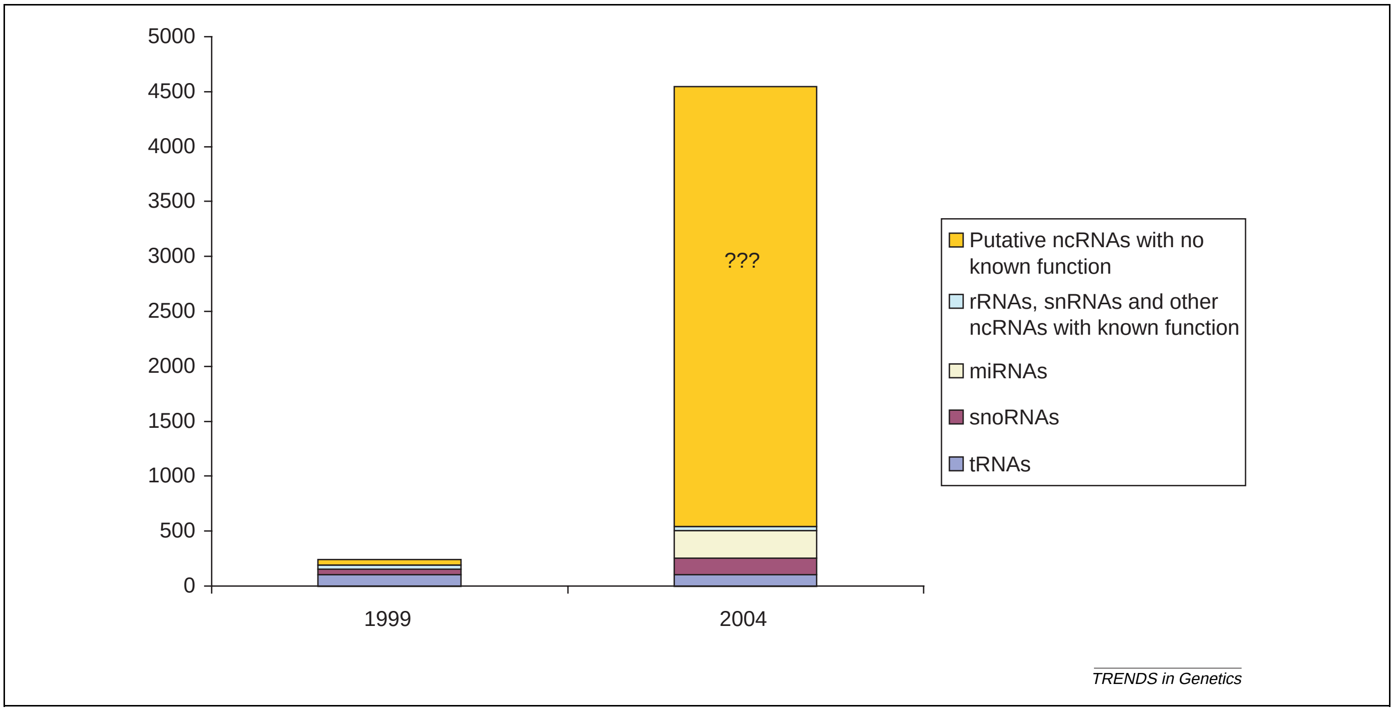


Figure 1. The rapidly increasing number of mammalian ncRNAs and ncRNA candidates from 1999 to 2004. Estimates of sizes of families of known RNAs based on experimental or computational studies of tRNAs, rRNAs, snRNAs [68], snoRNAs [19,68], miRNA and siRNAs [26]. Identical copies of rRNA, tRNA and snRNA genes are not included in these estimates. The numbers for ncRNA candidates are estimated from data from mouse cDNA-transcript libraries [3,70]. The question marks reflect the current uncertainty regarding the proportion of these transcripts that are actually functional.

