Evolutionary Patterns of Non-Coding RNAs

Athanasius F. Bompfünewerer [c,d], Christoph Flamm [c], Claudia Fried [a], Guido Fritzsch [b], Ivo L. Hofacker [c], Jörg Lehmann [a], Kristin Missal [a], Axel Mosig [a], Bettina Müller [e,a], Sonja J. Prohaska [a], Bärbel M. R. Stadler [f], Peter F. Stadler [a,b,c,g,*], Andrea Tanzer [b,c], Stefan Washietl [c], and Christina Witwer [c]

[a] *Bioinformatics Group, Department of Computer Science, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*

[b] *Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*

[c] *Department of Theoretical Chemistry University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*

[d] *Zentralfriedhof Wien, 3. Tor Simmeringer Haupstraße, A-1110 Wien, Austria*

[e] *Department of Biotechnology & Bioinformatics, University of Applied Sciences Weihenstephan, D-85350 Freising, Germany*

[f] *Max Planck Institute for Mathematics in the Sciences Inselstrasse 22-26, D-04103 Leipzig, Germany*

[g] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*

[*] *Corresponding author:*
Peter F. Stadler
Bioinformatics Group, Department of Computer Science, and
Interdisciplinary Center for Bioinformatics,
University of Leipzig,
Härtelstraße 7b, D-04107 Leipzig, Germany.
Tel: ++49 341 97 16691, Fax: ++49 341 97 16709,
Email: `peter.stadler@bioinf.uni-leipzig.de`

Abstract

A plethora of new functions of non-coding RNAs have been discovered in past few years. In fact, RNA is emerging as *the* central player in cellular regulation, taking on active roles in multiple regulatory layers from transcription, RNA maturation, and

RNA modification to translational regulation. Nevertheless, very little is known about the evolution of this "Modern RNA World" and its components. In this contribution we attempt to provide at least a cursory overview of the diversity of non-coding RNAs and functional RNA motifs in non-translated regions of regular messenger RNAs (mRNAs) with an emphasis on evolutionary questions. This survey is complemented by an in-depth analysis of examples from different classes of RNAs focusing mostly on their evolution in the vertebrate lineage. We present a survey of Y RNA genes in vertebrates, studies of the molecular evolution of the U7 snRNA, the snoRNAs E1/U17, E2, and E3, the Y RNA family, the *let-7* microRNA family, and the mRNA-like *evf-1* gene. We furthermore discuss the statistical distribution of microRNAs in metazoans, which suggests an explosive increase in the microRNA repertoire in vertebrates. The analysis of the transcription of non-coding RNAs (ncRNAs) suggests that small RNAs in general are genetically mobile in the sense that their association with a hostgene (e.g. when transcribed from introns of a mRNA) can change on evolutionary time scales. The *let-7* family demonstrates, that even the mode of transcription (as intron or as exon) can change among paralogous ncRNA.

Key words: evolution, non-coding RNA, mRNA, rRNA, snRNA, snoRNA, miRNA, Y-RNA, vault RNA, gRNA, RNA editing, UTR.

1 Introduction

Although it is still commonplace to speak of *"genes and their encoded protein products"*, thousands of human genes produce transcripts that exert their function without ever producing proteins. The diversity of sequences, sizes, structures, and functions of the known non-coding RNAs (ncRNAs) strongly suggests that we have seen only a small fraction of the functional RNAs. Most of the ncRNAs are small, they do not have translated ORFs, and they are not polyadenylated. Unlike protein coding genes, ncRNA gene sequences do not exhibit a strong *common* statistical signal, hence a reliable general purpose computational genefinder for non-coding RNA genes has been elusive [88].

The list of functional non-coding RNAs includes key players in the biochemistry of the cell. Many of them have characteristic secondary structures that are highly conserved in evolution. A non-exhaustive list is compiled in Tab. 1. In addition to these relatively well-described classes there is a diverse and rapidly growing list of ncRNAs with sometimes enigmatic function: The 17kb *Xist* RNA of humans and the smaller *roX* RNAs of *Drosophila* play a key role in dosage compensation and X chromosome inactivation [13, 109]. Several large ncRNAs are expressed from imprinted regions, see also [368]. Many of these are cis-antisense RNAs that overlap coding genes on the other genomic strand [96]. An RNA (meiRNA) regulates the onset of meiosis in fission yeast [306]. No precise function is known at present for the human H19 transcript, or the hrsω transcript induced by heat shock in *Drosophila*, see e.g. [97]. A recent survey of the slime mold *Dictyostelium discoideum* uncovered two novel classes of ncRNAs [12]. An experimental screen recovered hundreds of small ncRNAs from the mouse [175]. Ambros and coworkers [7] reported more than 30 *tiny non-coding RNAs* in a recent survey of *Caenorhabditis elegans* that are slightly shorter than microRNAs, are not processed from hairpin precursors, and are poorly conserved between related species.

Since the discovery of microRNAs [219, 228, 237] and the development of RNAi as a general technique for manipulating translation [93], there is mounting evidence that ncRNAs in fact dominate the regulatory networks of the cell [21, 157, 273, 274, 391]: The *E. coli* genome encodes more than 50 small RNA genes at least some of which (e.g. *MicF*, *OxyS*, *DsrA*, *Spot42*, *RhyB*) act by base-pairing to activate or repress translation [127, 383]. A large fraction of the mouse transcriptome consists of non-coding RNAs, many of them anti-sense to known protein-coding transcripts [389]. Similarly, about half of the transcripts from Human chromosomes 21 and 22 are non-coding [52, 194], see [290] for a discussion of the possible roles of anti-sense RNAs. *Leishmania* and related kinetoplastids have reduced transcriptional regulation of gene expression to a minimum, maybe to the point of having lost any specific polymerase II transcription initiation [62]. Instead, these organisms use an elaborate cleavage and trans-splicing mechanism based on the action of $\sim$ 40nt "spliced leader" RNA. *Tetrahymena* appears to use an RNA-based mechanism for directing its genome-wide DNA rearrangements [286, 444].

Another level of RNA function is presented by functional motifs within protein-coding RNAs. We briefly mention a few of the best-understood examples of structurally conserved RNA motifs in viral RNAs: An *IRES* (internal ribosomal entry site) region is used instead of a cap to initialize translation by Picornaviridae, some Flaviviridae including Hepatitis C virus, and a small number of mRNAs, see e.g. [352, 174, 326]. Viral RNAs contain a large number of structured binding motifs that are essential for the viral life cycles, e.g. the TAR and RRE motifs in HIV [75] or the CRE (cis-acting replication element) hairpin in Picornaviridae [439]. RNA-localization mechanisms involve specific sequences motifs in the localized RNA that cause certain proteins to mediate the interaction with cytoskeletal elements [307]. The localized *bicoid* mRNA, for instance, is responsible for laying down the body axes of the embryo [332].

RNA switches, i.e., RNAs that drastically change their structure, are important regulatory elements [386]. For instance, the terminator and anti-terminator, two alternative RNA hairpins, regulate gene expression in *E. coli* and *B. subtilis* by attenuation [15, 104, 337]. RNA switches can provide exact temporal control as in the *hok/sok* system of plasmid R1 which triggers programmed cell death [297, 287]. RNA switches also play a role in the spliced leader of trypanosomes and nematodes [234]. A theoretical study shows that RNAs exhibiting very different secondary structures with near-groundstate energy, i.e., potential riboswitches, are relatively frequent and easily accessible in

Table 1. Major classes of functional RNAs

Class		Size	Function	Phylogenetic Distribution	DB
tRNA		70-80	translation	ubiquitous	[379]
rRNA	16S/18S	1.5k	translation	ubiquitous	[416, 266]
	28S+5.8S/23S	3k	translation	ubiquitous	[443, 266]
	5S	130	translation	ubiquitous	[390]
RNase P	P	220-440	tRNA maturation	ubiquitous	[40]
	MRP	250-350	endonuclease, 5.8S rRNA maturation	eukarya	
snoRNA	H/ACA	∼130	pseudouridinylation in rRNAs	eukarya	[355]
	C/D	60-80	ribose 2'-O-methylation in rRNAs	eukarya, archaea	
	telomerase	400-550		eukarya	
snRNA	U1,U2,U4,U5,U6	100-160	major spliceosome, mRNA maturation	eukarya	
	U11,U12	130-140	minor spliceosome, mRNA maturation	eukarya	[131]
	SL	∼100	trans-splicing	lower eukaryotes	
	U7	∼65	histone mRNA maturation	eukarya	
	7SK	∼300	transcriptional regulation	vertebrata	
	7SL/SRP	300-400	signal recognition particle	ubiquitous	[124]
	vault	80-100	part of vault particle	vertebrata	
	Y	80-100	part of Ro particle	metazoa	
tmRNA		300-400	tags protein for proteolysis	bacteria, chloroplasts, cyanoplasts	[457]
miRNA		∼22	post-transcriptional regulation	multicellular organisms	[131]
gRNA		40-80	RNA editing	kinetoplastids	[156]

evolution [108]. Artificial riboswitches have been explored for biotechnological applications [378, 207, 364] and it has been demonstrated that such constructs can be specifically triggered by means of small "modifier" RNAs [277, 142].

Given the importance of ncRNAs and RNA-based mechanism in extant life-forms, it is surprising that we know relatively little about the evolutionary history of most RNA classes. There are strong reasons to conclude that the Last Common Ancestor (LCA) was preceded by simpler life forms that were based primarily on RNA. In this *RNA World* scenario [117, 116], the translation of RNA into proteins and, finally, the usage of DNA [110] as information storage device are later innovations. The wide range of catalytic activities that can be realized by relatively small ribozymes [22, 177, 188, 191, 236, 411] as well as the usage of RNA catalysis at crucial points of the information metabolism of modern cells [186, 86, 289] provides support for the RNA World hypothesis. Plausible ribozyme catalyzed pathways for a late-stage ribo-organism [191], the role and evolution of co-enzymes [180], and a rather detailed model of the steps leading from the RNA world to modern cellular architectures [333] have been the subject of detailed investigations.

Probably the best-studied group are the ribosomal RNAs (rRNAs) because of their utility in molecular phylogenetics. In fact, much of our knowledge about the deepest branches of the tree of life has been inferred from 16S/18S sequence data [85, 308, 415, 327, 51]. Besides the 16S/18S and the 28S/23S large subunit rRNA, other classes of RNAs, however, have been used only sporadically for these purposes, although it has been shown that they are phylogenetically informative [42, 66, 173]. Telomerase RNA structures were used to elucidate the phylogeny of tetrahymenine ciliates [445]. Nevertheless, relatively little information is available on the origins of various RNA classes. Apart from the ribosomal RNAs (see e.g. [43]) and tRNAs [91], an origin predating the last common ancestor is clear only for the RNase P/RNase MRP family.

The `Rfam` database [131, 132], the `noncode` database [253], and the RNAdb [312] collect the flood of information on such ncRNAs and functional RNA motifs that before has been distributed over a large number of specialized databases (referenced in Tab. 1) dedicated to individual ncRNA families. A specialized database for plant-specific ncRNAs is the Arabidopsis Small RNA Project Database (ASRP) [140].

The purpose of this contribution is two-fold. Firstly, we tried to compile an overview of the current (January 2005) knowledge on all the different levels of RNA activity in the cell, with an emphasis of what is (or is not) known about the evolution of individual classes of RNAs. Secondly, we use the framework of the review-like material to put new results on individual ncRNAs into perspective. Together, a picture emerges that on the one hand supports the picture of RNA as an ancient player in the cell, likely deriving from an RNA world pre-dating the last common ancestor of all extant life [186], while on the other hand many ncRNA families are probably relatively young innovations or have expanded dramatically, as for instance microRNAs, in certain lineages.

2 Detection of ncRNAs

Genome databases nowadays offer a wealth of annotation about protein coding genes and their putative functions. Annotation of ncRNA genes, however, is almost non-existent. The main reason for this is the lack of established and reliable methods to detect such ncRNA genes computationally in genomic sequences. Current approaches for ncRNA detection can be clearly separated into two classes: methods to detect new members of already known and well-characterized ncRNA families, and attempts to predict RNA genes *de novo* so that novel families of ncRNAs can also be found.

2.1 Members of Known Families

Large, highly conserved ncRNAs, in particular ribosomal RNAs, can easily be found using `blast` [4]. Similarly, `blast` can be used to find orthologous ncRNAs in closely related species, e.g. [395, 430]. In most cases, however, this approach is limited by the relatively fast evolution of most ncRNAs. Since RNA sequence often evolves much faster than structure, the sensitivity of search tools can be greatly improved by using both sequence and secondary structure information.

The simplest class of search tools uses regular or context free grammars to describe RNA motifs that are explicitly known to the user. There is no possibility to adapt the model to variations of the instance, and it is also very difficult for a user to define production rules for complicated motifs with a large number of exceptions.

With probabilistic models, such as stochastic context free grammars (SCFG), the user is able to assign probability distributions to production rules; noise in the dataset is handled easily because the model can adapt itself to variations. The main drawback of stochastic context free grammars is that most of the available implementations demand large computational resources. Hybrid languages, like `HyPaL` [129] or the language used in `RNAMotif` [264], connect pattern languages with user defined approximative rules, which rank the results according to their distance to the motif. Their advantage lies in a faster processing compared to SCFG. Nevertheless, the definition of approximative rules also requires explicit knowledge, at least to some extent. Table 2 sum-

Table 2
General Purpose Algorithms for RNA Motif Detection.
Tools that detect a special class of RNA motifs are not listed here.

Program	Comparative or single organism	Description
Approaches which search for instances of a motif		
`ERPIN` [114]	comparative	Input is a sequence alignment with consensus structure. For each helix and single strand a log-odds-score profile is defined which describes the motif.
`PATSearch` [324]	single	Motif is defined by a language inspired by regular expressions.
`fragrep` [293]	single	Detects patterns consisting of approximately matched gapless blocks with constrained inter-block distances
`Palingol` [29]	single	A constraint programming language particularly adapted for secondary structures. Allows both sequence and structure patterns, including pseudo-knots.
`RNAMotif` [264]	comparative	Description of structural motif in terms of helices and sequence patterns. Putative hits are ranked according to user defined rules.
`infernal` [89]	comparative	Toolkit for constructing covariance models and finding new members of a family. Input is a multiple alignment with structural annotation. With SCFGs a consensus model of RNA structure shared by these sequences is defined.
`Rsearch` [202]	single	Input is a single RNA sequence and its structural information. `Rsearch` is a local alignment algorithm which considers structural and sequence constraints. A base pair and single nucleotide substitution matrix for RNAs (RIBOSUM) defines alignment scores.
`FastR` [17]	single	Like `Rsearch` a pairwise alignment algorithm that addresses structural and sequence conservation. Running time is highly decreased by preprocessing the target sequences. Only those targets sharing similar structural features with the query RNA are aligned.
Approaches which search for motifs from scratch		
`SLASH` [123]	comparative	Input are unaligned sequences. `foldalign` defines highest scoring local alignments of these sequences according to sequence and structure constraints. `COVE` creates a SCFG model from those local alignments and does database searches.
`RNAProfile` [318]	comparative	Input is a set of unaligned sequences. Motif is defined by the number of single hairpins it may contain. Greedy heuristic to find sequences in the input set which share a common motif with defined number of hairpins.
GPRM [170]	comparative	Genetic programming approach to find structural RNA motifs that discriminate a set of input sequences from a set of randomized sequences
`HyPa` and `HyPaLib` [129]	single	A search engine and pattern library for “hybrid patterns”, consisting of sequence and structure elements. The language also includes thermodynamic constraints. Currently, however, `HyPaLib` contains only some 60 patterns.

marizes the most commonly used approaches.

`PatSearch` [324], `RNAMotif` [264], and `Palingol` [29] are tools which allow the user to specify a given motif with a particular description language and offer search approaches to identify instances of the motif in a set of sequences.

`Palingol` is a constraint programming language to describe arbitrary rules on primary and secondary structure. The user defines a series of boolean expressions which must be satisfied by a successful hit.

In `PatSearch`, a language similar to regular expressions is used to describe motifs. For patterns composed of a string, a weight matrix can be defined which enables ranking and searching for approximative hits.

`RNAMotif` combines a pattern language with an `awk`-like programming language that describes approximative user defined scores. Sequences which have been matched successfully are evaluated and ranked according to the scoring section.

`ERPIN` [114] is an example of tools that do not need an explicit definition of a descriptor to search for homologs of a motif. From a sequence alignment annotated with helix regions it extracts frequencies of nucleotides in single strands and base pair frequencies in helices. Those frequencies are compared to expected base frequencies in the target database by calculating log-odds ratios. The sum of log-odds ratios over all positions of a target sequence gives the final score.

`RNAProfile` [318] requires as input the number of hairpins of a motif to extract it from an unaligned set of sequences where some contain the same motif. All sequences are folded and only those subsequences forming minimum free energy structures with the specified number of hairpins are regarded during the search. In a greedy search approach the selected regions of the first two sequences are pairwise aligned, according to primary and secondary structure. For each alignment, a profile, composed of observed frequencies of unpaired and paired nucleotides at each position, is defined and the best scoring ones are kept. In the second step the best scoring pairwise profiles are aligned to the selected regions of the next sequence and again only the best updated profiles are kept, and so on. If all sequences of the input set are processed, the highest scoring profiles define the detected motif. A fitness value is assigned to each final hit assessing its statistical significance.

A number of large-scale surveys have been performed using one of the general purpose tools mentioned above. An non-exhaustive list includes a microRNA survey using `ERPIN` [241], a search for U5 snRNA and RNase P using `RNAmotif` [65], and a survey of RNase P RNAs in bacterial genomes [244].

`Fragrep` [293] is a simple sequence based tool that allows to specify a query of short sequence elements that are separated by poorly conserved regions of variable length. Local alignment algorithms such as `blast` are therefore ill-suited for the discovery of new homologs of such ncRNAs in genomic sequences. The `fragrep` tool instead implements an efficient algorithm for detecting pattern fragments that occur in a given order. For each pattern fragment a mismatch

Table 3
Survey of Y RNAs in completely sequenced genomes using `fragrep`.

Genome	Hs	Mm	Rn	Gg	Xl	Tr	Tn	Dr
# matches	148	6	8	4	4	3	3	2

Hs: *Homo sapiens*, Mm: *Mus musculus*, Rn: *Rattus norvegicus*, Gg: *Gallus gallus*, Xt: *Xenopus tropicalis*; Tr: *Takifugu rubripes*, Tn: *Tetraodon nigroviridis*, Dr: *Danio rerio.*

tolerance and bounds on the length of the intervening sequences can be specified separately.

The application of `fragrep` is demonstrated in Tab. 3 using Y RNAs, an abundant small ncRNA described in some more detail below, as an example. It is straightforward to extract a query from sequences and structures of Y1, Y3, Y4 and Y5 RNAs given in [304]; the conserved sequence fragments of Y RNAs have also been studied by other authors [103, 399]. The large number of human sequences indicates that Y RNAs are associated with a repeat family in the human genome. An analysis of the Y RNA candidate sequences will be given in section 3.5.

Specialized programs have been developed to detect members of particular ncRNA families. Examples of this approach include `miRseeker` for microRNAs [221], `BRUCE` for tmRNAs [227], `tRNAscan` for tRNAs [254], `snoScan` for box C/D snoRNAs [255], `fisher` for box H/ACA snoRNAs [90], as well as a heuristic for SRP RNAs [339, 351]. An improved method for box C/D snoRNAs was recently presented by Accardo *et. al.* [1]: starting from yeast rRNA methylation sites, they first identified homologous positions in *D. melanogaster* rRNAs and then use `snoScan` [255] to search for putative snoRNAs with binding motifs complementary to the putative methylation sites. MicroRNAs in plants can be found by extracting those hairpin structures that contain sequence motifs complementary to a mRNA, which is then a putative target [189, 34, 2].

2.2 Novel ncRNAs and RNA motifs

Detecting novel ncRNAs without any prior knowledge of sequence or structure is still a largely unsolved issue. In contrast to protein-coding genes, which show strong statistical signals like open reading frames or codon bias, ncRNAs lack any comparable signals in primary sequence that could be used for reliable detection.

Only in very special cases can ncRNAs be identified based on a significant bias in base composition. AT-rich hyper-thermophiles were successfully screened for ncRNAs simply by searching for GC rich regions [203, 360]. MicroRNAs can be detected based on their increased thermodynamic stability [35]. Carter *et al.* used machine learning techniques to extract common sequence features of known ncRNAs including GC content in *E. coli* [49].

Most ncRNAs do, however, depend on a well-defined structure for their function. This has led to various attempts to predict functional RNAs using predicted secondary structures. It was first suggested by Maizel and co-workers that functional RNA elements should have a more stable secondary structure than expected by chance [231, 57]. However, Rivas and Eddy had to conclude in an in-depth study on the subject that thermodynamic stability alone is generally not statistically significant enough for reliable ncRNA detection [342]. Some other characteristic measures derived from secondary structure predictions have been proposed [365, 233, 232] which, however, are also of limited value in the context of genome wide ncRNA prediction. A combination of gene expression data and high level sequence conservation was successful in discovering novel ncRNAs in the intergenic regions of the *E. coli* genome [429].

The reason for the limited success of these approaches is that the presence of secondary structure in itself does not indicate any functional significance, because almost all RNA molecules form secondary structures. In fact, most compelling evidence for functional significance comes from comparative studies that demonstrate evolutionary conservation of structure.

Extensive computer simulations, see e.g. [367, 135, 136, 176], showed that a small number of point mutations is very likely to cause large changes in the secondary structures. It follows that structural features will be preserved in RNA molecules with less than some 80% of sequence identity only if these features are under stabilizing selection, i.e., when they are functional.

This fact is exploited by the `alidot` [161] algorithm for searching conserved secondary structure patterns in large RNAs. Secondary structures are predicted independently for each sequence, typically using McCaskill's algorithm [275], which yields a list of thermodynamically plausible base pairs with their equilibrium probabilities. Next, a conventional multiple sequence alignment is computed, e.g. using `ClustalW`. By copying the gaps from the multiple sequence alignment into the predicted structures, a list of homologous base pairs is obtained. This list is then sorted by means of hierarchical credibility criteria that explicitly take into account both thermodynamic information and sequence covariation. A detailed description of the method can be found in [161, 164]. A similar approach is taken by the `ConStruct` tool [259, 258], which also features a graphical tool for manipulating the sequence alignment in order to achieve a better consensus structure. `Alidot` does not pre-suppose the existence of a global conserved structure. It is therefore particularly well suited when the sequences are expected to contain only small structurally conserved regions, as is the case for example in RNA viruses.

For predicting globally conserved structures a different technique, "folding the alignment", may be preferred. Here, the folding algorithm itself is modified to work on a sequence profile, or multiple sequence alignment, instead of a single sequence. The two best known implementations of this approach are `pfold` [205, 204], and `RNAalifold` [162]. `pfold` is based on an stochastic context free grammar, and thus uses parameters derived from a training set. It also makes explicit use of a predicted phylogenetic tree. `RNAalifold`, on the other hand, uses the standard energy model for RNA secondary structures, augmented with a covariation term that rewards consistent and compensatory mutations. Thus, for identical sequences, it gives the same result as the single sequence prediction from `RNAfold`. With a few (or even just two) related sequences these programs achieve prediction accuracies much higher than prediction methods for single sequences. The approach is limited by the accuracy of the input alignment.