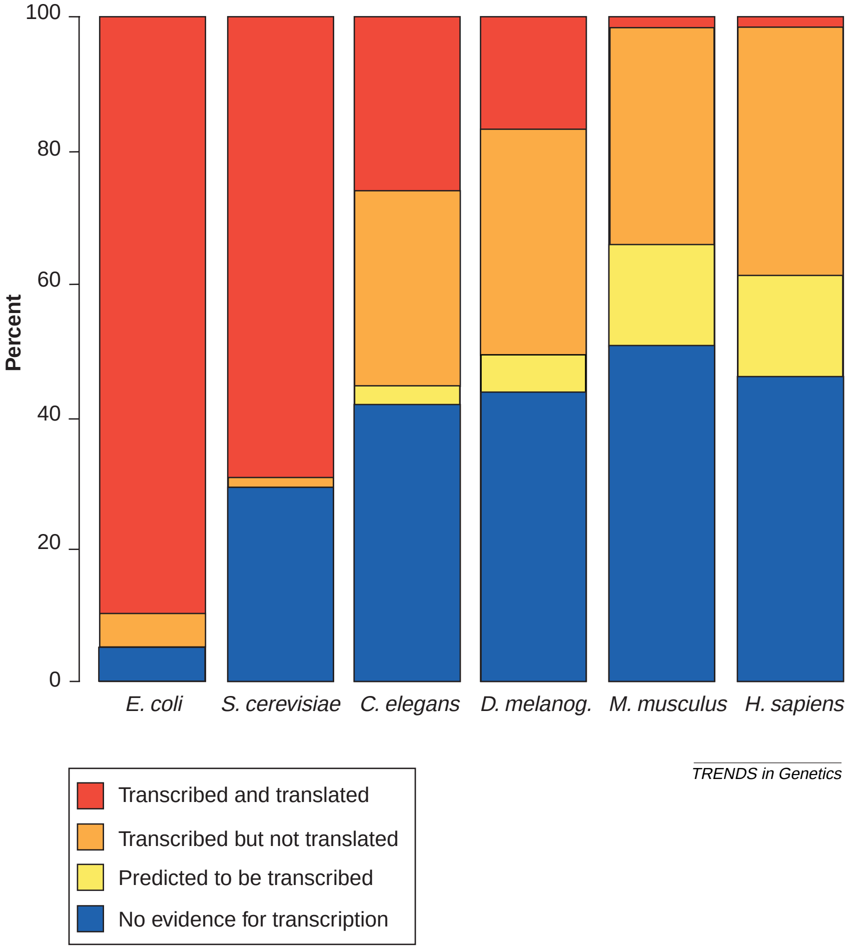


Figure 2. Genomic space for the discovery of novel ncRNAs in higher eukaryotes. Estimated sizes of RNA fractions of representative bacterial or eukaryal genomes, which are either protein-coding or non-protein coding, are given as percentages of the total size of the respective genome. The protein-coding estimates were obtained from computational gene-finding programs applied to completed genomic-sequence data [48,67,68,83]. Transcriptional fractions for bacteria are estimated from Ref. [83] and from unpublished data (J. Vogel, personal communication). For mammals, transcription estimates are based on tiling-microarray- and cDNA library-generation experiments [52,59,62] in addition to unpublished data (M. Pheasant and J.S. Mattick, personal communication). Tiling-array data from human chromosomes 21 and 22 have been extrapolated to the entire human genome. *Drosophila melanogaster* has been shortened to *D. melanog.* because of space restrictions.

antisense transcripts). It is possible that other ncRNAs regulate protein-coding genes in a similar manner. More remarkable is the recent example of genetic regulation by a transcribed ncRNA that does not even depend on the sequence of the transcript. At the genomic location described in this example [75], the transcription of essentially any ncRNA is sufficient to interfere with the expression of an adjacent protein-coding gene in a regulated manner. If this phenomenon is more widespread it might explain the function of many ncRNA candidates.

On the other hand, most of these RNA species might not be functional. Perhaps, they are merely random transcription products resulting from the genomic distribution of weak promoter and polyadenylation sites that occur by chance in a three billion nucleotide genome. Possibly, the cell does not ‘switch off’ the promoters that transcribe these ncRNAs, simply because it is too costly to completely downregulate all non-functional transcription. Moreover, because splicing and differential expression are also controlled by short cis-acting sequences that occur by chance even in a random sequence, the existence of spliced or differentially expressed transcripts also does not necessarily prove functionality. Finally, it is possible that the functional ncRNAs that are observed in antisense and pseudogene transcripts are really just isolated cases and not the tip of any ncRNA iceberg.

Some of the current data appear to support this non-functional-transcription hypothesis. If the ncRNA transcripts are functional, we might expect their interspecies conservation patterns to be similar to those of known functional ncRNAs. But that does not appear to be the case. Known functional ncRNAs are highly conserved between mouse and human. Moreover, they typically have conserved secondary structures, resulting in mutual covariance signals that are detectable by programs such as QRNA [14]. However, on average, the novel mouse ncRNA transcripts are no more conserved than random intergenic regions [76]. In addition, among 296 recent human ncRNA candidates, only 47 yielded positive QRNA signals [71]. Of course, it is possible that the new transcripts are functional and are less conserved than the majority of known ncRNAs [77]. However, there is little experimental data to support this hypothesis to date.

Additional clues to the functions of the novel ncRNA transcripts might soon emerge from microarray experiments. These experiments can be used to cluster transcripts on the basis of their expression patterns in different tissues, during different developmental stages, in varying cellular environments or disease states [78]. When combined with immuno-precipitation (IP) enrichment (Figure I in Box 1), microarray detection of ncRNAs should facilitate the identification of RNA-binding proteins (RNABPs) that are associated with specific ncRNAs. So far, this approach for genome-wide association of RNAs and RNABPs has only been applied to mRNAs [79]. Applying microarray detection with IP enrichment to ncRNAs is more challenging because ncRNAs typically have more internal secondary structure than mRNAs; however, it should still be feasible as demonstrated by a recent study using a microarray that tiled yeast ncRNAs [80].

Differential expression or RNABP binding does not prove transcript function. In principle, an alternative approach to testing the non-functional-transcription hypothesis could involve introducing large, artificial, random DNA sequences into a mammalian genome and investigating to what extent these sequences are transcribed. However, although conceptually appealing, such an approach would be challenging to implement. Moreover, it would require a credible underlying evolutionary model to design the artificial DNA sequence. In particular, the results of such an experiment would depend on the distribution of promoter and polyA sites in the artificial sequence. Because genome evolution has involved multiple duplications and translocations, the number of such sites in non-functional, or no longer functioning, parts of the genome might be different from the distributions predicted by a purely random-sequence model.

Ultimately, the *in vivo* functions of ncRNAs can only be assessed by gene knockout experiments. Clearly, testing thousands of transcript knockouts is a daunting task. However, recent experiments involving megabase deletions in mouse [81] indicate that at least small-scale experiments of this nature are feasible. In these experiments, 2.3 Mb of mouse genome – including sequence encoding four of the 11 665 novel ncRNA transcripts described in Ref. [62] – were deleted with no observable phenotypic consequences. Extrapolating a background rate for non-functional transcription from such limited data is clearly premature. However, with additional megabase-deletion data, an estimation of the level of background, cellular non-functional-genomic transcription might be feasible.

An alternative to ncRNA-transcript knockouts might be the systematic inactivation of ncRNA transcripts by RNAi 'knock-down'. Such RNAi knock-down has been shown to enable a medium-to high-throughput analysis of all of the protein-coding genes in a cell [27]. RNAi might not work as well for ncRNAs because many ncRNAs are located in the nucleus or nucleolus of the cell, whereas current RNAi methods generally function in the cytoplasm. However, recent results on knock-down of snoRNAs in trypanosomes suggest that, at least in some cases, silencing of non-protein-coding RNAs with RNAi is possible [82].

Concluding remarks

What can we conclude from all of the hype on the new ncRNA transcriptional data? Do these results demonstrate – as has been recently proposed – that the 'main output of the genomes of complex organisms is genetically active but non-coding RNA [4]'? Or are these transcripts primarily 'junk' RNA? The honest answer at this point is we still do not know. However, emerging computational and experimental approaches are beginning to lead not only to the identification of all ncRNAs in different model organisms but also to clues about their function. Only when the functions of these ncRNAs have been determined will it be possible to assess the overall contribution of ncRNAs to the genetic activity of the organism.

Acknowledgements

We apologize to all colleagues whose work could not be cited owing to space limitations. We thank Jörg Vogel, John Mattick and Michael Pheasant for providing unpublished data and Todd Lowe for helpful suggestions and for critical reading of the article. We also thank one of the referees for extensive valuable comments and suggestions. This work was supported by a grant from the Austrian Science Foundation, FWF (ZFP 171370), to A.H. and a FWF grant (P16932) to N.P.