For sequences with less than 60% identity, pure sequence alignments typically differ significantly from structurally correct alignments. In these cases, one can resort to using a variant of the Sankoff algorithm [357] which computes the alignment and consensus structure simultaneously. Notable implementations are `foldalign` [123, 125, 147], `dynalign` [271], `pmcomp / pmmulti` [160], and `dart` [167]. The Sankoff algorithm is computationally very expensive, scaling as $\mathcal{O}(n^6)$ in the unrestricted case. The above algorithms therefore use various restrictions to improve speed (`foldalign` for example considers only unbranched stem-loop structures). Nevertheless, they are generally not suitable for genome wide scans. A different approach to structural alignments is provided by making use of the tree representations of RNA secondary structures. Both `RNAforrester` [158] and `MARNA` [374] produce multiple alignments from pairwise structure-based alignments. For a recent comparison of techniques for consensus structure prediction see [112].

Accurate predictions of consensus structures can provide a stepping stone towards reliable detection of functional RNAs. However, an successful ncRNA finder must also provide a measure of significance, such as an p-value or E-value. A well-known program to classify pairwise sequence alignments as ncRNA, protein coding, or anything else, is `qrna` [343]. This progamm compares the score of three distinct models of sequence evolution to decide which one describes best the given alignment: a pair stochastic context free grammars (SCFG) is used to model the evolution of secondary structure, a pair hidden Markov model (HMM) describes the evolution of protein coding sequence, and a different pair HMM implements the null model of a non-coding sequence. `Qrna` was successfully used to predict ncRNAs candidates in *E. coli* and *S. cerevisiae* [344, 276], some of which could be verified experimentally. `Qrna` is, however, currently limited to pairwise alignments, and somewhat slow for large genomic scans. Other recent programs for detecting conserved RNA secondary structures include `ddbRNA` [80] and `MSARi` [67]. A phylogenetic shadowing approach specifically geared towards the detection of microRNA precursors is described in [27].

Currently, the sensitivity and/or specificity of all these programs is insufficient for screens of large eukaryotic genomes. Part of the problem is often oversimplification of the folding model (poor thermodynamics), as well as considering only compensatory mutations as signal for structural conservation. Typical data sets, however, do not always show enough sequence variation ensuring this to be a significant indicator.

Recently, it has been demonstrated that the comparative approach can give significant results even for alignments with only few sequence and high similarity [426]. This approach uses `RNAalifold` [162] to compute consensus structures, making best use of covariance information and thermodynamic stability. Significance is then measured by a z-score comparing the consensus folding energy of the native alignment (as computed by `RNAalifold`) with the folding energies of randomized alignments, obtained by a shuffling procedure. Although the results are promising in terms of accuracy, the practicability of this approach is limited by the high computational costs caused by the time consuming shuffling procedure. In a more recent contribution, this problem is solved resulting in a time efficient algorithm showing similar accuracy. The program `RNAz` [427] uses two independent criteria for classification: a z-score measuring thermodynamic stability of individual sequences, and a *structure conservation index* obtained by comparing folding energies of the individual sequences with the predicted consensus folding. The two criteria are combined by a support vector machine that detects conserved and stable RNA secondary structures with high sensitivity and specificity. Thus, `RNAz` seems to be the first program suitable for screening large eukaryotic genomes [427, 82].

`GPRM` [170] considers motif prediction a supervised learning problem. Coregulated mRNA sequences are used as positive examples, while the same number of randomly generated sequences form a set of negative examples. A genetic programming approach is used to learn the motifs in the predicted structures that can discriminate the positive set from the random sequences. Optimal discriminators are therefore good candidates for functionally important structural motifs [171].

It should be pointed out, however, that not all ncRNAs can be tracked down by searching for conserved secondary structures. To mention only a few examples, the U4 and U6 spliceosomal RNAs are known to form extensive *inter*-molecular interactions rather than forming stable *intra*-molecular secondary structures and are therefore missed by this approach. Also, most of the C/D-class snoRNAs lack an easily detectable secondary structure. Thus, while reliable structural RNA gene finding programs have come into reach, a general RNA gene finder remains elusive.

3 Sequence Evolution of ncRNA Families

3.1 Non-coding RNAs and Phylogenetic Inference

While, as we have seen in the previous section, sequence information alone is in general insufficient to detect non-coding RNAs, it can be used very well to elucidate the evolutionary relationships of these genes, at least within a given family of ncRNAs or RNA motifs. Since most known ncRNAs have evolutionarily conserved structures, however, they are only approximately described by models assuming independent evolution of sequence positions. A more accurate treatment explicitly takes into account that sequence positions that form conserved base pairs are highly correlated. Corresponding models of sequence evolution are described e.g. in [362, 205, 358, 311]. The `phase` package [190, 173] implements such a model and is specifically designed to infer phylogenies from RNAs that have a conserved secondary structure.

These secondary structures, however, have rarely been used in molecular phylogenetics so far. An exception is the investigation into the history of RNase P and RNase MRP RNAs by David Penny and co-workers [66]. This study uses RNA editing distances [371] implemented in `Vienna RNA package` [163] to show that "RNA secondary structure is useful for evaluating evolutionary relatedness, even with sequences that cannot be aligned with confidence". More recently, cladistic analyses based on RNA secondary structure [42, 43] have demonstrated this point convincingly, in particular at the level of deep phylogenies.

In the following we compile an overview of our current knowledge of the evolution of the best known classes of non-coding RNAs. Our focus therefore are gene phylogenies and the history of duplications and losses that led to the present ncRNA inventory. This review of the literature is complemented by a number of original results, for which we provide supplemental data in electronic form[1]. We have mostly used neighbor-joining [354] rather than the sophisticated maximum likelihood techniques mentioned above, since we are interested here in the large-scale patterns rather than subtle details of the ncRNA gene phylogenies.

3.2 tRNAs

Multiple copies of functional tRNA genes, the existence of numerous pseudogenes and tRNA-derived repeats are general characteristics of tRNA evolution

[1] `http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/05-001/`

[111]. Comparative sequence analysis of transfer RNA by means of statistical geometry provides strong evidence that transfer RNA sequences diverged long before the divergence of archaea and eubacteria [91]. In Fig. 1 we illustrate this using tRNAs coding for six of the twenty amino acids: tRNAs with the same anticodon form coherent subtrees. Models for the origin of tRNA from even simpler components are discussed e.g. in [92, 345, 81].

The evolution of mitochondrial tRNA was studied in detail by Paul Higgs and collaborators [173, 154, 183]. In particular, they present evidence that the two animal tRNA-Leu variants (one with anticodon UAG, the other with anticodon UAA) evolve by a peculiar mechanism of gene duplication, followed by mutation of the anticodon and subsequent gene loss. At least five such replacement events have been described in metazoan evolution [183].

3.3 Ribosomal RNAs

Evidence from both *in vitro* studies [199, 302] and the analysis of the atomic structure [338] reveals that the ribosome is in fact a ribozyme in which only rRNA is involved in the positioning of the A- site and P-site substrates, and only RNA is in a position to chemically facilitate peptide-bond formation [382]. Due to its ubiquity, size, and generally slow rate of evolution, the small-subunit ribosomal RNA has become the most sequenced of all genes and an invaluable tool for molecular phylogenetics [85, 308, 415, 327]. More recently, large subunit rRNA are increasingly used for this purpose as well, e.g. [268]. The evolution of the secondary structures of ribosomal RNAs with an emphasis on functional sites is discussed in detail in [43].

Most organisms have multiple copies of their rRNA genes. In *Escherichia coli*, for instance, there are seven operons encoding rRNAs 16S, 23S, and 5S [31]. Typical Eukaryotes contain tandemly repeated arrays of rRNAs genes each of which contains three of the four ribosomal RNA components separated by two “internally transcribed spacers” (18S/ITS1/5.8S/ITS2/28S) [155]. In most species the fourth rRNA gene, 5S rRNA, is also contained in this array, while it sometimes is dispersed throughout the genome (as in *Schizosaccharomyces pombe*, [441]), organized in its own tandem arrays (as in soybeans [128]), or both (as in humans [252]). For each of these genes, however, the rDNA sequences that are represented in fully processed rRNA are essentially identical in most organisms, i.e., rRNA genes are subject to *concerted evolution* [155, 361, 121]. This is the tendency of the different genes in a gene family or gene cluster to evolve “in concert”. As a consequence, one observes that paralogous sequence in the same species are more similar than orthologous sequences of different species. Multiple molecular mechanisms may account for this phenomenon: gene conversion (a non-reciprocal process in which two

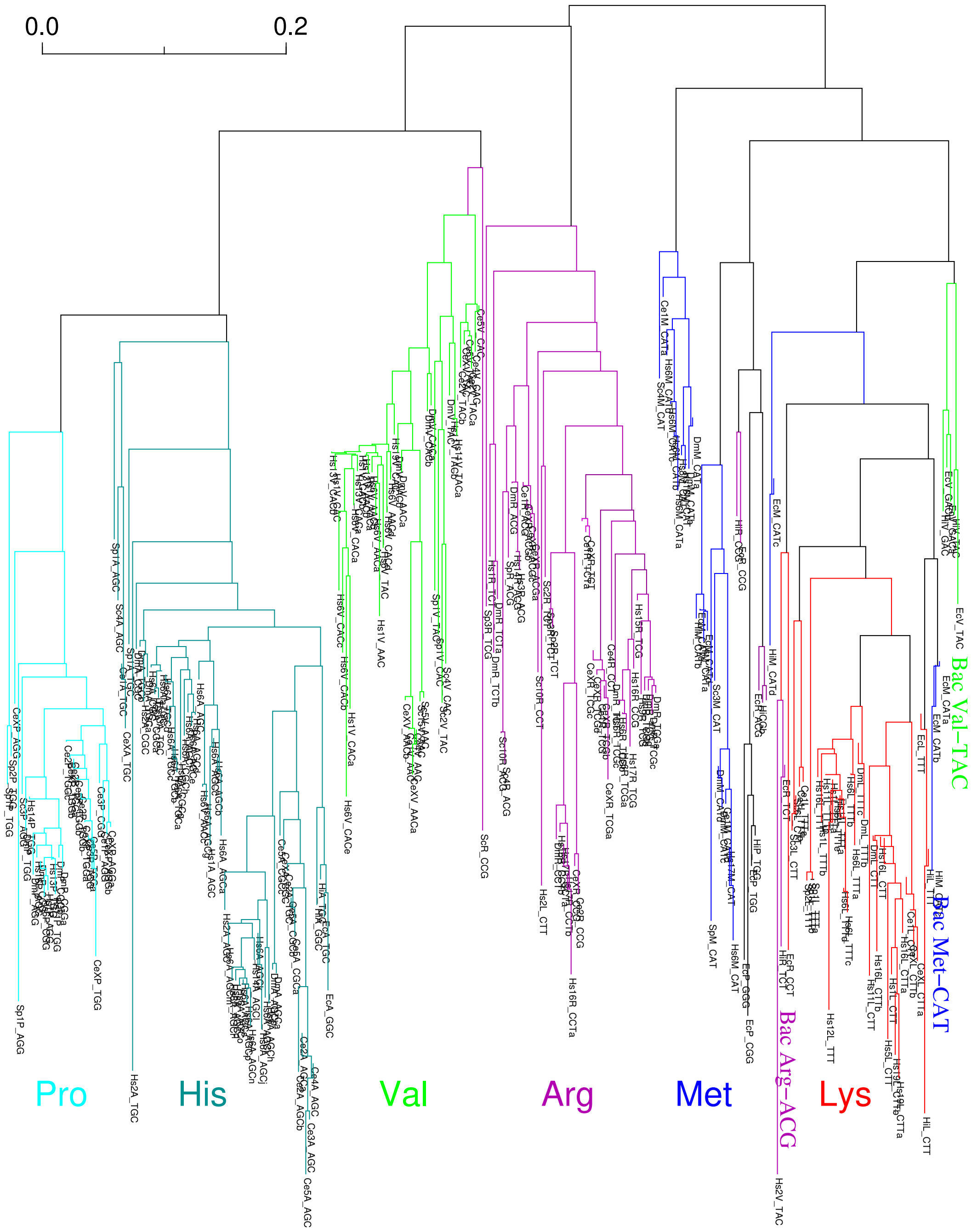


Fig. 1. Neighbor-joining tree [354] of nuclear tRNAs with anticodons for Alanine (Ala), Valine (Val), Prolin (Pro), Arginine (Arg), Leucine (Leu), and Methionine (Met) from Human, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces pombe*, *Saccharomyces cerevisiae*, *Escherichia coli* K-12, and *Haemophilus influenzae*. A few groups of bacterial tRNAs that fall outside the main groups are indicated. Sequence data are taken from the *Genomic tRNA Database*, `http://rna.wustl.edu/tRNAdb/`.

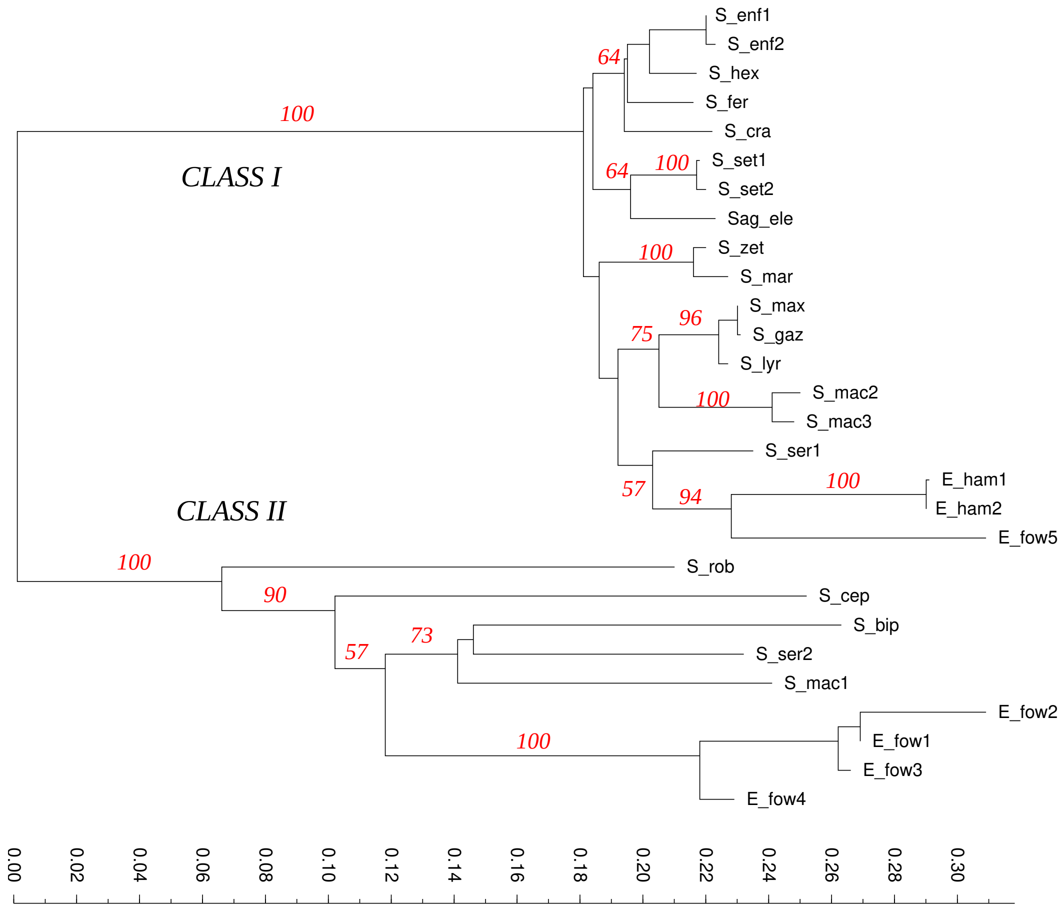


Fig. 2. Neighbor-joining phylogeny [354] of chaetognatha from partial 28S RNA sequences. The tree is recalculated from data published by M.J. Telford and P.W.H. Holland [397] using a `clustalw` alignment and the `phylip` package. The 28S sequences fall into two paralog groups that have separated at a common ancestor of the recent chaetognaths. For the species *Eukrohnia fowleri*, *Sagitta macrocephala* and *Sagitta serratodentata* both paralogs have been identified [397]. Bootstrap values in percent (1000 replicates) are marked at major branches.

sequences interact in such a way that one is converted by the other), repeated unequal crossover, and gene amplification (frequent duplications and losses within family), see [246] for a review.

There are, however, exceptions to the rule: two classes of ancient paralogs of the 28S rRNA have been reported in the chaetognaths [397], see also Fig. 2. Similarly, paralog 18S rRNA are known e.g. in the flatworm family *Dugesiidae* [48, 47] and in apicomplexans [348], intraspecific 5.8S RNA variations have been reported in the coral *Acropora* [269]. In Xenopus, a somatic and an oocyte class 5S RNA genes are differentially expressed in development due to changes in transcription factor and histone interactions with the two types of gene [440]. Distinct types of rRNA operons were also found in the *Bacillus cereus*

Table 4
Spliceosomal RNA components.

Mechanism	snRNAs				
	pol-II				pol-III
major spliceosome	U1	U2	U4	U5	U6
minor spliceosome	U11	U12	U4atac	U5	U6atac
transsplicing		U2	U4	U5	U6

group [45]. Divergent paralogs could, if undetected, misguide phylogenetic studies.

3.4 Spliceosomal RNAs

Most genes in higher eukaryotes contain introns that must be excised from the primary transcript to yield a mature mRNA. Intron removal and ligation of the exons occurs in a massive ribonucleoparticle (RNP), the *spliceosome*, see e.g. [301] and the references therein. Recently, there has been mounting evidence that main catalytic function in the spliceosome are indeed performed by its RNA components, i.e., that the spliceosome, like the ribosome, is essentially a ribozyme [413, 414, 407]. The spliceosomal RNA U1 has an additional function in the regulation of transcriptional initiation [216].

There are three distinct splicing mechanisms that are all dependent on a small set of RNA components of the spliceosome, Tab. 4: The major-spliceosome is the predominant mechanism e.g. in vertebrates, plants, and yeasts, which spliced introns with the "canonical" **GT**-**AG** boundaries. The minor-spliceosome processes introns with non-canonical boundaries [316], predominantly **AT**-**AC**. Trans-splicing, finally joins a small non-coding exon derived from the SL RNA to each coding exon of the pre-mRNA and is used to produce multiple mature mRNAs from a single poly-cistronic pre-mRNA [330, 404].

The evolutionary history of the spliceosome and its protein and RNA components is discussed in detail in [64]. In spliced leader *trans* splicing, a common 5'-terminal exon is added post-transcriptionally to mRNAs which is derived from the SL RNA. The evolutionary origin(s) of this mechanism are still poorly understood because there is no clear pattern in the phylogenetic distribution of species that have this mechanism and the SL RNAs of distant species are too different to decide whether they are indeed homologous [300].

Both the pol-II transcribed spliceosomal RNAs U1, U2, U4, and U5 and the pol-III transcribed U6 snRNA appear in multiple copies in many ver-

Table 5
Repetitive elements associated with U7 snRNA.
U7 RNA-like sequences are abundant in mammalian genomes, as determined by a `blast` search of the U7 sequence against the genomic sequence with a cutoff of $E = 10^{-10}$.

Species	Human	Mouse	Rat	Dog	Cow
# hits	21/91*	8	4	3	2

*21 hits when the U7 RNA sequence from [359] is used, 91 when using the consensus of all `Rfam` entries.

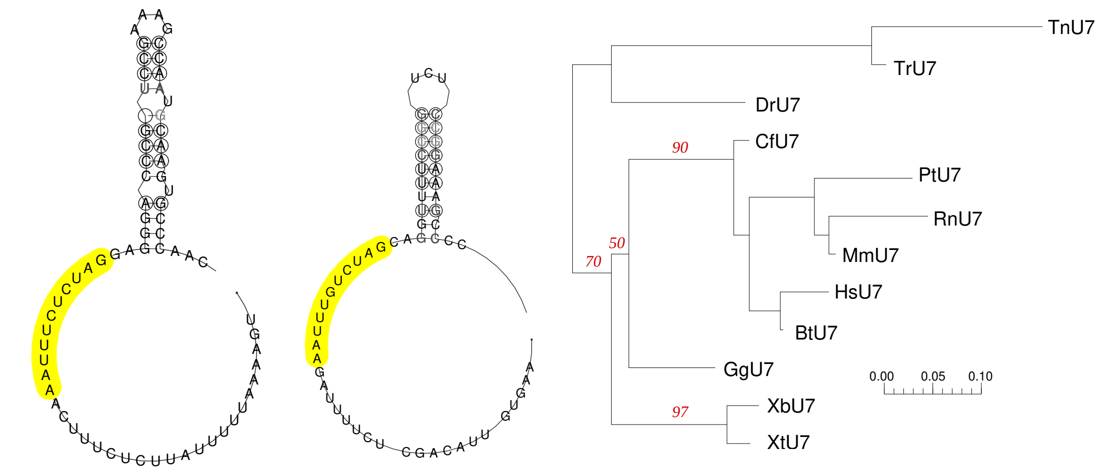
tebrates and are known to be subject to concerted evolution in some species [84, 247, 295, 431]. Divergent paralogs are also known in some species: For example, Xenopus has distinct embryonic and somatic classes of U1 snRNAs [70]. The evolution of U12 in vertebrates is considered in [396]. A comprehensive investigation of snRNA evolution in the light of the available genomic sequence data, however, is still missing.

3.5 Other snRNA-like Molecules

U7 RNA. Replication-dependent histone pre-mRNAs, in contrast to all other mRNAs, are not polyadenylated. Instead, they are processed at their 3'end by endonucleolytic cleavage between two conserved sequence elements located within about 100nt of the stop codon: a highly conserved stem loop structure and a purine-rich histone downstream element (HDE). The latter is recognized by the U7 small nuclear ribonucleoprotein which consists of the U7 snRNA, a common *Sm* protein, and two unique Sm-like protein, *Lsm10* and *Lsm11* [366].

The U7-snRNP-dependent histone RNA 3'end processing mechanism is a metazoan innovation [14]. Sequences of the U7 snRNA, which is only 60-70nt long, have been published for some mammals (e.g. [377, 451]), Xenopus [442], Fugu [295], an echinoderm [120], and more recently also for *Drosophila melanogaster* [83]. Using a simple `blast` search, we found additional homologs in the chick genome, in two additional teleosts and in *Drosophila pseudoobscura*. Like most other snRNAs, there are U7-derived repetitive sequences in some lineages, notably in human, while other species exhibit only a few scattered paralogs or pseudogenes [329], or even have only a single copy (e.g. in the fugu [295]), see Tab. 5.

The sequences evolve quickly, severely limiting the power of comparative approaches. Because of the short sequence length of only 60-70nt, one cannot



(a) (b) (c)

Fig. 3. Consensus secondary structures obtained from manual alignments of (a) 4 invertebrate and (b) 12 vertebrate U7 sequences calculated by `RNAalifold` [162]. The highly conserved Sm binding sequence is highlighted. Panel (c) shows a neighorjoining tree obtained from the vertebrate alignment using the `phylip` package. Resolution within the mammals is poor, otherwise the U7 RNA tree reflects the accepted species phylogeny. Species abbreviations are: Bt *Bos taurus*, Cf *Canis familiaris*, Dr *Danio rerio*, Hs *Homo sapiens*, Mm *Mus musculus*, Pt *Pan troglodytes*, Rn *Rattus norvegicus*, Tn *Tetraodon nigroviridis*, Tr *Takifugu rubripes*, Xb *Xenopus borealis*, Xt *Xenopus tropicalis*

expect a strong phylogenetic signal. Figure 3c shows, however, that the sequence evolution is at least consistent with established phylogeny.

The U7 snRNA forms a relatively well-conserved hairpin structure just downstream of the Sm binding sequence, see Fig. 3a,b. The U7 sequences were indeed used as an example to demonstrate the `ConStruct` approach to determining evolutionarily conserved secondary structures in [258]. The analysis in Fig. 3 using `RNAalifold` [162] shows that there are significant differences in the secondary structures of invertebrates and vertebrates: vertebrate have smaller stem-loop structure with smaller or no interior loops or bulges.

SRP RNA. The Signal Recognition Particle (SRP) is responsible for targeting nascent proteins to the ER membrane. In the process, protein synthesis is arrested when the SRP binds to the N-terminal signal of the nascent protein chain [196]. The SRP, components of which have been identified in all three domains of life [349], contains a non-coding RNA, which in higher metazoan is also known as 7SL RNA. While the secondary structure of archaeal SRP RNAs closely resembles those of higher eukaryotes, Fig. 4, protozoan and fungal sequences may deviate considerably, and only the S-domain is present in most bacterial sequences [455, 349, 351]. Chloroplast SRP RNA is described in [350]. A detailed comparative discussion of the structural features of SRP

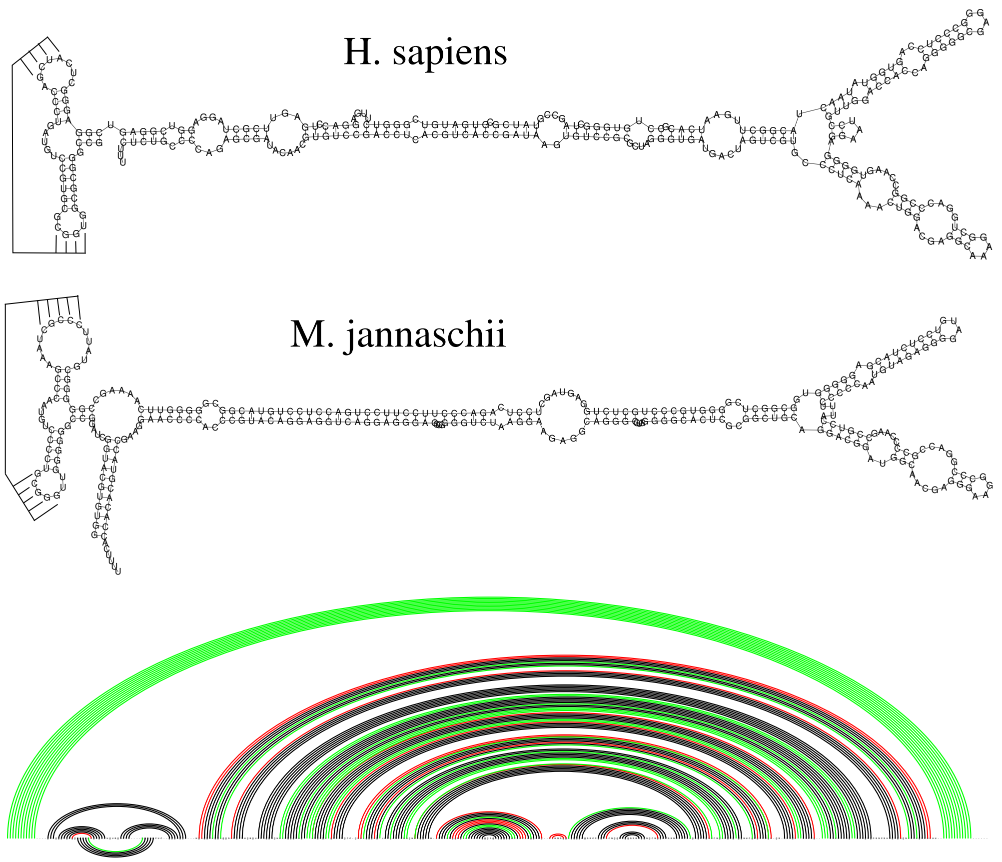


Fig. 4. Highly conserved secondary structure of SRP RNA from *Homo sapiens* and *Methanococus jannaschii* [455]. Bottom: superposition of both structures: base pairs contained in both species are drawn in black, base pairs only present in the *Homo sapiens* structure are drawn in red, and those only present in the *Methanococus jannaschii* are drawn in green.

RNAs from the different kingdoms can be found in [456].

Two small RNAs designated sRNA-85 (in *Leptomonas collosoma* [24]) and sRNA-76 (in *Trypanosoma brucei* [23]) co-isolate with the 7SL RNAs of these Trypanosomatids, and there are indications that they function in place of certain protein components of the signal recognition particle. Their evolutionary relationship with the 7SL RNAs, however, is unclear [456].

P and MRP RNA. The RNase P and RNase MRP RNAs are the catalytically active components of their respective RNPs, which both act as endonucleases. RNase P is essential for the maturation of tRNAs in Bacteria, Eukarya, and Archaea, see [331] for a summary of its phylogenetic distribution and structural evolution. MRP RNA, in contrast, has been found only in Eukarya where it cleaves the primers necessary for the initiation of mitochondrial DNA replication [292], but also has nuclear functions. RNase P and MRP appear to be ancient paralogs, albeit it remains unclear whether MRP RNA is a eukaryote innovation or an older invention [66]. In several ascomycete fungi the RNase MRP gene is located in the mitochondrial genome and vary considerably in size and sequence, see e.g. [392]. RNase P RNA is also encoded in the chloroplast genome of in some algae [76]. The absence of structural homology between bacterial and archaeal/eukaryotic RNase P proteins suggests that RNase P once was a pure ribozyme that pursued completely different strategies in the recruitment of protein subunits in the two different lineages [146]. A detailed investigation of bacterial RNase P RNAs [141] demonstrates an abrupt, dramatic restructuring in the common ancestor of the *Bacillus-Lactobacillus-Streptococcus* and the *Mycoplasma* groups of the low G+C Gram-positive bacteria. The latter shares the common ancestral "type A" structural architecture of bacterial RNase P RNAs, see also [214, 434].

Expressed paralogs of RNase P RNA have been found in the mouse [243], a systematic study of RNase P and MRP RNA variants, however, has not been performed to our knowledge.

7SK RNA. Despite its abundance in mammalian cells, the function of the 7SK RNP has remained unknown until recent studies implicated 7SK RNA as well as components of the splicing apparatus [216] in the regulation of transcriptional elongation, see [32, 280, 448]. Its secondary structure is known in detail from chemical probing experiments [428]. Interestingly, the 7SK RNA is very well conserved among vertebrates, while the lamprey sequence is already rather diverged [139]. D. Koper's PhD dissertation [209] reports divergent 7SK sequences from the hagfish *Myxine glutinosa* and from two invertebrate species: *Branchiostoma lanceolatum* and *Helix pomatia*.

Y RNAs are small eukaryotic RNAs that are part of the Ro ribonucleoprotein (Ro RNP) complex, whose function is not known at present. Four families of Y RNAs, Y1, Y3, Y4, and Y5, have been described in human and frog. Their secondary structure is very well conserved among vertebrates [304, 103, 399]. It consists of at least three stems, two of which form a stem-loop structure separated by a relatively short interior loop. The sequences in the stems, as well as parts of the loop regions, are highly conserved and probably serve as binding sites to the Ro60 protein in the Ro RNP complex and/or other cellular nucleic acids.

These conserved sequence patterns were used to scan genomic sequences for Y RNA candidates using the `fragrep` tool [293], see section 2.1. The phylogenetic tree resulting from an alignment of the matching sequences is shown in Figure 5. It allows a further classification of the Y RNA candidate matches. Several matches, classified as an outgroup in the tree, are likely to be random occurrences of the search pattern. Integration of known representatives of the

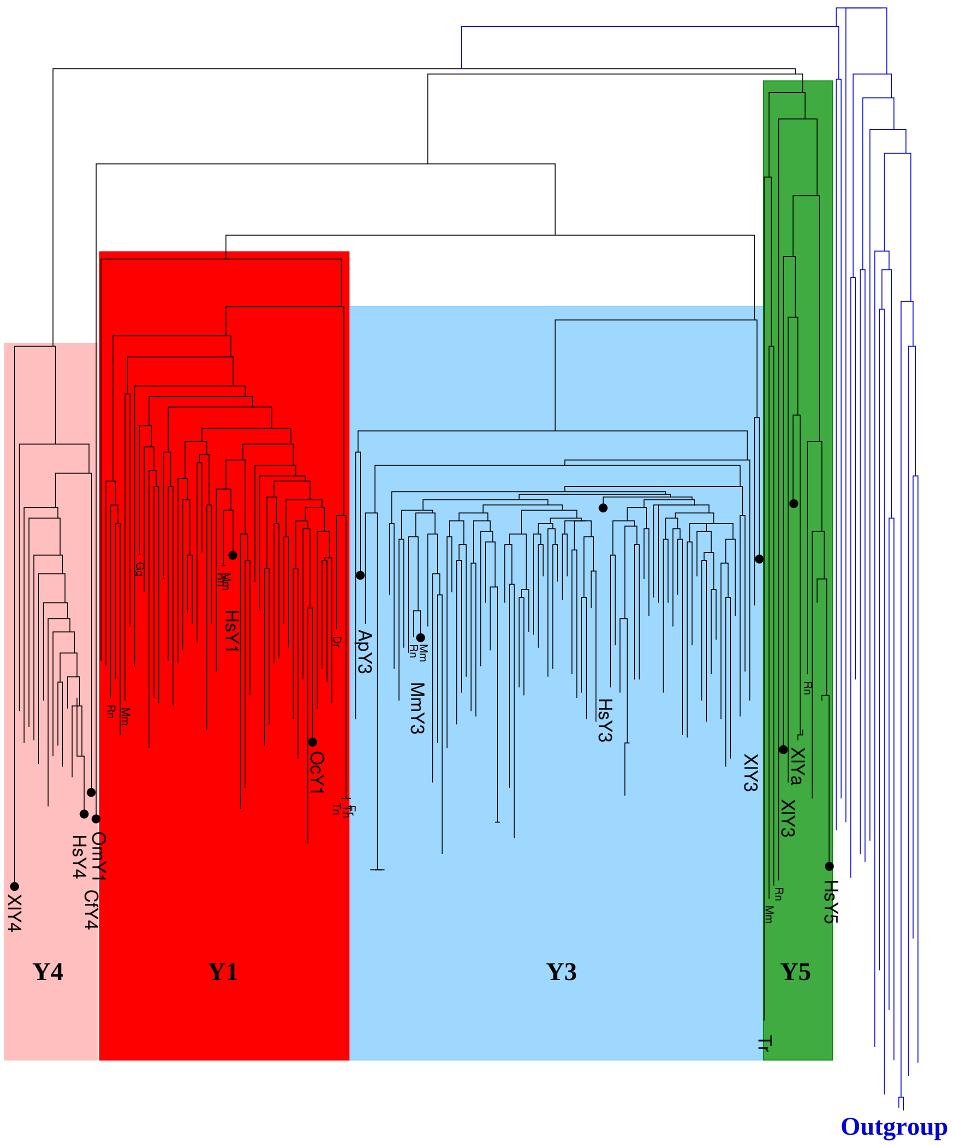


Fig. 5. Neighbor-joining tree derived from the candidate Y RNA matches obtained by `fragrep` using a `Clustalw` alignment [400]. Known Y1, Y3, Y4, and Y5 sequences were added to the candidate match sequences and are highlighted in the tree. Beside the outgroup on the left hand side, all matching sequences can be clearly assigned to one of the known groups of Y RNA.

known classes of Y RNA (Y1, Y3, Y4, Y5) allows all other matches to be assigned to one of these known Y RNA classes. The data suggest that the four Y RNA families are at least as old as the last common ancestor of tetrapoda and actinoperygian fishes. The Y RNA family as whole is much older: a single member has been found in *Caenorhabditis elegans* [417].

vault RNAs belong to a class of pol-III transcribed RNA genes with poorly understood function. Vaults are cytoplasmic ribonucleoprotein particles believed to be involved in multidrug resistance. The complex contains several small untranslated RNA molecules [418]. So far, vault RNAs have been described only for a few vertebrate species. Vault particles, however, are known also in the slime mold *Dictyostelium discoideum* [419], suggesting that vault RNAs are at least as old as Eukaryotes. The human genome contains at least 4 distinct vaultRNA genes, three of which are located in small cluster and share external promoter elements [418].

3.6 Small Nucleolar RNAs (snoRNAs)

Nascent rRNA transcript are matured in both eukarya and archaea [79, 309] with the help of a large number ribonucleoparticles that modify bases and direct cleavage. The human rRNAs, for instance, together contain more than 200 modified nucleotides [265]. The position of the snoRNA function is determined by the formation of a local snoRNA-rRNA duplex. Two major classes of snoRNA can be distinguished: The C/D box snoRNAs direct 2'-O-methylation of the ribose, while the H/ACA box snoRNAs guide the conversion of uridine nucleotides to pseudouridine. For details we refer to a series of reviews of snoRNA structure and function [432, 201, 16, 398, 149]

Besides their canonical roles in rRNA maturation, snoRNAs also target spliceosomal RNA. These snoRNAs perform their function in the Cajal bodies; for this reason they are sometime referred to as scaRNAs ("small Cajal-body associated RNAs") [201]. Most recently, three novel C/D box snoRNAs targeting U2, U4, and U12 snRNAs were identified, that, in contrast to all other known metazoan snoRNAs are independently transcribed [408]. In archaea, tRNAs are also targeted for modification [393], in trypanosomatids the spliced leader SL RNA is modified as well [245, 410]. An intriguing representative of this group is U85, a hybrid snoRNA that has both a functional C/D box and a functional H/ACA box domain that simultaneously modify the U5 snRNA [182]. Some snoRNAs lack complementarity to rRNAs or snRNAs. A small group of "orphan snoRNAs" (U3, U8, U22 and yeast snR10) directs rRNA cleavage instead of modification. The C/D box snoRNA U14, as well as the H/ACA box snoRNAs U17 (also called E1, and homologous to yeast sn30), E2 and E3, are both functional modification guides and play an additional role in pre-rRNA cleavage [95]. An increasing number of recently identified snoRNAs exhibits tissue-specific expression patterns in contrast to all snoRNAs that are known to modify rRNA or snRNA [50]. The genes of these, mostly brain-specific, RNAs are subject to genomic imprinting. Vertebrate telomerase [260], finally, contains a conserved H/ACA box snoRNA domain [284, 58].

The origin of the snoRNA machinery is still not well understood. The absence of snoRNAs from bacterial genomes suggests that snoRNPs arose in the archaeal and eukaryotic branch after the divergence of the bacteria. The K-turn motif, which forms the functional core of both classes of snoRNAs in archaea, on the other hand, also appears in bacterial RNAs including rRNA; it was probably present in the translation apparatus already before the last common ancestor [321]. This suggests a common origin of both modern ribosome and modern snoRNPs from a primitive translation apparatus [403]. The numerous box C/D and H/ACA RNPs of Archaea and Eukarya are likely to have arisen through duplication and variation of the guide sequence [217]. This scenario explains the lack of conservation of modified nucleotides shared between Archaea and Eukarya as well as the existence of tissue specific snoRNAs. In the following we demonstrate that this process is ongoing in vertebrate evolution.

The systematic investigation of snoRNA evolution is complicated by their fast evolution at sequence level. `blast` searches starting from human snoRNAs, for examples, are usually unsuccessful already in non-mammalian vertebrate genomes. As a starting point for investigating the evolution of snoRNAs we have therefore focused on the three snoRNAs that were first discovered [296], since sequences for these examples have been reported from a variety of different vertebrates. All three belong to the H/ACA class and are intron-encoded [369, 283].

The U17 (or E1) snoRNA is essential for the cleavage of pre-rRNA within the 5' external transcribed spacer (ETS) [95] with a length ranging from 200-230 nucleotides, longer than most snoRNAs; its secondary structure has been studied in detail [54]. Its sequence evolution in chelonians is discussed in [55]. Both E2 and E3 snoRNA are involved in the processing of eukaryotic pre-rRNA and have regions of complementarity to 28S rRNA. Gene trees reconstructed for these three examples are displayed in Fig. 6. While in many cases closely related paralogs are found, we can also identify ancient duplications that have been maintained in the genome over long times.

The evolutionary history of the paralog snoRNAs differs considerably between the three examples. The U17/E1 sequences for each species cluster together (with the exception of the Human and Chimp sequences), suggesting that the paralogs (which reside in adjacent introns) evolve via concerted evolution. In addition, however, the rodent genomes contain an additional paralog located on a different chromosome. In contrast, both for E2 and E3 we find two distinct evolutionary old paralog groups. In the case of E2 they separated before the advent of the tetrapods; the split between the two E3 groups predates the last common ancestor of the eutherian mammals. The six copies of E3 in the zebrafish apparently arose after the teleost-specific genome duplication [8].

The history of only a few other snoRNAs has been investigated in detail. Maybe the most interesting example is the C/D box snoRNA U36. It is homologous to snR47 in *Saccharomyces cerevisiae* and appears in two paralogs

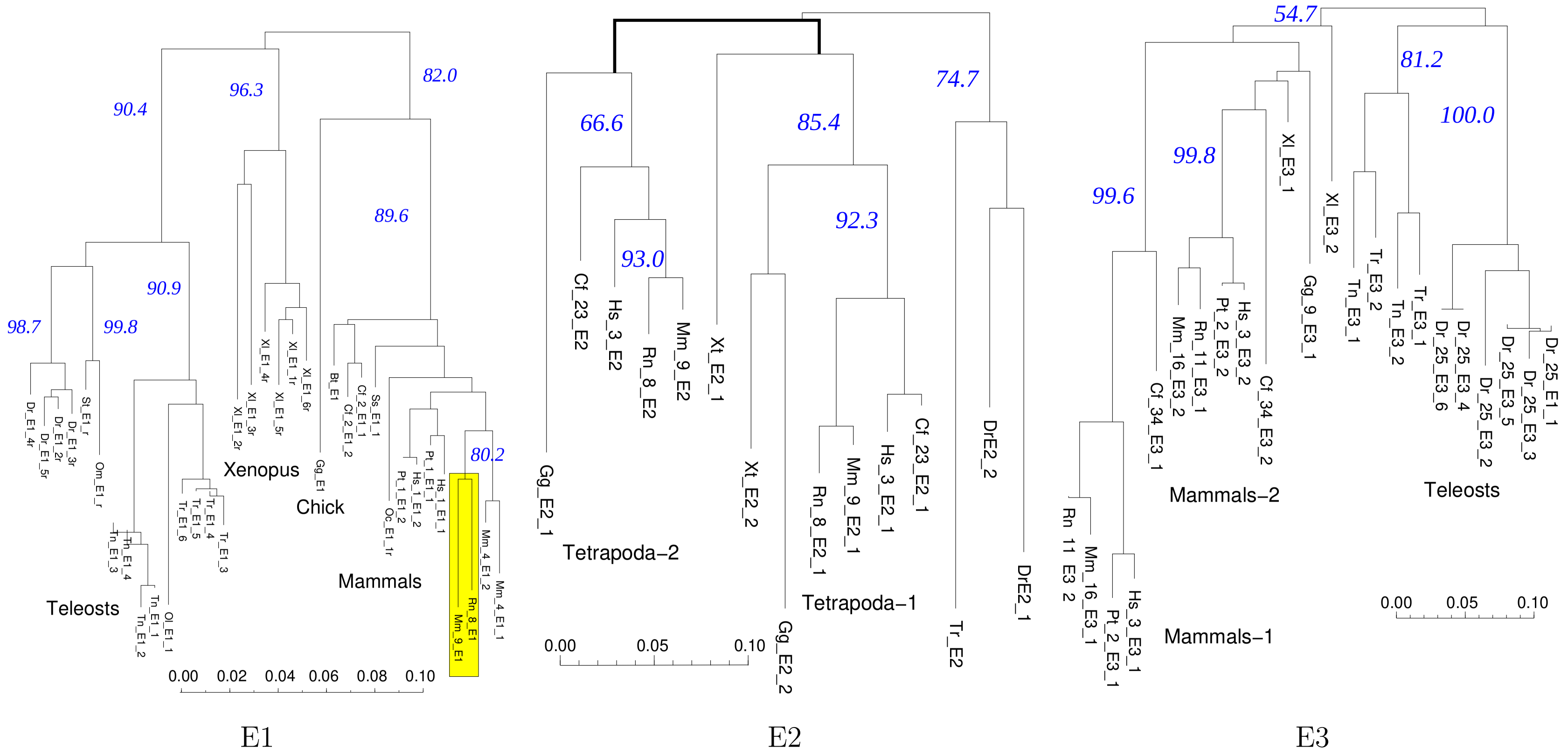


Fig. 6. Neighbor-joining trees of the E1, E2, and E3 snoRNAs. Bootstrap values from 1000 replicates are indicated in italics. The U17 sequences of *Takifugu* are taken from Acc.No. ***X94942*** [53]; Tr_E1_4 does not map unambiguously to a genomic location. The copies of the E1 snoRNAs that are located in a different host gene in rodents are highlighted.

in adjacent introns of the rpL7a gene in non-mammalian vertebrates. In mammals, however, U36a was duplicated with subsequent differential loss of function in mammals [119]. Other examples of snoRNAs whose evolution has been discussed in the literature include U14 [356] and U24 [119].

The patterns observed in Fig. 6 show that concerted evolution breaks down occasionally when two paralogs acquire functional or regulatory differences. The mechanism behind the concerted evolution of snoRNA copies is not well known. The identification of a retrogene with a poly(A) tail for H/ACA box snoRNA U99 [421], supports the idea that retro-transposition events play a substantial role in the mobility and diversification of snoRNA genes during evolution. This would argue for gene amplification [431].

3.7 Telomerase RNA

Telomeres are specialized protein-DNA complexes that cap chromosome ends that are essential for genome stability and cellular proliferation [106]. Sequence loss during replication is counteracted by specialized mechanism(s) in organisms with linear chromosomes [250]. In most organisms, the telomerase RNP extends chromosome ends by iterative reverse transcription of its RNA template, the telomerase RNA [198].

The secondary structures of the telomerase RNAs from vertebrates, ciliates, and yeast vary dramatically in sequence composition and in their size but share a common core structure [59, 72, 249, 452] that hints at an ancient origin. Plants also contain well-conserved telomerase, see [305] and the references therein; plant telomerase RNA, however, does not seem to have been studied systematically so far.

The vertebrate telomerase RNA apparently has co-opted a H/ACA box snoRNA domain [284] during its evolution, shares evolutionarily conserved proteins with H/ACA snoRNPs, and contains a Cajal body specific localization signal that is shares with a Cajal body specific subclass of H/ACA snoRNPs [181].