References

1 Eddy, S.R. (2002) Computational genomics of noncoding RNA genes. *Cell* 109, 137–140

2 Gottesman, S. (2002) Stealth regulation: biological circuits with small RNA switches. *Genes Dev.* 16, 2829–2842

3 Huttenhofer, A. *et al.* (2002) RNomics: identification and function of small, non-messenger RNAs. *Curr. Opin. Chem. Biol.* 6, 835–843

4 Mattick, J.S. (2004) RNA regulation: a new genetics? *Nat. Rev. Genet.* 5, 316–323

5 Storz, G. *et al.* (2004) Controlling mRNA stability and translation with small, noncoding RNAs. *Curr. Opin. Microbiol.* 7, 140–144

6 Cavaille, J. *et al.* (2000) Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *Proc. Natl. Acad. Sci. U. S. A.* 97, 14311–14316

7 Hüttenhofer, A. *et al.* (2001) RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.* 20, 2943–2953

8 Tang, T.H. *et al.* (2002) Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus. *Proc. Natl. Acad. Sci. U. S. A.* 99, 7536–7541

9 Marker, C. *et al.* (2002) Experimental RNomics: identification of 140 candidates for small non-messenger RNAs in the plant *Arabidopsis thaliana*. *Curr. Biol.* 12, 2002–2013

10 Yuan, G. *et al.* (2003) RNomics in *Drosophila melanogaster*: identification of 66 candidates for novel non-messenger RNAs. *Nucleic Acids Res.* 31, 2495–2507

11 Vogel, J. *et al.* (2003) RNomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res.* 31, 6435–6443

12 Vitali, P. *et al.* (2003) Identification of 13 novel human modification guide RNAs. *Nucleic Acids Res.* 31, 6543–6551

13 Tang, T-H. *et al.* (2005) Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*. *Mol. Microbiol.* 55, 469–481

14 Rivas, E. *et al.* (2001) Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr. Biol.* 11, 1369–1373

15 Argaman, L. *et al.* (2001) Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.* 11, 941–950

16 Wassarman, K.M. *et al.* (2001) Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev.* 15, 1637–1651

17 Klein, R.J. *et al.* (2002) Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc. Natl. Acad. Sci. U. S. A.* 99, 7542–7547

18 Bachellerie, J.P. *et al.* (2002) The expanding snoRNA world. *Biochimie* 84, 775–790

19 Kiss, A.M. *et al.* (2004) Human box H/ACA pseudouridylation guide RNA machinery. *Mol. Cell. Biol.* 24, 5797–5807

20 Lowe, T.M. and Eddy, S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science* 283, 1168–1171

21 Schattner, P. *et al*. (2004) Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res.* 32, 4281–4296

22 Decatur, W.A. and Fournier, M.J. (2003) RNA-guided nucleotide modification of ribosomal and other RNAs. *J. Biol. Chem.* 278, 695–698

23 Ambros, V. (2003) MicroRNA pathways in flies and worms: growth, death, fat, stress, and timing. *Cell* 113, 673–676

24 Nelson, P. *et al*. (2003) The microRNA world: small is mighty. *Trends Biochem. Sci.* 28, 534–540

25 Brennecke, J. and Cohen, S.M. (2003) Towards a complete description of the microRNA complement of animal genomes. *Genome Biol.* 4, 228

26 Bartel, D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 116, 281–297

27 Tuschl, T. and Borkhardt, A. (2002) Small interfering RNAs: a revolutionary tool for the analysis of gene function and gene therapy. *Mol. Interv.* 2, 158–167

28 Meister, G. and Tuschl, T. (2004) Mechanisms of gene silencing by double-stranded RNA. *Nature* 431, 343–349

29 Murchison, E.P. and Hannon, G.J. (2004) miRNAs on the move: miRNA biogenesis and the RNAi machinery. *Curr. Opin. Cell Biol.* 16, 223–229

30 Bernstein, E. *et al*. (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* 409, 363–366

31 Ketting, R.F. *et al*. (2001) Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*. *Genes Dev.* 15, 2654–2659

32 Tijsterman, M. and Plasterk, R.H. (2004) Dicers at RISC; the mechanism of RNAi. *Cell* 117, 1–3

33 Pfeffer, S. *et al*. (2004) Identification of virus-encoded microRNAs. *Science* 304, 734–736

34 Doench, J.G. and Sharp, P.A. (2004) Specificity of microRNA target selection in translational repression. *Genes Dev.* 18, 504–511

35 Pillai, R.S. *et al*. (2004) Tethering of human Ago proteins to mRNA mimics the miRNA-mediated repression of protein synthesis. *RNA* 10, 1518–1525