3.8 MicroRNAs

MicroRNAs (miRNAs) form a class of non-coding RNA genes whose products are small single-stranded RNAs with a length of about 22nt. These are involved in the regulation of translation and degradation of mRNAs. We refer to the recent review [299] for a discussion of their functions and mechanisms as well as their history of discovery. MicroRNAs are known in both multi-cellular animals

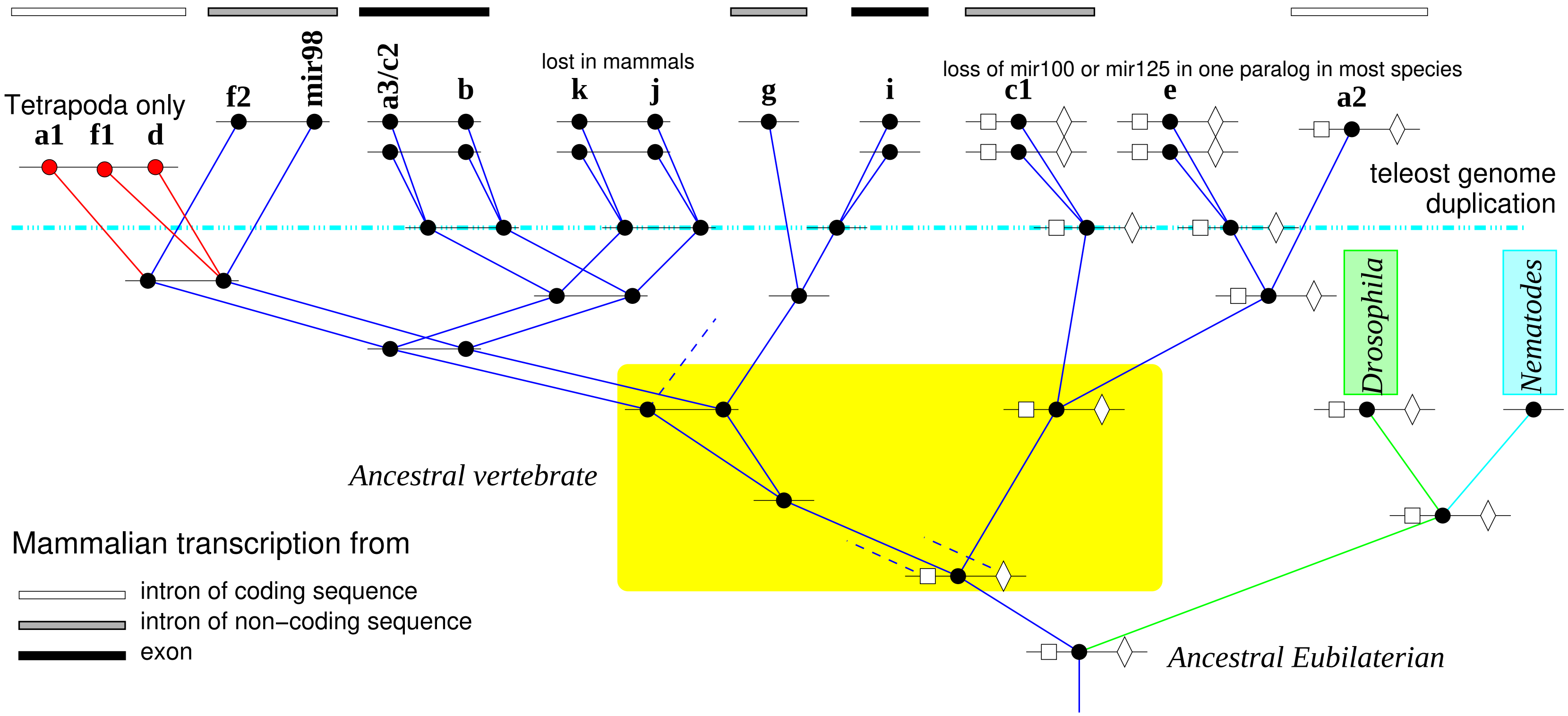


Fig. 7. A scenario for the expansion of the *let-7* family in vertebrates. There are few lineage-specific changes since the last common ancestor of teleosts and tetrapods; the inventory of *let-7* paralogs in tetrapods has essentially remained the same with the exception of the triple *let-7-a1*, *let-7-f1*, and *let-7-d1*. In teleost fishes we observe the loss of some paralog clusters and in particular the loss of linked *mir-100* (□) and/or *mir-125* (◊) copies in some of the sequences of the *let-7-a2/c1/e* family. In addition, the *let-7-j/k* pair has been deleted in mammals (in contrast to birds). The bars in the top line indicate the mode of transcription of the human and mouse sequences [347]. A recent computational survey [241] produced distant relatives of the vertebrate *let-7* sequences in *Ciona intestinalis* and *Ciona savigyi*. The sequences are too diverged, however, to determine whether they were produced by independent duplication of the ancestral *let-7* or whether they share part of the duplications shown in the figure.

and plants. A dedicated database, the `miRNA Registry` [130], at present [2] contains more than 1345 microRNA sequences from 12 species. Recently, several miRNAs were detected using the micro-array technique [19, 388].

Many of the known microRNAs appear in clusters on a single poly-cistronic transcript [239, 294, 220, 221]. The **mir-17** family, for instance, consists of numerous paralogs of three apparently non-homologous sequences. A detailed investigation of its evolutionary history [395] revealed a complicated sequence of tandem duplications within a cluster and duplications of entire clusters, which are probably linked to genome-wide duplications [166, 313]. Two microRNA families that are associated with the *Hox*-clusters have received considerable attention: *mir-10* and *mir-196* [446]. Again, an expansion of both families is observed that closely follow the vertebrate and teleostean genome duplications [394].

As a further example we consider the history of *let-7* family. The *let-7* gene was discovered in the *C. elegans* as a timing regulator in development [341]. The *let-7* microRNA is present in diverse animal phyla including chordates, echinoderms, mollusks, annelids, arthropods, nematods, chaetognaths, nemerteans, and platyhelminths, but it is absent in basal metazoa including cnidarians, poriferans, ctenophora, and acoel flatworms [315, 314]. In vertebrates a plethora of *let-7* paralogs are known. In Fig. 7 we present a reconstruction of the history of this microRNA family.

Mammals seem to share a more or less similar miRNA repertoire. More than 90% of the mammalian miRNAs listed in the `Rfam miRNA registry` v4.0 [131] can be found in human, mouse, and rat. In contrast, chicken and frog contain only 50-60% of the mammalian miRNAs, whereas teleost fishes harbor slightly more (50-65%). Since the chicken and frog genome sequencing and assembly is still incomplete, these numbers might change slightly in future studies. If the number of miRNAs increased linearly in evolution, *Ciona intestinalis*, an ascidian urochordate and hence close relative of the vertebrates, would be expected to contain about 30% of the miRNAs found in mammals. However, we were able to detect only about 15%. This suggests that the origin of vertebrates was associated with a dramatic expansion of the miRNA repertoire.

Fig. 8 summarizes statistical evidence. Only a small group of miRNAs, which includes *let-7* (discussed above), *mir-10* [394], and *mir-92* [395], can be found throughout most metazoans. These three families are characterized by numerous paralog miRNA genes at dispersed genomic locations and an additional expansion of families in teleosts. This points at a close association of the miRNA expansion with the genome duplications at the root of vertebrate tree

[2] Release 5.0, Sept. 2004

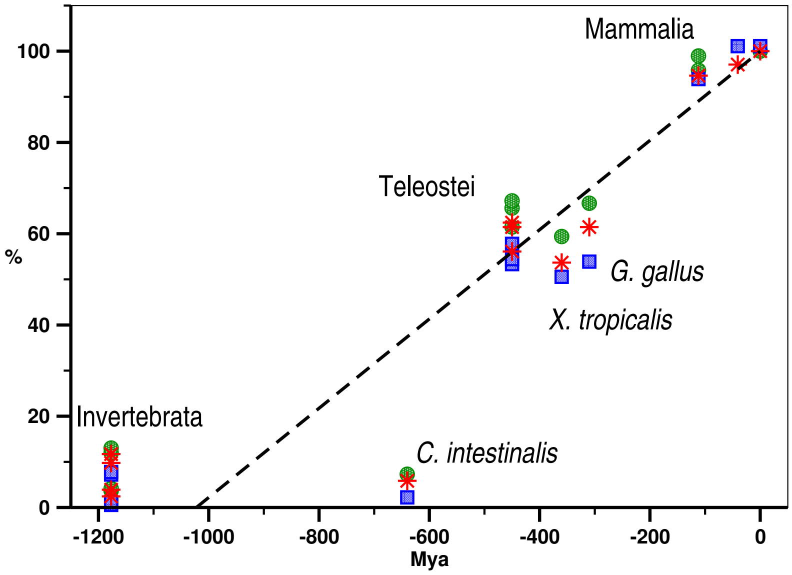


Fig. 8. Non-linear increase of miRNAs in Evolution. The human (green circles), rat (blue squares) and mouse (red stars) miRNAs listed in the `Rfam miRNA registry` v4.0 were blasted (cut off 10e-4) against the genomes of Invertebrata (*D. melanogaster, A. gambiae, C. elegans, C.briggsae*), *Ciona intestinalis*, Teleost fishes (*D. rerio, T. rubripes, T. nigroviridis*), *X. tropicalis*, *G. gallus* and Mammalia (*H. sapiens, M. musculus, R. norvegicus*). The percentage of mamalian miRNAs recovered are blotted against the evolutionary distance of those species.

[261, 99] and early in the evolution of the actinopterygian lineage [8].

The origin of microRNAs remains unknown. As yet, no microRNA with homologs in both animals and plants has been described so far, although the microRNA processing machinery is clearly homologous. In [395] it has been argued that microRNA could easily arise *de novo* since stem-loop structures resembling pre-miRNAs are very abundant secondary structures in genomic sequences. Most recently, a mechanism for the origin of new microRNA in plants from inverted duplications of expressed sequences has been proposed for the *Arabidopsis thaliana* sequences *mir161* and *mir163* [3]. In this scenario, the new microRNAs will target the mRNA they arose from. On the other hand, evolutionarily ancient microRNAs are also known in plants: miR166 is conserved between flowering plants, ferns, mosses, and hornworts. In addition to land-plants and metazoan animals, microRNAs have also been found in viral genomes, including the Epstein-Barr virus (Herpesviridae) [328] and HIV (Retroviridae) [26, 310].

RNA editing in trypanosome mitochondria is a unique post-transcriptional maturation process in which uridine residues are inserted and/or deleted at precise sites of mitochondrial mRNAs [38, 100, 126, 384]. Guide RNAs (***gRNAs***), which are usually transcribed from the kinetoplast DNA minicircles [168], provide the information for the editing.

In contrast, RNAediting mechanism (besides those the snoRNA-based base modifications) in other eukaryotes, prokaryotes, and viruses do not make use of RNA components [210, 30]. Models for the evolution of the gRNA-based editing process are discussed in [223], a phylogenetic analysis of U-insertion editing [224] suggests that extensive editing is a primitive genetic phenomenon that has disappeared in more modern organism, see also [376].

Probably the best-understood bacteria-specific non-coding RNA is the ***tm-RNA***, which is part of a ribonucleoprotein complex and combines the functions of tRNAs and mRNAs in order to rescue stalled ribosomes [143]. Usually tmRNA is a single molecule. At least three isolated clades in alpha-proteobacteria [197], cyanobacteria [113, 437], and beta-proteobacteria [372] have two-component tmRNAs, while jakobids have lost the mRNA-like region in their mitochondrial tmRNAs [179]. Reduction of the tmRNA structure in endosymbionts seems to be a common phenomenon [137]. The usefulness of tmRNA sequences for bacterial phylogenetics is demonstrated in [105] by revealing a structural feature that is characteristic for beta-proteobacteria.

Procaryots contain a diverse set of small non-coding ***sRNAs***. For example, a number of small (40-400nt) RNAs that neither encode proteins nor function as tRNAs or rRNAs, have been characterized in *E. coli* [152, 424]. The functions of many of these RNAs remain to be determined, while some of them are known to play crucial regulatory roles. There appear to be three general mechanisms: some are integral parts of RNP complexes, such as the 4.5S component of the signal recognition particle and RNase P RNA. A few, such as the 6S RNA, which regulated RNA polymerase activity [288], and the CsrB and CsrC RNAs mimic the structures of other nucleic acids, while a third class, reviewed in [383], acts by specific base-pairing with other RNAs. The co-evolution of the small RNA *micF* and its target mRNA *ompF* in Enterobacteria was studied in some detail [78]. A curious case are the MCS4 RNAs in mycoplasmas, which have a sequence similarity with eukaryotic U6 snRNAs. Homologs in other bacteria do not seem to exist [412], so that horizontal gene transfer from the host organism is a plausible explanation. Otherwise, very little is know about the origin and evolutionary relationships of the small ncRNAs in prokaryotes [127].

An increasing number of ***viral*** noncoding RNAs have been reported as well. Examples include the recently discovered viral microRNAs [26, 310, 328], the well-known VA1 RNA of adenoviruses [272], which is capable of inhibiting RNAi in human cells [256], the pRNA component of the packaging motor in some bacteriophages [18, 138]. One might suspect that at least some of the conserved RNA structure elements that were discovered in computations surveys of RNA virus genomes [165, 402, 439] are also non-coding RNAs rather than cis-acting elements.

3.10 mRNA-like ncRNAs

In eukaryotic cells, many RNA transcripts can be found that are not translated into protein. These so-called mRNA-like RNA transcripts are polyadenylated and spliced. In contrast to translated genes, they lack long ORFs [97, 98]. The best-known mammalian representatives of this rapidly expanding group are *H19* and *Xist*. Some of these large ncRNAs, including mammalian *Xist* and *Air*, and *roX* in Drosophila, have distinct roles in epigenetic gene regulation they are performed by means of chromatin modifications, reviewed in [9]. A number of plant specific mRNA-like ncRNAs are known experimentally; additional candidates were detected in a computational survey [3] of *Arabidopsis thaliana* ESTs [263].

The *Xist* (X-inactive specific transcript) gene is the only gene known to be specifically transcribed from the inactive X chromosome in female somatic cells [39]. It codes for a 17-kb spliced, polyadenylated non-coding RNA. *Xist* is necessary and sufficient for the initiation and spread of X inactivation [322]. The *Xist* gene is associated with an anti-sense transcript *Tsix* [235, 282] that is thought to be a repressor for *Xist*. A comparative analysis of the X-Inactivation Center (XIC) region, and the *Xist* gene in particular, in human, mouse, and cow is reported in [61]: while the *Xist* gene is well conserved among mammals with minor difference in the intron-exon structure, there is no apparent sequence conservation for the anti-sense transcript *Tsix*. Chureau *et al.* [61] also identified two new non-coding RNA gene, termed *Jpx* and *Ftx* in the XIC region, which are well-conserved in mammals.

The human *H19* gene is an imprinted gene that is exclusively expressed from the allele of maternal origin. It has a conserved secondary structure in mammals [193]. The *H19* gene is abundantly expressed in both extraembryonic and fetal tissues and is repressed after birth, except in a few adult organs. The possible functional relationship between H19 expression and tumorigenesis is still a matter of debate, as it seems to depend on the organ, the cell type and the cellular environment, see e.g. [28] and the references therein.

[3] `http://www.prl.msu.edu/PLANTncRNAs/`

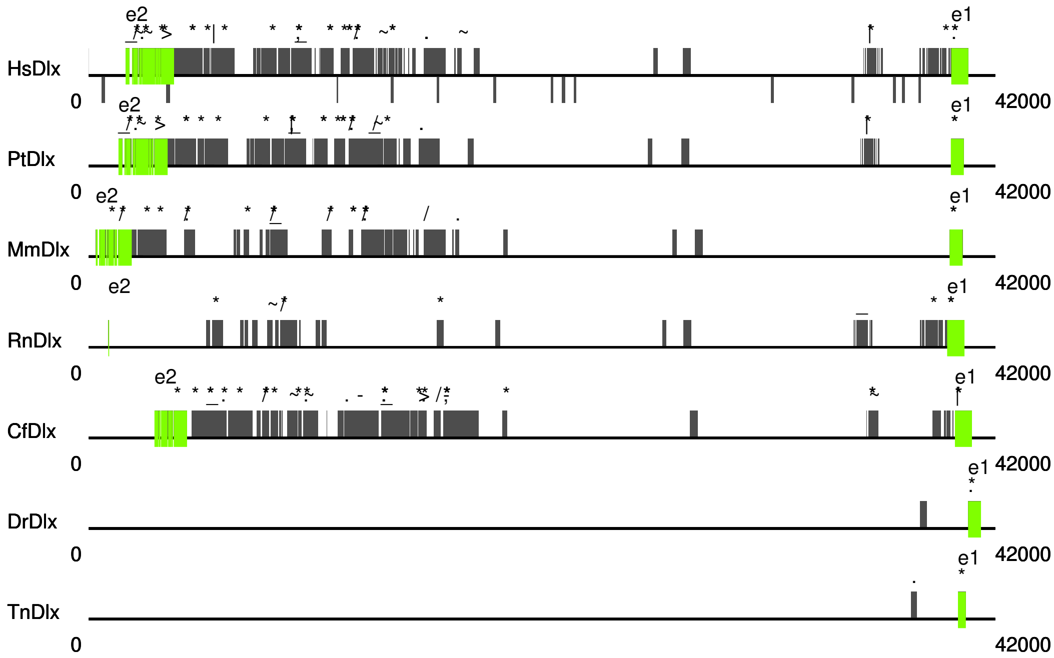


Fig. 9. Sequence conservation in the region 42kb upstream of the *Dlx6* gene. Boxes above the line mark phylogenetic footprints detected by `tracker` [336]. Conserved regions that lie in the known exons of *evf-1* are colored green and cross the line. Boxes below the line for the human sequence mark `RNAz` hits. Putative cis-acting elements as identified using `infernal` and the `Rfam` database are denoted by symbols above the `tracker` hits using the following symbols: * IRE; _ Hammerhead_1; / SECIS; ∼ REN-SRE; . Histone3; > U36; — Intron_gpII; , s2m; - tRNA. The following Sequences were used: HsDlx *Homo sapiens*, PtDlx *Pan troglodytes*, MmDlx *Mus musculus*, RnDlx *Rattus norvegicus*, CfDlx *Canis familiaris*, DrDlx *Danio rerio*, TnDlx *Tetraodon nigroviridis*.

As a third example, we describe here a computational analysis of the recently discovered mRNA-like ncRNA *evf-1*, which is located upstream of the *Dlx6* gene and its expression is linked to both Sonic hedgehog (*shh*) and *Dlx* genes [206]. *Dlx6* occurs clustered with *Dlx5*, another member of the same class of homeodomain transcription factors that are involved e.g. in the patterning and migration of ventral forebrain neurons, see [387]. Like *Xist* and *H19*, *evf-1* shows no homology to other known non-coding RNA sequences [206].

The *evf-1* genes consists of two exons that are divided by a single approximately 37.5kb large intron. We analyzed the DNA-sequence 42kb upstream of the *Dlx6* gene to find highly conserved regions in this genomic region, Fig. 9. The highly conserved regions were detected by `tracker` [336], a program for phylogenetic footprinting [453]. The so detected phylogenetic footprints were scanned for conserved RNA-secondary structures using `RNAfold` and `alidot` [161, 164, 163, 159] and assigned to known secondary structure elements according `Rfam` [131] using S. Eddy's `infernal` program [89]. `Infernal` suggests a possible annotation for 34 of the 79 `tacker` hits. A table listing all blocks of conserved sequence elements can be found in the electronic supplement, see also Fig. 9. The position of the two exons was inferred by `blast` comparison with the rat sequence (Acc. no. ***AY518691.1***).

In contrast to the mammalia-specific genes such as *Xist* and *H19* we find that *evf-1* shares at least *exon-1* and one large intronic sequence element with teleost fishes. A `blast` search also recovers *exon-1* from the xenopus and chicken genome. Since the genome assemblies of both the frog and the chick are incomplete in this region these sequences were not included in the analysis summarized in Fig. 9.

3.11 Antisense RNAs

Antisense RNAs predominantly act as post-transcriptional downregulators of gene expression [229]. Indeed, some of the RNA families discussed above can be viewed as antisense RNAs since they exert their function by binding complementarily to their target RNAs; examples are the microRNAs, snoRNAs, as well as many of the bacterial small RNAs [425]. The analysis of genomic sequence data, however, has revealed that a substantial fraction of transcribed DNA does not code for proteins and often derives from the anti-sense strand, see e.g. [194, 373, 447]. Antisense transcripts thus emerge as a common mechanism of regulating gene expression in eukaryotic cells, reviewed e.g. in [229]. Mechanistically, there are three major pathways: The formation of *double-stranded RNA* may trigger the RNAi pathway and lead to degradation of the sense transcript [144]. Binding of sense and anti-sense transcript may prevent the binding of other trans-acting factors (*RNA masking*). *Transcriptional interference* is the inhibition of transcriptional elongation due to a collision of the RNA Pol-II complexes on overlapping transcriptional units located at opposite strands [335]. Antisense RNAs are transcribed either *in cis* from the opposite strand, or *in trans* from a different genomic locus.

Many anti-sense transcripts are only poorly conserved in evolution, e.g. the *tsix* gene, which is the antisense transcript to the *Xist* ncRNA associated with X chromosome inactivation (sect. 3.10). On the other hand, a number of well-conserved antisense transcripts are known. Probably the best-studied example is the *HoxA11* antisense transcript, which is well-conserved between human and mouse and exhibits tissue-specific alternative splicing [334]. The Na/Pi co-transporter is essential in maintaining phosphate homeostasis in vertebrates. Antisense transcripts associated with the *npt* genes have been described in wide range of vertebrates [433] suggesting a conserved mode of transcription. Natural anti-sense transcripts have also been reported in mammals, insects, and fungi for genes that are part of the circadian clocks [68]. This system coordinates the expression of some 10% of the eukaryotic genes on a daily and seasonal timescale.

3.12 Natural Ribozymes

Until about 20 years ago, it was firmly believed that proteins were the only catalytic macromolecules in biology. The discovery of the first catalytic RNA molecules, or ribozymes, in the early 1980s, however, has changed this pictures considerably. We have already encountered several examples: RNase P, the spliceosome, and the ribosome are essentially ribozymes. In most cases, ribozymes serve an RNA-processing function using RNA as substrates. The majority of known ribozymes have been created in artifical selection experiments and hence are not a topic of this contribution; for a recent review of artifical ribozymes the interested reader is referred to [192] and the references therein.

A number of natural ribozymes, however, are not independently stable ncRNAs but rather are part of larger RNA molecules. For example, there are four distinct groups of nucleolytic ribozymes: hammerhead and hairpin ribozymes are mostly found in plant viruses, the Varkud satellite (VS) ribozyme was found in fungal mitochondria, and hepatitis delta virus contains another ribozyme. A recent study suggests a common origin of hammerhead, hairpin, and hepatitis delta ribozymes [145], although convergent evolution cannot be ruled out.

The second large class of naturally occurring ribozymes is involved in the self-splicing of introns in a wide range of species; these molecules belong to one of two structural classes known as *group I* and *group II* ribozymes. All these ribozymes perform different kinds of phosphoryl transfer reactions, in which a transesterification reaction results in breakage of the backbone in the first step [248]. Since they behave rather like mobile genetic elements they are outside the scope of this survey; indeed many group II introns carry their own ORFs, see e.g. [454].

4 Transcription of ncRNAs

Some non-coding RNAs can be found by searching for likely transcripts that do not contain an open reading frame. A survey of the *Escherichia coli* genome for DNA regions that contain a $\sigma 70$ promotor within a short distance of a *Rho*-independent terminator, for instance, resulted in 144 novel possible ncRNAs [60], see also [11, 429, 263] for similar studies. This approach is limited,

Table 6
Major modes of transcription

RNA Polymerase	Promoter	Location relative to start site	Transcript	Function
Pol I	core element UCE (upstream control element)	-45 to +20 -180 to -107	pre-rRNA (28S, 18S, 5.8S)	components of the ribosome; translation
Pol II	TATA-Box Initiator CpG islands	-25 to -35 -100	mRNA	protein coding genes
			snRNA (U1-4)	components of the spliceosome; mRNA splicing
	no		LINEs	Retrotransposon
Pol III	A-box, B-box, C-box	+50 to +80	5S rRNA	component of large ribosomal subunit
			tRNA	translation
			snRNA (U6)	components of the spliceosome; mRNA splicing
			7SL RNA	component of the SRP (signal recognition particle); protein transport to ER (endoplasmatic reticulum)
			SINEs	Retrotransposon

however, to functional RNAs that are transcribed in the "usual" manner, see Table 6. For many ncRNAs, however, the mode of transcription is unknown.

RNA Polymerase I transcribes rDNA transcription units to 18S, 5.8S and 28S rRNAs in the nucleolus. The rDNA promoters consist of the start site proximal core promoter (CP) resembling a TATA-box and an upstream control element (UCE). Both CP and UCE show poor sequence but strong structure conservation. Decreases in cell growth and protein production also reduce rRNA transcription; rRNA transcription activity oscillates during the cell cycle, showing maxima at S and G2 phase and is repressed during mitosis. In general, acetylation and phosphorylation of basal TFs regulate pol-I transcription. These modifications are performed e.g. by components of the MAPK pathway or tumor suppressors. For reviews see e.g. [317] and [133].

Another important class of non-coding RNA genes are transcribed by pol-III. Besides all canonical tRNAs and the 5S rRNA, this group includes the U6, and presumably U6acac snRNAs, RNase P and RNase MRP RNA, 7SK RNA, selenocystein tRNA, Y-RNAs, and vault RNAs [363]. Furthermore certain repetitive elements including SINEs are pol-III transcripts. For a detailed description we refer to [435].

The majority of transcripts is produced by pol-II, however. Most vertebrate snoRNAs are processed from introns of either protein coding genes or of "host genes" whose only known purpose is to carry an snoRNA in its intron(s) [409], see also [16]. Some snoRNAs, however, are transcribed directly from monocistronic or poly-cistronic genes, notably the U3, U8, and U13 snoRNAs. These share their promoter structure with a group of non-coding RNAs that contains the spliceosomal RNAs and the U7 RNA which is involved in histone mRNA processing [151]. In vertebrates almost all snoRNAs are encoded in introns of a specific subclass of pol-II transcripts, the TOP genes, whose promoter elements determine a specific ratio of snoRNA and mRNA production [77]. Many vertebrate snoRNAs appear in multiple copies in different introns of the same gene, sometimes paralogs are located even on different chromosomes, see Fig. 10. The recent discovery of H/ACA snoRNA clusters within individual introns in *Drosophila* a different expression strategy for a box H/ACA snoRNA compared to box C/D snoRNAs in this species [172].

The association of intron-encoded snoRNAs with their surrounding gene surprisingly is not stable over long time-scales. U17, for example, is located in Rps7 in tetrapoda, while it is associated with the unrelated CHC1 protein in teleost fishes, Fig. 10. In addition, rodents have an additional copy of E1 in intron 3 of the lamin gene, which carries the E2 snoRNA in vertebrates. E3 switches from the ribosomal protein RPL0 to a "host gene" whose exons do not code for a functional protein.

MicroRNAs are processed from long primary precursors (pre-miRNAs) [239, 238]. Unlike the majority of snoRNAs, neither the genomic location of miRNAs coincides with a specific genomic context, nor is their transcription performed by a single typical mechanism. A recent survey of mammalian genomes showed that there are five major classes [347]: About 30% are directly transcribed by RNA Polymerase II and a 5' cap as well as a poly(A) tail is added [44], as shown for the *mir-23* cluster [240]. 40% of the mammalian miRNAs are probably processed from introns [450, 449] of protein coding genes, 10% of the known microRNAs reside in introns and another 13% in exons of non-coding transcripts. Antisense transcripts account for 14% of all mammalian miRNAs.

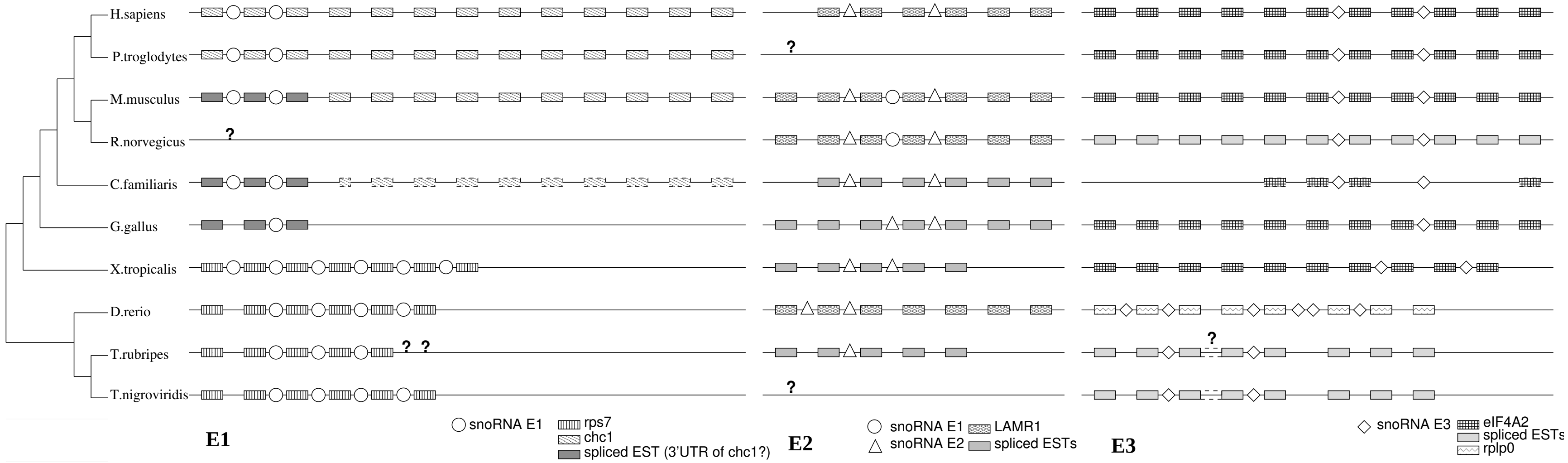


Fig. 10. Organization of snoRNA paralogs in the introns of their associated genes.
All SnoRNAs E1, E2 and E3 identified so far, reside within introns. E1 changed its host gene from rps7 (ribosomal protein S7) to chc1 (chromosome condensation 1); E3 switched from rplp0 (ribosomal protein, large, P0) to elf4A2 (E74-like factor 4, an ets-domain transcription factor). Only snoRNA E2 remained stable associated with its host gene LAMR1 (laminin receptor 1 [ribosomal protein SA]), which gained an additional copy of E1 in rodents. Question marks indicate incomplete genome data. For details like gene accession numbers or genome coordinates we refer to the supplemental material.

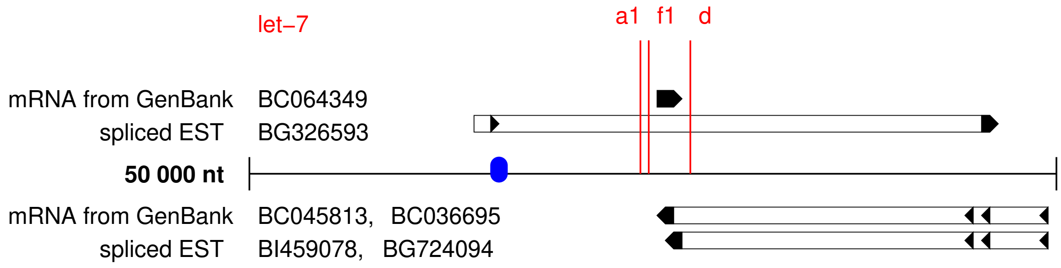


Fig. 11. Genomic environment of the human *let-7* family members *let-7-a1*, *let-7-f1*, and *let-7-d*. Known transcription units on the plus strand are shown above the line, the area below the line implies a location on the minus strand. The dot indicates a cluster of phylogenetic footprints (detected by `tracker`) that is conserved at least among amniotes. An ortholog of this particular *let-7* miRNA cluster was not found in the unfinished genome of the frog *Xenopus tropicalis* (both v2.0 and v3.0).

The remaining cases are of uncertain transcriptional origin [347].

Interestingly, our reconstruction of the duplication history of *let-7* in Fig. 7 shows that different mammalian members of this family occurs in introns of both protein-coding and non-protein-coding genes as well as in exons. Preliminary data [394] suggest that microRNA genes, like the genes of the three snoRNAs in Fig. 10, can "move around" in the genome: Only one fourth of the known human microRNAs are located in annotated genes with known homologs in mouse and chicken. Of these 56 human microRNAs, however, less than half have known orthologs (in the `Rfam Micro RNA Registry 4.0` which contains the results of a survey of the chicken genome) in the homologous genes.

We searched the genomic vicinity of *let-7-d*, which also contains *let-7-a1* and *let-7-f1*, for conserved non-coding DNA using the phylogenetic footprinting tool `tracker` [336]. This cluster of 3 let-7 miRNAs appears to be located within an intron of a single transcription unit (suggested by spliced EST data) with an approximate start site about 10kb upstream of *let-7-a1*, Fig. 11. An other transcript on the plus strand with its start site between *let-7-f1* and *let-7-d* is also a possible host transcript for *let-7d*. The transcription unit OTTHUMG00000020259 (Vega database[4]), which is implicated in [347] as carrying *let-7d* in its intron, is located on the opposite strand, however.

Approximately 500 nucleotides downstream of the transcription start site, we detected a large phylogenetic footprint cluster (100-150nt) that is conserved among amniotes, while it is not conserved in the most closely related *let-7* cluster in Actinopterygian fishes (consisting of orthologs of *let-7-f2* and *mir-98*).

[4] `http://vega.sanger.ac.uk/`

The footprint cluster does not correspond to a additional microRNA or another unannotated ncRNA in the cluster, since `RNAz` [427] classifies this region unambiguously as not containing a conserved RNA structure. A search for transcription factor binding sites within the footprint cluster using `tfsearch` revealed a set of common sites CREB, MZF1, GATA-1 /2, Nkx-2, NrF-2 or c-Ets and Elk-1, suggesting a function in the regulation of this *let-7* microRNA family.

5 Modifications of ncRNAs

Many, if not most, of the non-coding RNAs are post-transcriptionally modified. We have already encountered the snoRNA guided pseudouridinylation and ribose 2'-O-methylation of ribosomal RNAs and spliceosomal RNAs. The target sites of these modification are well conserved over long evolutionary time-scales, a fact that allowed the usage of yeast rRNA methylation sites in the search for snoRNAs that modify homologous position of fruitfly rRNAs [1].

More than 80 different nucleotide modifications are listed in the "Compilation of tRNA sequences and sequences of tRNA genes" [5] of various organisms [380, 379, 381]. These are achieved by a large family of often highly conserved enzymes, see [262] and [169] for reviews. In Archaea these modifications have been shown to require four snoRNAs, one of them encoded within the intron of tRNATrp [63]. In *S. cerevisiae* tRNA genes were shown to co-locate with the nucleolus [401]. Within this nuclear structure rRNA transcription and processing including modification by snoRNAs takes places. Recently, a selenocystein tRNA was co-immunoprecipitated with Cbf5p from Euglena, a putative pseudouridine synthase usually associated with H/ACA snoRNAs modifying rRNAs, [353]. However, no tRNA modifying snoRNAs have been detected in eukaryotes so far. Studies of the evolutionary aspects of RNA editing have focused on the enzymes. Ref. [262], for instance, describes the evolution of the superfamily of RNA-dependent deaminases. We are not aware, however, of a systematic study of the evolution of the chemical modifications themselves. The 5' part of tRNAs is edited in some organisms by replacing mismatched nucleotides with nucleotides capable of forming Watson-Crick in order to obtain a canonical terminal stem. This mechanism is known e.g. in the rhizopod amoeba *Acanthamoeba castellanii* and the chytridiomycete fungus *Spizellomyces punctatus* and appears to have arisen independently at least twice [218].

In *C. elegans*, the pre-miRNA of let-7 has been shown to undergo *trans*-splicing

[5] `http://www.trna.uni-bayreuth.de/`

to the spliced leader 1 (SL1) RNA. This process allows folding of the pre-miRNA such that the miRNA precursor (pre-miRNA) forms a stem-loop structure, which in turn is cleaved by nuclear RNase III Drosha [107, 46, 239, 238]. Both, the mature miRNAs and the pre-miRNA can undergo A to I RNA editing by an RNA-specific adenosine deaminase (ADAR) [257].

6 RNA Motifs Associated With Protein-Coding mRNAs

6.1 mRNA Structure

In contrast to non-coding RNAs, the primary function of messenger RNAs is to encode in its exons the information that allows the translation machinery to generate proteins. Exon recognition by the spliceosome can be affected by many features of the pre-mRNA including exon length, promoter architecture, the presence of enhancer and silencer elements, the strength of splicing signals, and RNA processivity. It has also been proposed repeatedly that pre-mRNA secondary structures influence splicing activity. A recent review of the topic [41] strongly suggests that many pre-mRNA sequences contain selected regions folding *in vivo* into well-defined secondary structures that are likely to play a role in the splicing process.

Eukaryotic mature mRNA exhibit a tripartite structure: an untranslated region at the 5' end (5'UTR), the coding regions which is translated into amino acids, and an untranslated region at the 3' end (3'UTR). With the exception of the replication dependent histone genes mentioned above, the 3' end of the mature mRNA carries a poly(A) tail. Both untranslated regions are involved in the post-transcriptional regulation of gene expression processes, like subcellular localization, mRNA stability and translation efficiency [230, 73, 307, 326, 436]. These processes are mainly controlled by cis-acting functional elements in the UTRs, which comprise both sequence motifs and RNA structure motifs. Short sequence motifs may be potential binding sites for trans-acting factors, while longer sequence motifs found in UTRs have been hypothesized to be antisense RNA binding sites [251]. In addition, a number of motifs are known that are determined by structural features rather than nucleic acid sequence.

In general, the protein coding region of mRNAs is much better conserved than the UTRs. The distribution of conserved sequence motifs in the UTRs is not uniform: the 3'UTR is typically better conserved than 5'UTR and introns [185]. For example, 30% of 3'UTRs in different vertebrate mRNAs contain highly conserved regions which are at least 100 nt long and show at least 70% similarity [87]. The overall higher conservation of the 3'UTR may be a consequence of the observation that post-transcriptional regulation in 3'UTRs is rather based on protein complexes than on single proteins [436]. Antisense binding, as e.g. in the case of microRNAs, also leads us to expect many non-structural binding motifs. This is in particular the case in plants, where microRNA targets typically closely match the corresponding microRNA, see [200] and the references therein. (In mammals, however, the requirements for mRNA-miRNA interactions appear to be much more complex [420].) In contrast, 5'UTRs regulatory motifs might be mostly structural motifs. It is known, for example, that translation initiation is essentially controlled by RNA structures in 5'UTRs [73, 326].

A more detailed analysis showed, however, that pattern of conservedness is reversed at the border of the coding region. The 30nt of the 5'UTR immediately upstream of the start codon is the best conserved regions with $70\% - 80\%$ sequence identity between human and mouse; in contrast, the 3'UTR is very poorly conserved immediately downstream of the stop codon [285, 370]. This pattern can be explained by the specific interaction with sequence specific binding factors initiating translation in the 5'UTR on the one hand, and the fact that the first segment of the 3'UTR is covered by ribosome and hence inaccessible to specific factors at the termination of translation.

The UTRs and the coding regions are subject to different functional constraints and hence evolve differently; even the 5'UTRs and 3'UTRs of the same gene do not necessarily share the same evolutionary dynamics [226]. In addition there are also mRNAs that encode nearly identical proteins but have highly diverged UTRs [87, 185, 226], suggesting that the divergent UTRs form specific translational regulation patterns which enables them to reply differently to variable stimuli.

6.2 Detection of UTR Motifs

A handful of cis-acting regulatory motifs in mRNAs have been characterized experimentally; these are collected in the `UTRsite` [325] and `Transterm` databases [71]. Functional RNA structures in UTRs are in general not as long as ncRNAs, since they are limited by the size of their UTRs. The average length of human UTRs is about 210 for 5'UTRs and about 1027 for 3'UTRs [326]. RNA structure motifs in UTRs can thus be expected to be relatively small, simple structures. This limits the usable information, and hence the fast and reliable prediction structural regulatory elements in UTRs has remained a largely unsolved problem.

Standard sequence alignment procedures usually fail to align UTRs in a meaningful way [185]. Detecting structural motifs in UTRs will therefore require algorithms that optimize sequence alignment and secondary structure simultaneously. Existing methods, which characterize putative motifs automatically, can be classified in (a) methods which require the description of a motif and search for similar instances of such a motif or (b) methods which search for motifs that are significantly overrepresented in a dataset.

While most regulatory motifs found in UTRs are conserved in secondary structure, some known motifs show conservation on the sequence level. In general such sequence motifs do not require an exact nucleotide substring, but allow some variation in nucleotide composition, or may consist of several conserved fragments separated by unconserved regions. Fragmented motifs may for example occur in regulatory structures, which exhibit sequence conservation in loops.

Many of the tools developed to identify functional RNA motifs in general have also been applied to UTRs. Among these `Palingol` [29], `PatSearch` [324] and `RNAMotif` [264] identify instances of a previously defined motif descriptor. Because of the limited information available in a UTR motif it is hard to define descriptors that are both specific and sensitive by hand. Other tools, therefore, are designed to require only limited information about a known motif and recognize motif features, which discriminate sequences containing the motif from sequences not containing it, automatically (see `ERPIN` [114]).

An even harder problem has to be solved when the motif is completely unknown. The detection problem in such a case can be treated as a classification problem: From an arbitrarily given set of UTRs all UTRs sharing a common motif shall be classified in the same group. The first approaches to this problem are implemented in the tools `comRNA` [187], which identifies novel mRNA structure motifs by clustering similar stems, and `RNAProfile` [318], which identifies the most conserved motif in a set of sequences where at least some share the same common motif.

6.3 Important Regulatory Motifs in UTRs of mRNAs

Gene expression is controlled by cis- and trans-acting factors during both transcription and during translation. By regulating translation a cell is able to respond quickly to environmental changes. The mature mRNA already resides in the cellular plasm but the amount and type of protein which will be translated depends on several cellular conditions. Cis-acting elements in the untranslated regions of mature mRNA bind trans-acting factors and control in this way translational efficiency, mRNA stability and subcellular localization. A selection of examples of such regulatory motifs in UTRs will be given here, see also table 7.

Iron response elements (IRE) are short hairpin structures with an internal loop and a conserved sequence in the hairpin loop, which are observed in 5' UTRs of ferritin mRNAs in 3'UTRs of of transferrin receptor mRNAs [150]. They can be classified in two slightly different instances, the first containing an internal loop of length three, which is replaced by a bulge loop in the second. Both have the primary consensus motif `CNNNNNCAGWGH` [325]. The IRE motif can be readily described with regular grammars; because of the highly redundant sequence pattern and frequent, simple secondary structure one has to expect a large number of false positives, however.

Translation control elements (TCE) are short elements (~ 90 nt) found in the 3'UTR of nanos mRNA of drosophila [69]. Its secondary structure is composed of a helix and a multiloop with two hairpin loops branching off, one with a conserved primary structure in the hairpin.

Internal ribosome entry site (IRES) elements were first described in the 5'-untranslated region of picornavirus RNA [184]. The IRES element enables cap-independent initiation of translation starting at an internal initiation codon. In addition to several types of viruses, which contain an IRES element, a small group of eukaryotic mRNA can be translated by internal ribosome entry. IRES-containing mRNAs mostly encode regulatory proteins such as, e.g., growth factors and transcription factors. Several studies have reported that under stress conditions, where cap-dependent translation is blocked, translation of specific mRNAs is enabled through IRES elements ([270] and references therein). Another function of IRESs involves the control of alternative initiation of translation. For example, the human fibroblast growth factor 2 contains 5 translation initiation codons. Translation initiation of the codon proximal to the 5'-end is initiated by a cap-dependent process, whereas initiation of the remaining codons depends on the IRES [33]. IRES elements are defined by functional criteria and cannot yet be predicted by the presence of characteristic RNA sequence or structural motifs. In general, there are no significant similarities between individual IRESs unless they are from related sources.

Selenocysteine insertion sequences (SECIS) is found in the coding region of some eubacterial mRNAs and in 3' untranslated regions of some mRNAs in archaea and eukaryotes [215]. In eubacteria, it forms a hairpin structure of conserved length with the selenocysteine codon in the outer helix. In archaea, the primary rather than the secondary structure is conserved. The consensus is a hairpin structure that differs in stem length, occurrence of internal loops and size of the hairpin loop, but it has a very conserved sequence motif in the helix beneath the apical loop. In eukaryotes, the secondary structure contains most of the information while only small sequence motifs are conserved. The core secondary structure is composed of a long hairpin structure consisting of two (type 1) or three (type 2) consecutive helices [102, 215].

At present, it is unclear whether large regulatory motifs such as IRES, IRE,

Table 7
Important general regulatory motifs in mature eukaryotic mRNA [326]. Most regulatory elements influence initiation of translation. Regulatory elements specific to mRNAs of particular genes are not listed.

Motif	Function	Description	References
		5'UTR	
m^7G cap structure	stabilization, initiation	prevents processing of mRNA from 5' to 3' end, hence stabilizes mRNA; eIFs bind to cap, which governs pre-initiation complex with small ribosomal subunit and initiates scanning	[73, 115]
initiation codon	initiation, translational efficiency	efficiency of translation start recognition depends on primary sequence context of `AUG` codon; optimal context for vertebrates is `(A/G)CCAUGG`	[211, 73, 323]
uORF	translational efficiency	inhibits translation by leaky scanning: scanning complex may either bypass upstream start codon depending on sequence context and mean ORF is translated or may start translation at upstream start codon	[73, 115]
IRES	initiation	alternative to ribosomal scanning; pre-initiation complex interacts with IRES element and scanning starts at this site	[230, 213, 281]
stable RNA structures	translational efficiency	very stable secondary structures in 5'UTRs can impede scanning	[212, 281, 436, 115]
Repeats	e.g. initiation, translational efficiency	*Alu*-elements in 5'UTRs e.g. repress translational efficiency; reason may be repression of initiation by *Alu*-elements forming stable secondary structures or containing weak start codons	[222]
		3'UTR	
Zipcodes	localization	RNA binding proteins bind to different zip codes and direct mRNA to subcellular region where corresponding protein is translated; proteins recognize zipcode by primary and tertiary structure	[307, 153]
poly(A)-tail	stabilization, initiation	prevents processing of mRNA from 3' to 5'end; interaction with pre-initiation complex via PABP activates translational initiation	[73, 436, 115]
CPE	stabilization, translational efficiency	CPEB binds at CPE and induces polyadenylation; a complex of CPEB and Maskin bound to CPE interacts with `cap` structure by binding to the *eIF4F* complex and translation is repressed	[278, 436, 115]
AREs	stabilization	influence rate of deadenylation depending on type of `AU`-rich element (ARE 1, ARE 2 or ARE 3)	[281, 277]
miRNA target sites	stabilization	miRNAs evoke degradation of mRNA by imperfect base-pairing interactions with 3'UTR	[436, 115]
Repeats	e.g. localization	`CAG/CUG` repeats in 3'UTRs e.g. result in very long mRNAs, which show in yeast different subcellular distribution	[101, 281]

Abbreviations: CDS = coding sequence, IRES = internal ribosome entry site, AREs = AU-rich elements, uORF = upstream open reading frame, CPE = cytoplasmic polyadenylation element, CPEB = cytoplasmic polyadenylation element binding protein, ACE = adenylate control element, eIFs = eukaryotic initiation factors, PABP = poly(A)-binding protein, Repeats found in UTRs include short interspersed elements (SINEs), long interspersed elements (LINEs), mini- and micro-satellites [281]; these are not listed above.

or SECIS elements, arose independently in different genes or gene families or whether there are mechanisms that allow their lateral spread within a genome.