36 Elbashir, S.M. *et al*. (2001) Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* 411, 494–498

37 Soutschek, J. *et al*. (2004) Therapeutic silencing of an endogenous gene by systemic administration of modified siRNAs. *Nature* 432, 173–178

38 Lorenz, C. *et al*. (2004) Steroid and lipid conjugates of siRNAs to enhance cellular uptake and gene silencing in liver cells. *Bioorg. Med. Chem. Lett.* 14, 4975–4977

39 Liu, J. *et al*. (2004) Argonaute2 is the catalytic engine of mammalian RNAi. *Science* 305, 1437–1441

40 Lim, L.P. *et al*. (2003) Vertebrate microRNA genes. *Science* 299, 1540

41 Lim, L.P. *et al*. (2003) The microRNAs of *Caenorhabditis elegans*. *Genes Dev.* 17, 991–1008

42 Jones-Rhoades, M.W. and Bartel, D.P. (2004) Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol. Cell* 14, 787–799

43 Enright, A.J. *et al*. (2003) MicroRNA targets in *Drosophila*. *Genome Biol.* 5, R1

44 John, B. *et al*. (2004) Human microRNA targets. *PLoS Biol* 2, e363

45 Lewis, B.P. *et al*. (2003) Prediction of mammalian microRNA targets. *Cell* 115, 787–798

46 Rhoades, M.W. *et al*. (2002) Prediction of plant microRNA targets. *Cell* 110, 513–520

47 Bartel, D.P. and Chen, C.Z. (2004) Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. *Nat. Rev. Genet.* 5, 396–400

48 Szymanski, M. *et al*. (2003) Noncoding RNA transcripts. *J. Appl. Genet.* 44, 1–19

49 Runte, M. *et al*. (2001) The IC-SNURF-SNRPN transcript serves as a host for multiple small nucleolar RNA species and as an antisense RNA for UBE3A. *Hum. Mol. Genet.* 10, 2687–2700

50 Reik, W. and Walter, J. (2001) Evolution of imprinting mechanisms: the battle of the sexes begins in the zygote. *Nat. Genet.* 27, 255–256

51 Erdmann, V.A. *et al*. (2001) The non-coding RNAs as riboregulators. *Nucleic Acids Res.* 29, 189–193

52 Mattick, J.S. (2003) Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *BioEssays* 25, 930–939

53 Benzow, K.A. and Koob, M.D. (2002) The KLHL1-antisense transcript (KLHL1AS) is evolutionarily conserved. *Mamm. Genome* 13, 134–141

54 Thenie, A.C. *et al*. (2001) Identification of an endogenous RNA transcribed from the antisense strand of the HFE gene. *Hum. Mol. Genet.* 10, 1859–1866

55 Millar, J.K. *et al*. (2000) Disruption of two novel genes by a translocation co-segregating with schizophrenia. *Hum. Mol. Genet.* 9, 1415–1423

56 Ji, P. *et al*. (2003) MALAT-1, a novel noncoding RNA, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene* 22, 8031–8041

57 Madison-Antenucci, S. *et al*. (2002) Editing machines: the complexities of trypanosome RNA editing. *Cell* 108, 435–438

58 Aspegren, A. *et al*. (2004) Novel non-coding RNAs in *Dictyostelium discoideum* and their expression during development. *Nucleic Acids Res.* 32, 4646–4656

59 Kampa, D. *et al*. (2004) Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.* 14, 331–342

60 Lavorgna, G. *et al*. (2004) In search of antisense. *Trends Biochem. Sci.* 29, 88–94

61 Yano, Y. *et al*. (2004) A new role for expressed pseudogenes as ncRNA: regulation of mRNA stability of its homologous coding gene. *J. Mol. Med.* 82, 414–422

62 Okazaki, Y. *et al*. (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420, 563–573

63 Zhang, Z. *et al*. (2004) Comparative analysis of processed pseudogenes in the mouse and human genomes. *Trends Genet.* 20, 62–67

64 Korneev, S.A. *et al*. (1999) Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. *J. Neurosci.* 19, 7711–7720

65 Mandal, M. and Breaker, R.R. (2004) Gene regulation by riboswitches. *Nat. Rev. Mol. Cell Biol.* 5, 451–463

66 Winkler, W.C. *et al*. (2004) Control of gene expression by a natural metabolite-responsive ribozyme. *Nature* 428, 281–286

67 Venter, J.C. *et al*. (2001) The sequence of the human genome. *Science* 291, 1304–1351

68 Lander, E.S. *et al*. (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921

69 Rodriguez, A. *et al*. (2004) Identification of mammalian microRNA host genes and transcription units. *Genome Res.* 14, 1902–1910

70 Numata, K. *et al*. (2003) Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. *Genome Res.* 13, 1301–1306

71 Imanishi, T. *et al*. (2004) Integrative annotation of 21 037 human genes validated by full-length cDNA clones. *PLoS Biol* 2, e162

72 Blake, W.J. *et al*. (2003) Noise in eukaryotic gene expression. *Nature* 422, 633–637

73 Cawley, S. *et al*. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* 116, 499–509

74 Nekrutenko, A. *et al*. (2003) An evolutionary approach reveals a high protein-coding capacity of the human genome. *Trends Genet.* 19, 306–310

75 Martens, J.A. *et al*. (2004) Intergenic transcription is required to repress the *Saccharomyces cerevisiae* SER3 gene. *Nature* 429, 571–574

76 Wang, J. *et al*. (2004) Mouse transcriptome: neutral evolution of 'non-coding' complementary DNAs. *Nature*, 431. DOI: 10.1038/nature03016 (http://www.nature.com/)

77 Suzuki, M. and Hayashizaki, Y. (2004) Mouse-centric comparative transcriptomics of protein coding and non-coding RNAs. *BioEssays* 26, 833–843

78 Peng, W.T. *et al*. (2003) A panoramic view of yeast noncoding RNA processing. *Cell* 113, 919–933

79 Gerber, A.P. *et al*. (2004) Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS Biol* 2, E79

80 Hiley, S.L. *et al*. (2005) Detection and discovery of RNA modifications using microarrays. *Nucleic Acids Res.* 33, e2

81 Nobrega, M.A. *et al*. (2004) Megabase deletions of gene deserts result in viable mice. *Nature* 431, 988–993

82 Liang, X.H. *et al*. (2003) Small nucleolar RNA interference induced by antisense or double-stranded RNA in trypanosomatids. *Proc. Natl. Acad. Sci. U. S. A.* 100, 7521–7526

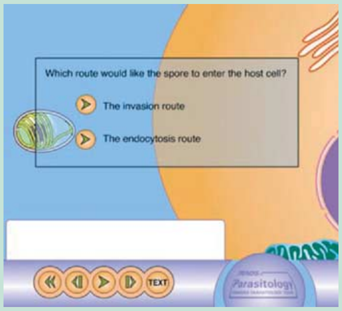
83 Blattner, F.R. *et al*. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science* 277, 1453–1474

84 Huttenhofer, A. *et al*. (2004) Experimental RNomics: a global approach to identifying small nuclear RNAs and their targets in different model organisms. *Methods Mol. Biol.* 265, 409–428

85 Schattner, P. (2003) Computational gene-finding for non-coding RNAs. In *NonCoding RNAs: Molecular Biology and Molecular Medicine* (Barciszewski, J., ed.), pp. 33–49, Kluwer Academic and Plenum Publishers

Getting animated with parasites!

Interested in the molecular cell biology of host–parasite interactions? Then take a look at the online animations produced by *Trends in Parasitology*, one of our companion *TRENDS* journals. The pictures below are snapshots from two of our collection of animations revealing the latest advances in understanding parasite life cycles. Check them out today!



Which route would like the spore to enter the host cell?

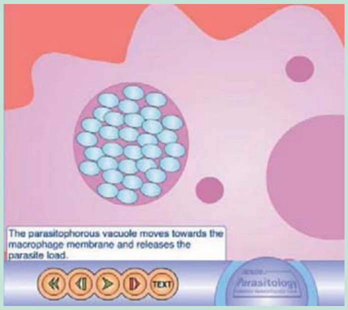
The invasion route

The endocytosis route

Microsporidia: how can they invade other cells?

By C. Franzen [(2004) *Trends Parasitol.* 20, 10.1016/j.pt.2004.04.009]

http://archive.bmn.com/supp/part/franzen.html



The parasitophorous vacuole moves towards the macrophage membrane and releases the parasite load.

Interaction of *Leishmania* with the host macrophage

By E. Handman and D.V.R. Bullen [(2002) *Trends Parasitol.* 18, 332–334]

http://archive.bmn.com/supp/part/swf012.html