Novel UTR Motifs. In addition to post-transcriptional regulatory mechanisms that are specific to a particular gene or gene family, there exist also mechanisms which are observed in a broader range of mRNAs (table 7). Such relatively non-specific regulatory processes are characterized by similar primary and/or secondary structures in mRNAs of different genes of the same organism. We performed a search for sequence elements of this type in the human genome.

Using `NCBI blast` [5], we compute pairwise alignments of repeat masked human UTRs from `Ensembl` database (release 24). The majority of UTRs were not conserved on the sequence level, suggesting that also non-specific regulatory motifs show large sequence divergence. Furthermore, we found many more conserved sequence blocks in 3'UTRs than in 5'UTRs. From the pairwise alignments we built a weighted similarity graph to identify clusters of UTRs with conserved regions by complete linkage clustering [74]. Alternative transcripts were not allowed to occur in the same cluster. Sequences of each cluster were aligned using `dialign2` [291] in order to identify putative regulatory sequence motifs.

We then used `RNAz` [427] to check whether some of these multiple alignments contain correspond to conserved RNA secondary structures. Among 481 5'UTR multiple alignments, 10% had regions forming with high probability stable RNA structures. Among the set of 1223 multiple 3'UTRs alignments, 21% alignments contain stable RNA structures. Table 8 lists the annotation of the best RNA predictions using `infernal` and the `Rfam` database. All significant hits matched the iron response element. The corresponding genes, however, are not known to be involved in the iron metabolism. We suspect that at least some of these cases form IRE-like structures that do not function as IREs, indicating that still more specific descriptors for UTR elements including IRE are desirable.

The small fraction of RNA motifs with known function that was recovered in our survey suggests that most non gene-specific mRNA motifs have very little well-conserved sequence information and most of them, including IRES, SECIS, IRE, and many others depend crucially on secondary structure. On the other hand, we detected hundreds of statistically significant sequence patterns that occur in multiple RNAs for which so far no function has been described. A pattern is defined by the consensus sequence of a run of gapless columns in the multiple alignments.

One possible function of sequence patterns in the 3'UTR of mRNAs is to act as target sites for microRNAs [436, 115]. We therefore tested all gapless

Table 8
RNA structure annotation based on `infernal` and description of corresponding genes.

Rfam model	Score	Genes	Description (Ensembl release 24)
		5′ UTR	
IRE	6.8	ENSG00000120853 ENSG00000166104	- -
IRE	6.31	ENSG00000092199 ENSG00000159267	Heterogeneous nuclear ribonucleoprotein Biotin-protein ligase
IRE	9.89	ENSG00000129873 ENSG00000172288 ENSG00000172353	Testis-specific chromodomain protein Y2 Testis-specific chromodomain protein Y1 Testis-specific chromodomain protein Y1
		3′ UTR	
IRE	8.24	ENSG00000066294 ENSG00000134822	CD84 antigen fatty acid desaturase
IRE	8.26	ENSG00000165282 ENSG00000152056	Phosphatidylinositol-glycan biosynthesis Sigma-adaptin 1C
IRE REN-SE	8.28 12.1	ENSG00000110436 ENSG00000171596 ENSG00000166676 ENSG00000090659	Amino acid transporter 2 G protein-coupled receptor 66 - CD209 antigen
IRE	8.33	ENSG00000181894 ENSG00000149451	- ADAM 33 precursor
IRE	11.29	ENSG00000185753 ENSG00000064115	- Transmembrane 7 superfamily protein member 3 precursor
IRE	12.51	ENSG00000012048 ENSG00000156675 ENSG00000142687	Breast cancer type 1 susceptibility protein Rab coupling protein Polycystic kidney disease 1-related
IRE SECIS	9.19 10.87	ENSG00000181719 ENSG00000178887 ENSG00000180747	- - -

Table 9. Potential miRNA target sites in human 3' untranslated regions.
We report z-scores for `miRanda` and p-values for `RNAhybrid` as computed by these tools. Gene annotation is taken from `Ensembl` (release 24).

miRNA	Gene		Score		Protein family	
	ENSG00000...	Protein	`miRanda`	`RNAhybrid`	ENSF0000000...	Name
hsa-miR-187	183850	Zinc finger protein 254	11.65	0.000096	0001	Zinc Finger
hsa-miR-187	181342	-	11.11	0.000081	0001	Zinc Finger
hsa-miR-134	065371	AKAP-binding sperm protein ropporin	10.50	0.000914	5520	-
hsa-miR-134	114547	AKAP-binding sperm protein ropporin	10.50	0.000914	5520	-
hsa-miR-324-5p	129277	Small inducible cytokine A4 precursor	8.41	0.000062	0592	-
hsa-miR-324-5p	189315	Small inducible cytokine A4 precursor like	8.36	0.000062	0592	-
hsa-miR-184	177111	-	11.49	0.000143	2097	DPY19
hsa-mir-184	177990	-	11.03	0.000143	2097	DPY19

regions in the multiple alignments for potential miRNA target sites. To this end we used the collection of all human microRNAs from `Rfam` (release 5.0, Sept. 2004) [130] and two different miRNA target prediction tools: `miRanda` [94] and `RNAhybrid` [340]. Table 9 lists the best-scoring candidates. In contrast, an analysis of the 5'UTRs and their flanking regions did not yield a potential site located within the untranslated region of mRNA that was predicted by both methods.

6.4 RNA Structures in Coding Regions

It is widely believed that RNA structures in ORFs can interfere with translation, although this phenomenon has not been studied systematically to our knowledge [195]. It is plausible to assume that coding regions are therefore largely devoid of secondary structures. There are, however, a number of well-known exceptions to this rule. A variety of conserved secondary structure elements have been detected in computational surveys of single stranded RNA virus genomes [165, 406, 405, 402, 439]. A comparative study of 28 different species [195] provides evidence for wide-spread selection for local secondary structures in mRNAs, in particular in eubacteria. Most recently, Pedersen *et al.* [320, 319] devised an SCFG-based algorithm for detecting conserved secondary structures motifs specifically within coding sequences.

The *Rev Response Element* (RRE), for example, forms a five-fingered motif spanning some 300nt [75], located in the *env* gene of HIV. The structure is well conserved among diverse HIV strains, see e.g. [161, 164, 208]. The interaction of RRE with the *Rev* protein reduces splicing and increases the transport of unspliced and single-spliced transcripts to the cytoplasm, which is necessary for the formation of new virion particles [267].

A *cis-acting regulation element* (CRE) within the coding region of several picornaviruses has been described in a number of different picornaviruses. The function of the CRE probably involves the initiation of the synthesis of the negative-sense strand template RNA during virus replication [122]. The CRE has been found as in a computational survey [439] in most genera of the picornaviridae. Interestingly, it genomic location varies between genera.

The best known example in a higher organism is the stem-loop structure in the coding region of the *ASH1* gene of yeast which localizes the *ASH1* mRNA to the bud tip [56]. With the exception of a viral elements, however, the functions, as well as possible evolutionary relationships, of structured RNA motifs within ORFs remain unknown.

Some RNA molecules exhibit two competing conformations, whose equilibrium can be shifted easily by molecular events such as the binding of another molecule. This can be used to regulate of gene expression, when the two mutually exclusive alternatives correspond to an active and in-active conformation of the transcript [279]. Mechanistically, one fold of the mRNA, the repressing conformation, contains a terminator hairpin or some other structural element which conceals the translation initiation site, whereas in the alternative conformation, the non-repressing one, the gene can be expressed [148]. An early computational study concluded that RNA switches are readily accessible in evolution and are therefore probably not exceptional instances of unusual RNA behavior [108]. The use of two competing RNA conformations allows molecular events like the binding of a target metabolite by a protein to influence which of the alternative conformations the terminator or the anti-terminator is formed, hence coupling the gene expression to the concentration of the target metabolite.

The best known example of such behavior are the riboswitches [423]. These are autonomous structural elements primarily found within the 5'-UTRs of bacterial mRNAs, which, upon direct binding of small organic molecules, can trigger conformational changes, leading to an alteration of the expression for the downstream located gene. Their general architecture shows two modular units [438], a ligand-binding one, which function as a "sensor" for a small metabolite and a unit which "interprets" the signal from the "sensor" unit and interfaces to those RNA elements involved in gene expression regulation. The size of the "sensor"-unit ranges typically from 70-170 nucleotides, which is unexpectedly large compared to artificial aptamers obtained by in vitro directed evolutionary experiments. While for most riboswitches the ligand-binding domain is highly conserved among various organisms, the "interpretation" module varies strongly in sequence, structure and mechanism by which it controls the appended gene. Riboswitches and engineered allosteric ribozymes [37, 375] demonstrate impressively that RNA is indeed capable of maintaining a complex metabolic state without the help of proteins.

Riboswitches regulate several key metabolic pathways [36, 303] in bacteria including those for coenzyme B_{12}, thiamine, pyrophosphate, flavin monophosphate, S-adenosylmethionine and a couple of important amino acids. The search for additional elements is ongoing, e.g. [20, 242]. The program `Riboswitch finder` [25] utilizes consensus motifs of known elements to detect new prokaryotes riboswitches.

A recent paper by Vitreschak *et. al.* [422] applied comparative and phylogenetic analysis to vitamin B_{12}-related genes using 200 sequences from 66 bacterial genomes. They identified a highly conserved regulatory RNA structure, the B12-element a cobalamin riboswitch, which is widely distributed in 5'-UTRs of vitamin B_{12} related genes in eubacteria. Comparison of the reconstructed phylogenetic tree for the B12-element with standard trees showed both lineage- and gene-specific branches, as well as a large number of recent gene duplications and horizontal gene transfer events. A related study is reported in [298]. Comparative approaches were also used to study the L-box regulon regulating the the lysine synthesis pathway [134] and the S- and T-boxes in the methionine metabolism of Gram-positive bacteria [346]. While most riboswitches were found in bacteria, such metabolite-binding RNA domains are also present in some eukaryotic genes [385]. These findings, and the fact that riboswitches bind their effectors directly without the need of additional factors, suggest that riboswitches represent one of the oldest regulatory systems.

7 Concluding Remarks

The recent discoveries in the "modern RNA World" have made it obvious that large-scale mRNA expression profiling data can provide only a partial picture of gene expression. Most post-transcriptional events are mediated by the association of RNAs with specific proteins or macromolecular protein complexes. Comprehensive determination of the RNA targets of RNA-binding proteins is therefore likely to be important in deciphering the complex events at this level of gene regulation. Approaches to exploring the post-transcriptional RNA world with DNA microarrays are discussed e.g. in [178].

Fig. 12 gives a sketch of the probably most ancient part of the RNA-based regulation system of the eukaryotic cell: Non-coding RNAs in Eukaryotic cells seem to fall into two major groups according to their subcellular localization and thus function. The nuclear fraction mainly performs ncRNA processing and maturation. SnRNAs, snoRNAs, and scaRNAs seem to be the major players, forming the central part of the nuclear RNA regulatory network. They modify themselves as well as other ncRNAs including rRNAs, but maybe even tRNAs. Another RNA processing mechanism, RNAediting, in general does not require guide RNAs as in the case of kinetoplasts. Besides, the snRNAs act on coding hnRNA (pre-mRNA) by splicing introns. Upon export to the cytoplasm, the majority of ncRNAs is involved in protein translation. MicroRNAs regulate protein expression by translational inhibition or RNAi, 7SL RNA transports mRNA of secretory proteins to the ER (endoplasmatic reticulum). Coding and non-coding RNAs thus share a similar "life cycle" depending on their subcellular localization: regulation and maturation is performed in the nucleus, their work is done in the cytoplasm.

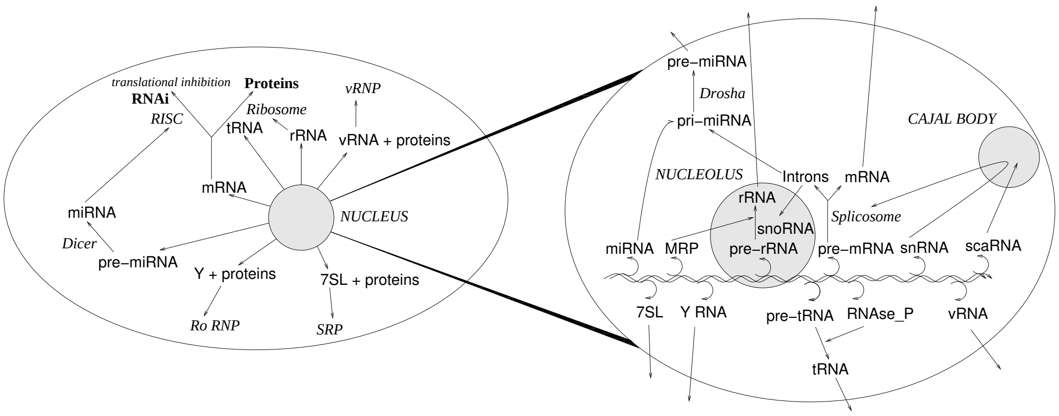


Fig. 12. Subcellular localization of ncRNAs in eukaryotic cells. The figure contains all eukaryotic ncRNA mentioned in this publication. The nuclear fraction of these regulators (right) functions in processing of ncRNAs, whereas the cytoplasmic ones (right) are involved in translation.

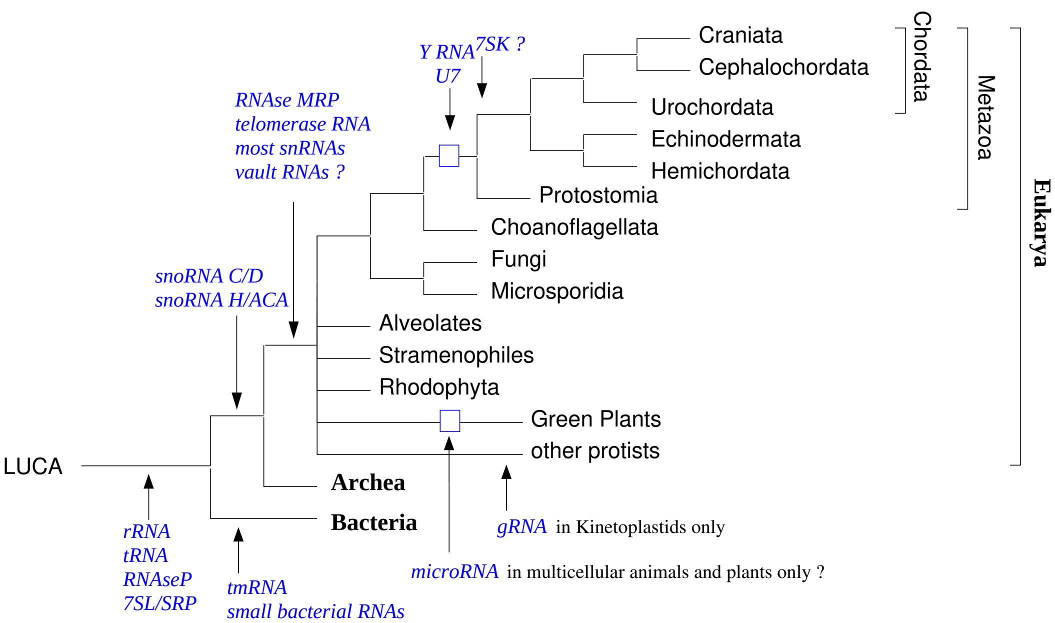


Fig. 13. Evolutionary origin of the most prominent ncRNA families.

It is commonly assumed that the primordial cell looked much more like a bacterial than a eukaryotic cell. For a discussion of the origin of the eukaryotic cell and its mitochondria we refer the reader to [225, 10]. Because of the lack of a nuclear membrane, transport mechanisms were not required. Furthermore, intronless genomes did not require a splicing-like mechanism. Instead, polycistronic transcripts of ncRNA and/or mRNA might have been processed by RNA modification and subsequent endonucleolytic cleavage. This picture is consistent with our present knowlege of the evolutionary history of the major ncRNA families summarized in Fig, 13.

While some RNAs, in particular those involved in protein synthesis predate the Last Universal Common Ancestor of all extant life forms, novel RNA families with novel — mostly regulatory — function have been invented throughout the history of life. The picture in Fig. 13 is almost certainly incomplete due to a bias in the available data which are concentrated on a small number of well-studied model organisms (mainly vertebrates, arthropods, nematods, yeast, rice, arabidopsis and bacteria). The recent discovery of a novel class of expressed ncRNAs with unknown function in *Dictyostelium discoideum* [12] and the large number of still poorly understood bacterial sRNAs, see e.g. [152, 6], suggests that quite a few ncRNA innovations in less-studied lineages could have escaped our attention so far.

The evidence compiled in this contribution indicates an explosive expansion of some ncRNA families, in particular of microRNAs, in the vertebrate lineage. Higher plants might show a similar pattern. In both cases, genome duplications are a plausible mechanism that at least contributed to expansion. Multiple dispersed copies of some snRNAs, in contrast, can be explained by the recent observation that certain retroviruses package and reverse-transcribe snRNAs [118]. Usually, this mechanism produces pseudogenes that are associated with LTRs of endogenous retroviruses. The mechanism or mechanisms that lead to duplicates of intron-encoded snoRNAs, or the processes leading to a change from intronic to exonic expression in paralogous microRNAs, on the other hand, still remain in the dark.

We close our discussion by emphasizing that it is by no means complete: topics such as the relationships of non-coding RNAs and repetitive elements (e.g. Alus) or mobile genetic elements (e.g. group II introns or endogeneous retroviruses) have been neglected here.

Acknowledgements

We thank Rolf Backofen and Daniel Gautheret for their comments on an earlier version of this manuscript. This work was supported in part by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Project No. P15893, by the German *DFG* Bioinformatics Initiative BIZ-6/1-2, and by the Austrian *Gen-AU bioinformatics integration network* sponsored by BM-BWK and BM-WA.

References

[1] M. C. Accardo, E. Giordano, S. Riccardo, F. A. Digilio, G. Iazzetti, R. A. Calogero, and M. Furia. A computational search for box C/D snoRNA

genes in the *D. melanogaster genome*. *Bioinformatics*, 20:3293–3301, 2004.

[2] A. Adai, C. Johnson, S. Mlotshwa, S. Archer-Evans, V. Manocha, V. Vance, and V. Sundaresan. Computational prediction of miRNAs in *Arabidopsis thaliana*. *Genome Res.*, 15:78–91, 2005.

[3] E. Allen, Z. Xie, A. M. Gustafson, G.-H. Sung, J. W. S. Spatafora, and J. C. Carrington. Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. *Nature Genetics*, 36:1282–1290, 2004.

[4] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J. Mol. Biol.*, 215:403–410, 1990.

[5] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped `BLAST` and `PSI-BLAST`: a new generation of protein database search programs. *Nucl. Acids Res.*, 25:3389–3402, 1997.

[6] S. Altuvia. Regulatory small RNAs: the key to coordinating global regulatory circuits. *J. Bacteriol.*, 186:6679–6680, 2004.

[7] V. Ambros, R. C. Lee, A. Lavanway, P. T. Williams, and D. Jewell. MicroRNAs and other tiny endogenous RNAs in *C. elegans*. *Curr. Biology*, 13:807–818, 2003.

[8] A. Amores, A. Force, Y. L. Yan, L. Joly, C. Amemiya, A. Fritz, R. K. Ho, J. Langeland, V. Prince, Y. L. Wang, M. Westerfield, M. Ekker, and J. H. Postlethwait. Zebrafish *Hox* clusters and vertebrate genome evolution. *Science*, 282:1711–1714, 1998.

[9] A. A. Andersen and B. Panning. Epigenetic gene regulation by noncoding RNAs. *Curr. Op. Cell Biol.*, 15:281–289, 2003.

[10] S. G. E. Andersson, O. Karlberg, B. Canbäck, and C. G. Kurland. On the origin of mitochondria: a genomics perspective. *Philos. Trans. R. Soc. Lond. B: Biol. Sci.*, 358:165–177, 2003.

[11] L. Argaman, J. Vogel, G. Bejerano, E. Wagner, H. Margalit, and S. Altuvia. Novel small RNA-encoding genes in the intergenic regions of escherichia coli. *Curr. Biol.*, 11:941–950, 2001.

[12] A. Aspegren, A. Hinas, P. Larsson, A. Larsson, and F. Söderbom. Novel non-coding RNAs in *Dictyostelium discoideum* and their expression during development. *Nucl. Acids Res.*, 32:4646–4656, 2004.

[13] P. Avner and E. Heard. X-chromosome inactivation: counting, choice, and initiation. *Nat. Rev. Genet.*, 2:59–67, 2001.

[14] T. N. Azzouz and D. Schümperli. Evolutionary conservation of the U7 small nuclear ribonucleoprotein in drosophila *melanogaster*. *RNA*, 9:1532–1541, 2003.

[15] P. Babitzke and C. Yanofsky. Reconstitution of Bacillus subtilis Trp attenuation in vitro with TRAP, the Trp RNA-binding attenuation protein. *Proc. Natl. Acad. Sci. USA*, 90:133–137, 1993.

[16] J.-P. Bachellerie, J. Cavaillé, and A. Hüttenhofer. The expanding snoRNA world. *Biochimie*, 84:775–790, 2002.

[17] V. Bafna and S. Zhang. FastR: Fast database search tool for non-coding RNA. *Proc. IEEE Comp. Systems Bioinformatics Conference*, 2004.

[18] S. Bailey, J. Wichitwechkarn, D. Johnson, B. E. Reilly, D. L. Anderson, and B. J. W. Phylogenetic analysis and secondary structure of the *Bacillus subtilis* bacteriophage RNA required for DNA packaging. *J. Biol. Chem.*, 265:22365–22370, 1990.

[19] O. Barad, E. Meiri, A. Avniel, R. Aharonov, A. Barzilai, I. Bentwich, U. Einav, S. Gilad, P. Hurban, Y. Karov, E. Lobenhofer, E. Sharon, Y. Shiboleth, M. Shtutman, Z. Bentwich, and P. Einat. MicroRNA expression detected by oligonucleotide microarrays: system establishment and expression profiling in human tissues. *Genome Res*, 14:2486–2494, 2004.

[20] J. E. Barrick, K. A. Corbino, W. C. Winkler, A. Nahvi, M. Mandal, J. Collins, M. Lee, A. Roth, N. Sudarsan, I. Jona, J. K. Wickiser, and R. R. Breaker. New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc. Natl. Acad. Sci. USA*, 101:6421–6426, 2004.

[21] D. P. Bartel and C.-Z. Chen. Micromanagers of gene expression: the potentially wide-spread influence of metazoan microRNAs. *Nature Genetics*, 5:396–400, 2004.

[22] S. Baskerville and D. P. Bartel. A ribozyme that ligates RNA to protein. *Proc. Natl. Acad. Sci. USA*, 99:9154–9159, 2002.

[23] O. Beja, E. Ullu, and S. Michaeli. Identification of a tRNA-like molecule that copurifies with the 7SL RNA of *Trypanosoma brucei. Mol. Biochem. Parasitol.*, 57:223–229, 1993.

[24] H. Ben-Shlomo, A. Levitan, N. E. Shay, I. Goncharov, and S. Michaeli. RNA editing associated with the generation of two distinct conformations of the trypanosomatid leptomonas collosoma 7SL RNA. *J. Biol. Chem.*, 274:25642–25650, 1999.

[25] P. Bengert and T. Dandekar. Riboswitch finder: a tool for identification of riboswitch RNAs. *Nucl. Acids Res.*, 32:W154–W159, 2004. Web Server Issue.

[26] Y. Bennasser, S. Y. Le, M. L. Yeung, and K. T. Jeang. HIV-1 encoded candidate micro-RNAs and their cellular targets. *Retrovirology*, 1:43, 2004. Epub.

[27] E. Berezikov, V. Guryev, J. van de Belt, E. Wienholds, R. H. A. Plasterk, and E. Cuppen. Phylogenetic shadowing and computational identification of human microRNA genes. *Cell*, 120:21–24, 2005.

[28] N. Berteaux, S. Lottin, E. Adriaenssens, F. Van Coppennolle, X. Leroy, J. Coll, T. Dugimont, and J.-J. Curgy. Hormonal regulation of H19 gene expression in prostate epithelial cells. *J. Endocrinology*, 183:69–78, 2004.

[29] B. Billoud, M. Kontic, and A. Viari. `Palingol`: a declarative programming language to describe nucleic acids' secondary structures and to scan sequence databases. *Nucl. Acids Res.*, 24:1395–1403, 1996.

[30] K. N. Bishop, R. K. Holmes, A. M. Sheehy, and M. H. Malim. APOBEC-

mediated editing of viral RNA. *Science*, 305(5684):645–645, Jul 2004.
[31] F. R. Blattner, G. Plunkett III, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, and *et al.* The complete genome sequence of *Escherichia coli* k-12. *Science*, 277:1453–1474, 1997.
[32] B. J. Blencowe. Transcription: surprising role for an elusive small nuclear RNA. *Curr. Biol.*, 12:R147–R149, 2002.
[33] S. Bonnal, C. Schaeffer, L. Creancier, S. Clamens, H. Moine, A. C. Prats, and S. Vagner. A single internal ribosome entry site containing a G quartet RNA structure drives fibroblast growth factor 2 gene expression at four alternative translation initiation codons. *J. Biol. Chem*, 278:39330–6, 2003.
[34] E. Bonnet, J. Wuyts, P. Rouzé, and Y. Van de Peer. Detection of 91 potential conserved plant microRNAs in *Arabidopsis thaliana* and *Oryza sativa* identifies important target genes. *Proc. Natl. Acad. Sci. USA*, 101:11511–11516, 2004.
[35] E. Bonnet, J. Wuyts, P. Rouzé, and Y. Van de Peer. Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, 20:2911–2917, 2004.
[36] S. Brantl. Bacterial gene regulation: from transcription attenuation to riboswitches and ribozymes. *Trends Microbiol.*, 12:473–475, 2004.
[37] R. R. Breaker. Engineered allosteric ribozymes as biosensores components. *Curr. Opin. Biotechnol.*, 13:31–39, 2002.
[38] A. Brennicke, A. Marchfelder, and S. Binder. RNA editing. *FEMS Microbiol Rev*, 23:297–316, 1999.
[39] C. J. Brown, A. Ballabio, J. L. Rupert, R. G. Lafrenière, M. Grompe, R. Tonlorenzi, and H. F. Willard. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature*, 349:38–44, 1991.
[40] J. Brown. The ribonuclease P database. *Nucl. Acids Res.*, 27:314–314, 1999.
[41] E. Buratti and F. E. Baralle. Influence of RNA secondary structure on the pre-mRNA splicing process. *Mol. Cell Biol.*, 24:10505–10514, 2004.
[42] G. Caetano-Anollés. Evolved RNA secondary structure and the rooting of the universal tree. *J. Mol. Evol.*, 54:333–345, 2002.
[43] G. Caetano-Anollés. Tracing the evolution of RNA structure in ribosomes. *Nucl. Acids Res.*, 30:2575–2587, 2002.
[44] X. Cai, C. H. Hagedorn, and B. R. Cullen. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA*, 10:1957–1966, 2004.
[45] B. Candelon, K. Guilloux, S. D. Ehrlich, and A. Sorokin. Two distinct groups of rRNA operons in the *Bacillus cereus* group. *Microbiol.*, 150:601–611, 2004.
[46] M. A. Carmell and G. J. Hannon. RNase III enzymes and the initiation of gene silencing. *Nat. Struct. Mol. Biol*, 11:214–218, 2004.

[47] S. Carranza, J. Baguñà, and M. Riutort. Origin and evolution of paralogous rRNA gene clusters within the flatworm family dugesiidae (platyhelminthes, tricladida). *J. Mol. Evol.*, 49:250–259, 1999.

[48] S. Carranza, G. Giribet, C. Ribera, J. Baguñà, and M. Riutort. Evidence that two types of 18S rDNA coexist in the genome of *Dugesia (Schmidtea) mediterranea* (platyhelminthes, turbellaria, tricladida). *Mol. Biol. Evol.*, 13:824–832, 1996.

[49] R. J. Carter, I. Dubchak, and S. R. Holbrook. A computational approach to identify genes for functional RNAs in genomic sequences. *Nucl. Acids Res.*, 29:3928–3938, 2001.

[50] J. Cavaillé, K. Buiting, M. Kiefmann, M. Lalande, C. I. Brennan, B. Horsthemke, J.-P. Bachellerie, and A. Hüttenhofer. Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. *Proc. Natl. Acad. Sci. USA*, 97:14311–14316, 2000.

[51] T. Cavalier-Smith and E. E.-Y. Chao. Phylogeny of Choanozoa, Apusozoa, and other protozoa and the early eukaryote megaevolution. *J. Mol. Evol.*, 56:540–563, 2003.

[52] S. Cawley, S. Bekiranov, H. H. Ng, P. Kapranov, E. A. Sekinger, D. Kampa, A. Piccolboni, V. Sementchenko, J. Cheng, A. J. Williams, R. Wheeler, B. Wong, J. Drenkow, M. Yamanaka, S. Patel, S. Brubaker, H. Tammana, G. Helt, K. Struhl, and T. R. Gingeras. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, 116:499–509, 2004.

[53] F. Cecconi, C. Crosio, P. Mariottini, G. Cesareni, M. Giorgi, S. Brenner, and F. Amaldi. A functional role for some fugu introns larger than the typical short ones: the example of the gene coding for ribosomal protein S7 and snoRNA U17. *Nucl. Acids Res.*, 24:3167–3172, 1996.

[54] M. Cervelli, F. Cecconi, M. Giorgi, F. Annesi, M. Oliverio, and P. Mariottini. Comparative structure analysis of vertebrate U17 small nucleolar RNA (snoRNA). *J. Mol. Evol.*, 54:166–179, 2002.

[55] M. Cervelli, M. Oliverio, A. Bellini, M. Bologna, F. Cecconi, and P. Mariottini. Structural and sequence evolution of U17 small nucleolar RNA (snoRNA) and its phylogenetic congruence in chelonians. *J. Mol. Evol.*, 57:73–84, 2003.

[56] P. Chartrand, X. H. Meng, S. R. H., and R. M. Long. Structural elements required for the localization of ASH1 mRNA and of a green fluoreszent protein reporter particle *in vivo*. *Curr. Biol.*, 9:333–336, 1999.

[57] J. H. Chen, S. Y. Le, B. Shapiro, K. M. Currey, and J. V. Maizel Jr. A computational procedure for assessing the significance of RNA secondary structure. *Comput. Appl. Biosci.*, 6:7–18, 1990.

[58] J. L. Chen, M. A. Blasco, and C. W. Greider. Secondary structure of vertebrate telomerase RNA. *Cell*, 100:503–514, 2000.

[59] J.-L. Chen and C. W. Greider. An emerging consensus for telomerase

RNA structure. *Proc. Natl. Acad. Sci. USA*, 101:14683–14684, 2004.

[60] S. Chen, E. A. Lesnik, T. A. Hall, R. Sampath, R. H. Griffey, D. Eker, and L. Blyn. A bioinformatics based approach to discover small RNA genes in the *Escherichia coli* genome. *Biosystems*, 65:157–177, 2002.

[61] C. Chureau, M. Prissette, A. Bourdet, V. Barbe, L. Cattolico, L. Jones, A. Eggen, P. Avner, and L. Duret. Comparative sequence analysis of the X-inactivation center region in mouse, human, and bovine. *Genome Res.*, 12:894–908, 2002.

[62] C. E. Clayton. Life without transcriptional control? From fly to man and back again. *EMBO J.*, 21:1881–1888, 2002.

[63] B. Clouet d'Orval, M. L. Bortolin, C. Gaspin, and J. P. Bachellerie. Box C/D RNA guides for the ribose methylation of archaeal tRNAs. The tRNATrp intron guides the formation of two ribose-methylated nucleosides in the mature tRNATrp. *Nucl. Acids Res.*, 29:4518–4529, 2001.

[64] L. J. Collins. *Lost in the RNA World*. PhD thesis, Allan Wilson Center, Massey University, Palmerston North, New Zealand, 2004.

[65] L. J. Collins, T. J. Macke, and D. Penny. Searching for ncRNAs in eukaryotic genomes: Maximizing biological input with RNAmotif. *J. Integ. Bioinf.*, #6:15p, 2004. `http://journal.imbio.de/`.

[66] L. J. Collins, V. Moulton, and D. Penny. Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP. *J. Mol. Evol.*, 51:194–204, 2000.

[67] A. Coventry, D. J. Kleitman, and B. Berger. MSARI: Multiple sequence alignments for statistical detection of RNA secondary structure. *Proc. Natl. Acad. Sci. USA*, 101:12102–12107, 2004.

[68] S. K. Crosthwaite. Circadian clocks and natural antisense RNA. *FEBS Let.*, 567:49–54, 2004.

[69] S. Crucs, S. Chatterjee, and E. R. Gavis. Overlapping but distinct RNA elements control repression and activation of nanos translation. *J. Mol. Cell*, 3:457–467, 2000.

[70] J. E. Dahlberg and E. Lund. The genes and transcription of the major small nuclear RNAs. In M. L. Birnstiel, editor, *Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles*, pages 38–70. Springer-Verlag, Berlin, 1988.

[71] E. Dalphin, P. A. Stockwell, W. P. Tate, and C. M. Brown. `TransTerm`, the translational signal database, extended to include full coding sequences and untranslated regions. *Nucl. Acids Res.*, 27:293–294, 1999.

[72] A. T. Dandjinou, N. Lévesque, S. Larose, J.-F. Lucier, S. A. Elela, and R. J. Wellinger. A phylogenetically based secondary structure for the yeast telomerase RNA. *Curr. Biol.*, 14:1148–1158, 2004.

[73] D. A. Day and M. F. Tuite. Post-transcriptional gene regulatory mechanisms in eukaryotes: an overview. *J. Endocrinol.*, 157:361–371, 1998.

[74] W. H. E. Day and H. Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classifications*, 1:7–24, 1984.

[75] E. T. Dayton, D. A. Konings, D. M. Powell, B. A. Shapiro, L. Butini, J. V. Maizel, and A. I. Dayton. Extensive sequence-specific information throughout the CAR/RRE, the target sequence of the human immunodeficiency virus type 1 Rev protein. *J. Virol.*, 66:1139–1151, 1992.

[76] J. de la Cruz and A. Vioque. A structural and functional study of plastid RNAs homologous to catalytic bacterial RNase P RNA. *Gene*, 321:47–56, 2003.

[77] V. de Turris, G. Di Leva, S. Caldarola, F. Loreni, F. Amaldi, and I. Bozzoni. TOP promoter elements control the relative ratio of intron-encoded snoRNA versus spliced mRNA biosynthesis. *J. Mol. Biol.*, 344:383–394, 2004.

[78] N. Delihas. Annotation and evolutionary relationships of a small regulatory RNA gene *micF* and its target *ompF* in *Yersinia* species. *BMC Microbiology*, 3:13 [15 pp.], 2003.

[79] P. P. Dennis, A. Omer, and T. Lowe. A guided tour: small RNA function in archaea. *Mol. Microbiol.*, 40:509–519, 2001.

[80] D. di Bernardo, T. Down, and T. Hubbard. ddbRNA: detection of conserved secondary structures in multiple alignments. *Bioinformatics*, 19:1606–1611, 2003.

[81] M. Di Giulio. The origin of the tRNA molecule: implications for the origin of protein synthesis. *J. Theor. Biol.*, 226:89–93, 2004.

[82] C. Dieterich, S. Grossmann, A. Tanzer, S. Ropcke, P. F. Arndt, P. F. Stadler, and M. Vingron. Comparative promoter region analysis powered by CORG. *BMC Genomics*, 2005. submitted.

[83] Z. Dominski, X.-c. Yang, M. Purdy, and W. F. Marzluff. Cloning and characterization of the *Drosophila* U7 small nuclear RNA. *Proc. Natl. Acad. Sci. USA*, 100:9422–9427, 2003.

[84] A. M. Domitrovich and G. R. Kunkel. Multiple, dispersed human U6 small nuclear RNA genes with varied transcriptional efficiencies. *Nucl. Acids Res.*, 31:2344–2352, 2003.

[85] W. F. Doolittle and J. R. Brown. Tempo, mode, the progenote, and the universal root. *Proc. Natl. Acad. Sci. USA*, 91:6721–6728, 1994.

[86] J. A. Doudna and T. R. Cech. The chemical repertoire of natural ribozymes. *Nature*, 418:222–228, 2002.

[87] L. Duret, F. Dorkeld, and C. Gautier. Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression. *Nucl. Acids Res.*, 21:2315–2322, 1993.

[88] S. R. Eddy. Non-coding RNA genes and the modern RNA world. *Nature Genetics*, 2:919–929, 2001.

[89] S. R. Eddy. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics*, 3:18, 2002.

[90] S. Edvardsson, P. P. Gardner, A. M. Poole, M. D. Hendy, D. Penny, and V. Moulton. A search for H/ACA snornas in yeast using MFE secondary

structure prediction. *Bioinformatics*, 19:865–873, 2003.

[91] M. Eigen, B. F. Lindemann, M. Tietze, R. Winkler-Oswatitsch, A. W. M. Dress, and A. von Haeseler. How old is the genetic code? Statistical geometry of tRNA provides an answer. *Science*, 244:673–679, 1989.

[92] M. Eigen and R. Winkler-Oswatitsch. Transfer-RNA, an early gene? *Naturwissenschaften*, 68:282–292, 1981.

[93] W. Elbashir, S. Lendeckel and T. Tuschl. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev.*, 15:188–200, 2001.

[94] A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks. MicroRNA targets in Drosophila. *Genome Biology*, 5(1), 2003. Article R1.

[95] C. A. Enright, E. S. Maxwell, G. L. Elicieri, and B. Sollner-Webb. 5'ETS rRNA processing facilitated by by four small RNAs: U14, E3, U17, and U3. *RNA*, 2:1094–1099, 1996.

[96] V. Erdmann, M. Barciszewska, A. Hochberg, N. de Groot, and J. Barciszewski. Regulatory RNAs. *Cell. Mol. Life Sci.*, 58:960–977, 2001.

[97] V. Erdmann, M. Szymański, A. Hochberg, N. de Groot, and J. Barciszewski. Collection of mRNA-like non-coding RNAs. *Nucleic Acids Res.*, 27:192–195, 1999.

[98] V. A. Erdmann, M. Szymański, A. Hochberg, N. de Groot, and J. Barciszewski. Non-coding, mRNA-like RNAs database Y2K. *Nucl. Acids Res.*, 28:197–2000, 2000.

[99] H. Escriva, L. Manzon, J. Youson, and V. Laudet. Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. *Mol. Biol. Evol.*, 19:1440–1450, 2002.

[100] A. M. Estévez and L. Simpson. Uridine insertion/deletion RNA editing in trypanosome mitochondria — a review. *Gene*, 240:247–260, 1999.

[101] E. Fabre, B. Dujon, and G. Richard. Transcription and nuclear transport of CAG/CTG trinucleotide repeats in yeast. *Nucl. Acids Res.*, 30:3540–3547, 2002.

[102] D. Fagegaltier, A. Lescure, R. Walczak, P. Carbon, and A. Krol. Structural analysis of new local features in SECIS RNA hairpins. *Nucl. Acids Res.*, 28(14):2679–2689, 2000.

[103] A. D. Farris, G. Koelsch, G. J. Pruijn, W. J. van Venrooij, and H. J. B. Conserved features of Y RNAs revealed by automated phylogenetic secondary structure analysis. *Nucl. Ac. Res.*, 27:1070–8, Feb 1999.

[104] G. Fayat, F. J. Mayaux, C. Sacerdot, M. Fromant, M. Springer, M. Grunberg-Manago, and S. Blanquet. Escherichia coli phenylalanyl-tRNA synthetase operon region. Evidence for an attenuation mechanism. Identification of the gene for the ribosomal protein L20. *J. Mol. Biol.*, 171:239–261, 1983.

[105] B. Felden, C. Massire, E. Westhof, J. F. Atkins, and R. F. Gesteland. Phylogenetic analysis of tmRNA genes within a bacterial subgroup reveals a specific structural signature. *Nucl. Acids Res.*, 29:1602–1607, 2001.

[106] M. G. Ferreira and J. P. Miller, Kyle M.and Cooper. Indecent exposure: When telomeres become uncapped. *Mol. Cell*, 13:7–18, 2004.

[107] V. Filippov, V. Solovyev, M. Filippova, and S. Gill. A novel type of RNase III family proteins in eukaryotes. *Gene*, 245:213–221, 2000.

[108] C. Flamm, I. L. Hofacker, S. Maurer-Stroh, P. F. Stadler, and M. Zehl. Design of multi-stable RNA molecules. *RNA*, 7:254–265, 2000.

[109] A. Franke and B. Baker. Dosage compensation rox! *Curr. Opin. Cell Biol.*, 12:351–354, 2000.

[110] S. J. Freeland, R. D. Knight, and L. F. Landweber. Do proteins predate DNA? *Science*, 286:690–692, 1999.

[111] F. E. Frenkel, M. B. Chaley, E. V. Korotkov, and K. G. Skryabin. Evolution of tRNA-like sequences and genome variability. *Gene*, 335:57–71, 2004.

[112] P. P. Gardner and R. Giegerich. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatic*, 5:140, 2004.

[113] C. Gaudin, X. Zhou, K. P. Williams, and B. Felden. Two-piece tmRNA in cyanobacteria and its structural analysis. *Nucl. Acids Res.*, 30:2018–2024, 2002.

[114] D. Gautheret and A. Lambert. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. *J. Mol. Biol.*, 313:1003–1011, 2001.

[115] F. Gebauer and M. W. Hentze. Molecular mechanisms of translational control. *Nat. Rev. Mol. Cell Biol.*, 5:827–835, October 2004.

[116] R. F. Gesteland and J. F. Atkins, editors. *The RNA World*. Cold Spring Harbor Laboratory Press, Plainview, NY, 1993.

[117] W. Gilbert. The RNA world. *Nature*, 319:618, 1986.

[118] K. E. Giles, M. Caputi, and K. Beemon. Packaging and reverse transcription of snRNAs by retroviruses may generate pseudogenes. *RNA*, 10:299–307, 2004.

[119] J. Gilley and M. Fried. Evolution of U24 and U36 snoRNAs encoded within introns of vertebrate *rpL7a* gene homologs: Unique features of mammalian U36 variants. *DNA Cell Biol.*, 17:591–602, 1998.

[120] G. M. Gilmartin, F. Schaufele, G. Schaffner, and M. L. Birnstiel. Functional analysis of the sea urchin U7 small nuclear RNA. *Mol. Cell Biol.*, 8:1076–1084, 1988.

[121] I. L. Gonzalez and J. E. Sylvester. Human rDNA: Evolutionary patterns within the genes and tandem arrays derived from multiple chromosomes. *Genomics*, 73:255–263, 2001.

[122] I. G. Goodfellow, D. Kerrigan, and D. J. Evans. Structure and functional analysis of the poliovirus cis-acting replication element (CRE). *RNA*, 9:124–137, 2003.

[123] J. Gorodkin, L. J. Heyer, and G. D. Stormo. Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucl. Acids Res.*, 25(18):3724–3732, 1997.

[124] J. Gorodkin, B. Knudsen, C. Zwieb, and T. Samuelsson. SRPDB (signal recognition particle database). *Nucleic Acids Res.*, 29:169–170, 2001.

[125] J. Gorodkin, S. L. Stricklin, and G. D. Stormo. Discovering common stem-loop motifs in unaligned RNA sequences. *Nucl. Acids Res.*, 29(10):2135–2144, 2001.

[126] J. M. Gott and R. B. Emeson. Functions and mechanisms of RNA editing. *Annu. Rev. Genet*, 34:499–531, 2000.

[127] S. Gottesman. The small RNA regulators of *Escherichia coli*: roles and mechanisms. *Annu. Rev. Microbiol.*, 58:303–328, 2004.

[128] S. G. Gottlob-McHugh, M. Levesque, K. MacKenzie, M. Olson, O. Yarosh, and D. A. Johnson. Organization of the 5S rRNA genes in the soybean *Glycine max (L.) Merrill* and conservation of the 5S rDNA repeat structure in higher plants. *Genome*, 33:486–494, 1990.

[129] S. Gräf, D. Strothmann, S. Kurtz, and G. Steger. HyPaLib: a database of RNAs and RNA structural elements defined by hybrid patterns. *Nucl. Acids. Res.*, 29:196–198, 2001.

[130] S. Griffiths-Jones. The microRNA Registry. *Nucl. Acids Res.*, 32:D109–D111, 2004. Database issue.

[131] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. Eddy. Rfam: an RNA family database. *Nucl. Acids Res.*, 31:439–441, 2003.

[132] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy, and A. Bateman. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res*, 33 Database Issue:121–124, 2005.

[133] I. Grummt. Life on a planet of its own: regulation of RNA polymerase i transcription in the nucleolus. *Genes Dev.*, 17:1691–1702, 2003.

[134] F. J. Grundy, S. C. Lehman, and T. M. Henkin. The L box regulon: lysine sensing by leader RNAs of bacterial lysine biosynthesis genes. *Proc. Natl. Acad. Sci. USA*, 100:12057–12062, 2003.

[135] W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. I. neutral networks. *Monath. Chem.*, 127:355–374, 1996.

[136] W. Grüner, R. Giegerich, D. Strothmann, C. Reidys, J. Weber, I. L. Hofacker, P. F. Stadler, and P. Schuster. Analysis of RNA sequence structure maps by exhaustive enumeration. II. structures of neutral networks and shape space covering. *Monath. Chem.*, 127:375–389, 1996.

[137] P. Gueneau de Novoa and K. P. Williams. The tmRNA website: reductive evolution of tmRNA in plastids and other endosymbionts. *Nucl. Acids Res.*, 32:D104–D108, 2004. Database issue.

[138] P. Guo. Structure and function of phi29 hexameric RNA that drives the viral DNA packaging motor: review. *Prog. Nucleic Acid Res. Mol. Biol.*, 72:415–472, 2002.

[139] H.-C. Gürsoy, D. Koper, and B.-J. Benecke. The vertebrate 7S K RNA separates hagfish (*Myxine glutinosa*) and lamprey (*Lampetra fluviatilis*). *J. Mol. Evol.*, 50:456–464, 2000.

[140] A. M. Gustafson, E. Allen, S. Givan, D. Smith, J. C. Carrington, and K. D. Kasschau. ASRP: the Arabidopsis Small RNA Project Database. *Nucl. Acids Res.*, 33:D637–D640, 2005.

[141] E. S. Haas, A. B. Banta, J. K. Harris, N. R. P. Pace, and J. W. Brown. Structure and evolution of ribonuclease P RNA in Gram-positive bacteria. *Nucl. Acids Res.*, 24:4775–4782, 1996.

[142] J. Hackermüller, N.-C. Meisner, M. Auer, M. Jaritz, and P. F. Stadler. The effect of RNA secondary structures on RNA-ligand binding and the modifier RNA mechanism: A quantitative model. *Gene*, 2005. doi:10.1016/j.gene.2004.11.043.

[143] P. W. Haebel, S. Gutmann, and N. Ban. Dial tm for rescue: tmRNA engages ribosomes stalled on defective mRNAs. *Curr. Op. Struct. Biol.*, 14:58–65, 2004.

[144] G. J. Hannon. RNA interference. *Nature*, 418:244–251, 2002.

[145] R. J. Harris and D. Elder. Ribozyme relationships: The hammerhead, hepatitis delta, and hairpin ribozymes have a common origin. *J. Mol. Evol.*, 51:182–184, 2000.

[146] E. Hartmann and R. K. Hartmann. The enigma of ribonuclease P evolution. *Trends Genet.*, 19:561–569, 2003.

[147] J. H. Havgaard, R. Lingsø, G. D. Stormo, and J. Gorodkin. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, 2005. Epub Jan 18 2005.

[148] T. M. Henkin and C. Yanofsky. Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcribtion termination/antitermination decision. *BioEssays*, 24:700–707, 2002.

[149] A. K. Henras, C. Dez, and Y. Henry. RNA structure and function in C/D and H/ACA s(no)RNAs. *Curr. Op. Struct. Biol.*, 14:335–343, 2004.

[150] M. W. Hentze and L. C. Kühn. Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress. *Proc. Natl. Acad. Sci. USA*, 93:8175–8182, 1996.

[151] N. Hernandez. Small nuclear RNA genes: a model system to study fundamental mechanisms of transcription. *J. Biol. Chem.*, 276:26733–26736, 2001.

[152] R. Hershberg, S. Altuvia, and H. Margalit. A survey of small RNA-encoding genes in *Escherichia coli*. *Nucl. Acids Res.*, 31:1813–1820, 2003.

[153] J. Hesketh. 3'-untranslated regions are important in mRNA localization and translation: lessons from selenium and metallothionein. *Biochem Soc Trans.*, 32:990–993, 2004.

[154] P. G. Higgs, D. Jameson, H. Jow, and M. Rattray. The evolution of tRNA-leu genes in animal mitochondrial genomes. *J. Mol. Evol.*, pages 435–445, 2003.

[155] D. M. Hillis and M. T. Dixon. Ribosomal DNA: molecular evolution and phylogenetic inference. *Q. Rev. Biol.*, 66:411–453, 1991.

[156] S. Hinz and H. U. Göringer. The guide RNA database (3.0). *Nucl. Acids Res.*, 27:168, 1999.

[157] O. Hobert. Common logic of transcription factor and microRNA action. *Trends Biochem. Sci.*, 29:462–468, 2004.

[158] M. Höchsmann, T. Töller, R. Giegerich, and S. Kurtz. Local similarity in RNA secondary structures. In *Proc of the Computational Systems Bioinformatics Conference, Stanford, CA, August 2003 (CSB 2003)*, pages 159–168, 2003.

[159] I. L. Hofacker. Vienna RNA secondary structure server. *Nucl. Acids Res.*, 31:3429–3431, 2003.

[160] I. L. Hofacker, S. H. F. Bernhart, and P. F. Stadler. Alignment of RNA base pairing probability matrices. *Bioinformatics*, 20:2222–2227, 2004.

[161] I. L. Hofacker, M. Fekete, C. Flamm, M. A. Huynen, S. Rauscher, P. E. Stolorz, and P. F. Stadler. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucl. Acids Res.*, 26:3825–3836, 1998. Santa Fe Institute Preprint 98-03-020.

[162] I. L. Hofacker, M. Fekete, and P. F. Stadler. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, 319:1059–1066, 2002.

[163] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.

[164] I. L. Hofacker and P. F. Stadler. Automatic detection of conserved base pairing patterns in RNA virus genomes. *Comp. & Chem.*, 23:401–414, 1999. Santa Fe Institute preprint 98-06-058.

[165] I. L. Hofacker, R. Stocsits, and P. F. Stadler. Conserved RNA secondary structures in viral genomes: A survey. *Bioinformatics*, 20:1495–1499, 2004.

[166] P. W. H. Holland, J. Garcia-Fernández, N. A. Williams, and A. Sidow. Gene duplication and the origins of vertebrate development. *Development*, (Suppl.):125–133, 1994.

[167] I. Holmes. A probabilistic model for the evolution of RNA structure. *BMC Bioinformatics*, 5:166, 2004.

[168] M. Hong and L. Simpson. Genomic organization of *Trypanosoma brucei* kinetoplast DNA minicircles. *Prostist*, 154:265–279, 2003.

[169] A. K. Hopper and E. M. Phizicky. tRNA transfers to the limelight. *Genes Devel.*, 17:162–180, 2003.

[170] Y. Hu. GPRM: a genetic programming approach to finding common RNA secondary structure elements. *Nucl. Acids Res.*, 31(13):3446–3449, 2003.

[171] Y.-J. Hu. Prediction of consensus structural motifs in a family of coregulated RNA sequences. *Nucl. Acids Res.*, 30:3886–3893, 2002.

[172] Z. P. Huang, H. Zhou, D. Liang, and L. H. Qu. Different expression strategy: multiple intronic gene clusters of box H/ACA snoRNA in *Drosophila melanogaster*. *J. Mol. Biol.*, 341:669–683, 2004.

[173] C. Hudelot, V. Gowri-Shankar, H. Jow, M. Rattray, and P. G. Higgs.

RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences. *Mol. Phylogenet. Evol.*, 28:241–252, 2003.

[174] I. Huez, L. Créancier, S. Audigier, M. Gensac, A. Prats, and H. Prats. Two independent internal ribosome entry sites are involved in translation initiation of vascular endothelial growth factor mRNA. *Mol. Cell. Biol.*, 18:6178–6190, 1998.

[175] A. Hüttenhofer, M. Kiefmann, S. Neier-Ewert, J. O'Brien, H. Lehrach, J. Bachellerie, and J. Brosius. Rnomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.*, 20:2943–2953, 2001.

[176] M. A. Huynen, P. F. Stadler, and W. Fontana. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci. (USA)*, 93:397–401, 1996.

[177] M. Illangasekare and M. Yarus. A tiny RNA that catalyzes both aminoacyl-RNA and peptidyl-RNA synthesis. *RNA*, 5:1482–1489, 1999.

[178] V. R. Iyer. Exploring the post-transcriptional RNA world with DNA microarrays. *Trends Biotech.*, 22:498–500, 2004.

[179] Y. Jacob, E. Seif, P.-O. Paquet, and F. B. Lang. Loss of the mRNA-like region in mitochondrial tmRNAs of jakobids. *RNA*, 10:605–614, 2004.

[180] V. R. Jadhav and M. Yarus. Coenzymes as coribozymes. *Biochimie*, 84:877–888, 2002.

[181] B. E. Jády, E. Bertrand, and T. Kiss. Human telomerase RNA and box H/ACA scaRNAs share a common Cajal body specific localization signal. *J. Cell Biol.*, 164:647–652, 2004.

[182] B. E. Jády and T. Kiss. A small nucleolar guide RNA functions both in 2'-O-methylation and pseudouridylation of U5 spliceosomal RNA. *EMBO J.*, 20:541–551, 2001.

[183] D. Jameson, A. P. Gibson, C. Hudelot, and P. G. Higgs. OGRe: a relational database for comparative analysis of mitochondrial genomes. *Nucl. Acids Res.*, 31:202–206, 2003.

[184] S. K. Jang, H. G. Krausslich, M. J. Nicklin, G. M. Duke, A. C. Palmenberg, and E. Wimmer. A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation. *J. Virol*, 62:2636–43, 1988.

[185] N. Jareborg, E. Birney, and R. Durbin. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Research*, 9:815–824, 1999.

[186] A. M. Jeffares, Daniel C. andPoole and D. Penny. Relics from the RNA world. *J. Mol. Evol.*, 46:18–36, 1998.

[187] Y. Ji, X. Xing, and G. D. Stormo. A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics*, 20:1591–1602, 2004.

[188] W. K. Johnston, P. J. Unrau, M. J. Lawrence, M. E. Glasner, and D. P. Bartel. RNA-catalyzed RNA polymerization: Accurate and general RNA-templated primer extension. *Science*, 292:1319–1325, 2001.

[189] M. W. Jones-Roades and D. P. Bartel. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol. Cell*, 14:787–799, 2004.

[190] H. Jow, C. Hudelot, M. Rattray, and P. G. Higgs. Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. *Mol. Biol. Evol.*, 19:1591–1601, 2002.

[191] G. F. Joyce. The antiquity of RNA-based evolution. *Nature*, 418:214–221, 2002.

[192] G. F. Joyce. Directed evolution of nucleic acid enzymes. *Annu. Rev. Biochem.*, 73:791–836, 2004.

[193] V. Juan, C. Crain, and C. Wilson. Evidence for evolutionarily conserved secondary structure in the H19 tumour suppressor RNA. *Nucl. Acids Res.*, 28:1221–1227, 2000.

[194] D. Kampa, J. Cheng, P. Kapranov, M. Yamanaka, S. Brubaker, S. Cawley, J. Drenkow, A. Piccolboni, S. Bekiranov, G. Helt, H. Tammana, and T. R. Gingeras. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.*, 14:331–342, 2004.

[195] L. Katz and C. B. Burge. Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Res.*, 13:2042–2051, 2003.

[196] R. J. Keenan, D. M. Freyman, R. M. Stroud, and P. Walter. The signal recognition particle. *Annu. Rev. Biochem.*, 70:755–775, 2001.

[197] K. C. Keiler, L. Shapiro, and K. P. Williams. tmRNAs that encode proteolysis-inducing tags are found in all known bacterial genomes: a two-piece tmRNA functions in caulobacter. *Proc. Natl. Acad. Sci. USA*, 97:7778–7783, 2000.

[198] C. Kelleher, M. T. Teixeira, K. Förstemann, and J. Lingner. Telomerase: biochemical considerations for enzyme and substrate. *Trends Biochem. Sci.*, 27:572–579, 2002.

[199] P. Khaitovich, A. S. Mankin, R. Green, L. Lancaster, and H. F. Noller. Characterization of functionally active subribosomal particles from *Thermus aquaticus*. *Proc. Natl. Acad. Sci. U.S.A.*, 96:85–90, 1999.

[200] C. A. Kidner and R. A. Martienssen. The developmental role of microRNA in plants. *Curr. Op. Plant Biol.*, 8:38–44, 2005.

[201] T. Kiss. Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *EMBO J.*, 20:3617–3622, 2001.

[202] R. J. Klein and S. R. Eddy. RSEARCH: Finding homologs of single structured RNA sequences. *BMC Bioinformatics*, 4(44):1471–2105, 2003.

[203] R. J. Klein, Z. Misulovin, and S. R. Eddy. Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc. Natl. Acad. Sci. USA*, 99:7542–7547, 2002.

[204] B. Knudsen and J. Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucl. Acids Res.*, 31:3423–3428,

2003.

[205] B. Knudsen and J. J. Hein. Using stochastic context free grammars and molecular evolution to predict RNA secondary structure. *Bioinformatics*, 15:446–454, 1999.

[206] J. Kohtz and G. Fishell. Developmental regulation of EVF-1, a novel non-coding RNA transcribed upstream of the mouse Dlx6 gene. *Gene Expr. Patterns*, 4:407–412, 2004.

[207] Y. Komatsu. Regulation of ribozyme activity with short oligonucleotides. *Biol. Pharma. Bull.*, 27:457–462, 2004.

[208] J. Konecny, M. Schöninger, I. L. Hofacker, M.-D. Weitze, and G. L. Hofacker. Concurrent neutral evolution of mRNA secondary structures and encoded proteins. *J. Mol. Evol.*, 50:238–242, 2000.

[209] D. Koper-Emde. *Phylogenetische Heterogenitt der 7S-RNAs von Eukaryonten*. PhD thesis, Univ. Bochum, 2004.

[210] D. Korencic, I. Ahel, J. Schelert, M. Sacher, B. Ruan, C. Stathopoulos, P. Blum, M. Ibba, and D. Söll. A freestanding proofreading domain is required for protein synthesis quality control in archaea. *Proc. Natl. Acad. Sci. USA*, 101:10260–10265, 2004.

[211] M. Kozak. The scanning model for translation: an update. *J. Cell. Biol.*, 108:229–241, February 1989.

[212] M. Kozak. An analysis of vertebrate mRNA sequences: intimations of translational control. *J. of Cell Biol.*, 115(4):887–903, 1991.

[213] M. Kozak. New ways of initiating translation in eukaryotes? *Mol. and Cellular Biol.*, 21(6):1899–1907, 2001.

[214] A. S. Krasilnikov, Y. Xiao, T. Pan, and A. Mondragón. Basis for structural diversity in homologous RNAs. *Science*, 306:104–107, 2004.

[215] A. Krol. Evolutionarily different RNA motifs and RNA-protein complexes to achieve selenoprotein synthesis. *Biochimie*, 84:765–774, 2002.

[216] K. Y. Kwek, S. Murphy, A. Furger, B. Thomas, W. O'Gorman, H. Kimura, N. J. Proudfoot, and A. Akoulitchev. U1 snRNA associates with TFIIH and regulates transcriptional initiation. *Nat. Struct. Biol.*, 9:800–805, 2002.

[217] D. Lafontaine and D. Tollervey. Birth of the snoRNPs: the evolution of the modification-guide snoRNAs. *Trends Biochem. Sci.*, 23:383–388, 2002.

[218] M.-J. Laforest, C. E. Bullerwell, L. Forget, and F. B. Lang. Origin, evolution, and mechanism of 5'tRNA editing in chytridiomycete fungi. *RNA*, 10:1191–1199, 2004.

[219] M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl. Identification of novel genes coding for small expressed RNAs. *Science*, 294:853–857, 2001.

[220] M. Lagos-Quintana, R. Rauhut, J. Meyer, A. Borkhardt, and T. Tuschl. New microRNAs from mouse and human. *RNA*, 9:175–179, 2003.

[221] E. C. Lai, P. Tomancak, R. W. Williams, and G. M. Rubin. Computational identification of Drosophila microRNA genes. *Genome Biol.*,

4:R42, 2003.

[222] J. Landry, P. Medstrand, and D. L. Mager. Repetitive elements in the 5' untranslated region of a human zinc-finger gene modulate transcription and translation efficiency. *Genomics*, 76(1-3), August 2001.

[223] L. F. Landweber. The evolution of RNA editing in kinetoplastid protozoa. *Biosystems*, 28:41–45, 1992.

[224] L. F. Landweber and W. Gilbert. Phylogenetic analysis of RNA editing: a primitive genetic phenomenon. *Proc. Natl. Acad. Sci. USA*, 91:918–921, 1994.

[225] B. F. Lang, M. W. Gray, and G. Burger. Mitochondrial genome evolution and the origin of eukaryotes. *Annu. Rev. Genet.*, 33:351–397, 1999.

[226] A. Lariza, W. Makalowski, G. Pesole, and C. Saccone. Evolutionary dynamics of mammalian mRNA untranslated regions by comparative analysis of orthologous human, artiodactyl and rodent gene pairs. *Computers and Chemistry*, 26:479–490, 2002.

[227] D. Laslett, B. Canback, and S. Andersson. `BRUCE`: a program for the detection of transfer-messenger RNA genes in nucleotide sequences. *Nucl. Acids Res.*, 30:3449–3453, 2002.

[228] N. C. Lau, L. P. Lim, E. G. Weinstein, and D. P. Bartel. An abundant class of tiny RNAs with probable regulatory roles in *caenorhabditis elegans*. *Science*, 294:858–862, 2001.

[229] G. Lavorgna, D. Dahary, B. Lehner, R. Sorek, C. M. Sanderson, and G. Casari. In search of antisense. *Trends Biochem. Sci.*, 29, 2004.

[230] S. Le and J. V. Maizel. A common RNA structural motif involved in the internal initiation of translation of cellular mRNAs. *Nucl. Acids Res.*, 25(2):362–369, 1997.

[231] S. V. Le, J. H. Chen, K. M. Currey, and J. V. Maizel Jr. A program for predicting significant RNA secondary structures. *Comput. Appl. Biosci.*, 4:153–159, 1988.

[232] S. Y. Le, J. H. Chen, D. Konings, and J. V. Maizel Jr. Discovering well-ordered folding patterns in nucleotide sequences. *Bioinformatics*, 19:354–361, 2003.

[233] S. Y. Le, K. Zhang, and J. V. Maizel Jr. RNA molecules with structure dependent functions are uniquely folded. *Nucl. Acids Res.*, 30:3574–3582, 2002.

[234] D. M. LeCuyer, Karen A.and Crothers. Kinetics of an RNA conformational switch. *Proc. Natl. Acad. Sci. USA*, 91:3373–3377, 1994.

[235] J. T. Lee, L. S. Davidow, and D. Warshawsky. *Tsix*, a gene antisense to *Xist* at the X-inactivation centre. *Nat. Genet.*, 21:400–404, 1999.

[236] N. Lee, Y. Bessho, K. Wei, J. W. Szostak, and H. Suga. Ribozyme-catalyzed tRNA aminoacylation. *Nat. Struct. Biol*, 7:28–33, 2000.

[237] R. Lee and V. Ambros. An extensive class of small RNAs in *caenorhabditis elegans*. *Science*, 294:862–864, 2001.

[238] Y. Lee, C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Rådmark, S. Kim, and V. N. Kim. The nuclear RNase III Drosha

initiates microRNA processing. *Nature*, 425:415–419, 2003.

[239] Y. Lee, K. Jeon, J. T. Lee, S. Kim, and V. N. Kim. MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J.*, 21:4663–4670, 2002.

[240] Y. Lee, M. Kim, J. Han, K. H. Yeom, S. Lee, S. H. Baek, and V. N. Kim. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J*, 23:4051–4060, 2004.

[241] M. Legendre, A. Lambert, and D. Gautheret. Profile-based detection of microRNA precursors in animal genomes. *Bioinformatics*, 2005. Epub ahead of print.

[242] E. A. Lesnik, G. B. Fogel, D. Weekes, T. J. Henderson, H. B. Levene, R. Sampath, and D. J. Ecker. Identification of conserved regulatory RNA structures in prokaryotic metabolic pathway genes. *Biosystems*, 2005. doi:10.1016/j.biosystems.2004.11.002.

[243] K. Li and R. S. Williams. Cloning and characterization of three new murine genes encoding short homologues of RNAse P RNA. *J. Biol. Chem.*, 270:25281–25285, 1995.

[244] Y. Li and S. Altman. In search of RNase P RNA from microbial genomes. *RNA*, 10:1533–1540, 2004.

[245] X. H. Liang, Y. X. Xu, and S. Michaeli. The spliced-leader associated RNA is a trypanosome-specific sn(o)RNA that has the potential to guide pseudouridine formation on SL RNA. *RNA*, 8:237–246, 2002.

[246] D. Liao. Concerted evolution: Molecular mechanisms and biological implications. *Am. J. Hum. Genet.*, 64:24–30, 1999.

[247] D. Liao, T. Pavelitz, J. R. Kidd, K. K. Kidd, and A. M. Weiner. Concerted evolution of the tandemly repeated genes encoding human U2 snRNA (the RNU2 locus) involves rapid intrachromosomal homogenization and rare interchromosomal gene conversion. *EMBO J.*, 16:588–598, 1997.

[248] D. M. J. Lilley. The origins of RNA catalysis in ribozymes. *Trends Biochem. Sci.*, 28:495–501, 2003.

[249] J. Lin, H. Ly, A. Hussain, M. Abraham, S. Pearl, Y. Tzfati, and E. H. Parslow, Tristram G.and Blackburn. A universal telomerase RNA core structure includes structured motifs required for binding the telomerase reverse transcriptase protein. *Proc. Natl. Acad. Sci. USA*, 101:14713–14718, 2004.

[250] J. Lingner, J. P. Cooper, and T. R. Cech. Telomerase and DNA end replication: no longer a lagging strand problem? *Science*, 269:1533–1534, 1995.

[251] D. J. Lipman. Making (anti)sense of non-coding sequence conservation. *Nucl. Acids Res.*, 25(18):3580–3583, 1997.

[252] R. D. Little and B. C. Braaten. Genomic organization of human 5 S rDNA and sequence of one tandem repeat. *Genomics*, 4:376–383, 1989.

[253] C. Liu, B. Bai, G. Skogerbø, L. Cai, W. Deng, Y. Zhang, D. Bu, Y. Zhao, and R. Chen. NONCODE: an integrated knowledge database of non-

coding RNAs. *Nucl. Acids Res.*, 33:D112–D115, 2005. Database issue.

[254] T. Lowe and S. Eddy. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucl. Acids Res.*, 25:955–964, 1997.

[255] T. M. Lowe and S. R. Eddy. A computational screen for methylation guide snoRNAs in yeast. *Science*, 19:1168–1171, 1999.

[256] S. Lu and B. R. Cullen. Adenovirus VA1 noncoding RNA can inhibit small interfering RNA and MicroRNA biogenesis. *J. Virol.*, 78:12868–12876, 2004.

[257] D. J. Luciano, H. Mirsky, N. J. Vendetti, and S. Maas. RNA editing of a miRNA precursor. *RNA*, 10:1174–1177, 2004.

[258] R. Lück, S. Gräf, and G. Steger. Construct: A tool for thermodynamic controlled prediction of conserved secondary structure. *Nucl. Acids Res.*, 27:4208–4217, 1999.

[259] R. Lück, G. Steger, and D. Riesner. Thermodynamic prediction of conserved secondary structure: Application to the RRE element of HIV, the tRNA-like element of CMV, and the mRNA of prion protein. *J. Mol. Biol.*, 258:813–826, 1996.

[260] N. F. Lue. Adding to the ends: what makes telomerase processive and how important is it? *Bioessays*, 26:955–962, 2004.

[261] M. Lynch and J. S. Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290:1151–1155, 2000.

[262] S. Maas and A. Rich. Changing genetic information through RNA editing. *BioEssays*, 22:790–802, 2000.

[263] G. C. MacIntosh, C. Wilkerson, and P. J. Green. Identification and analysis of *Arabidopsis* expressed sequence tags characteristic of non-coding RNAs. *Plant Physiol.*, 127:765–776, 2001.

[264] T. J. Macke, D. J. Ecker, R. R. Gutell, D. Gautheret, D. A. Case, and R. Sampath. RNAMotif, an RNA secondary structure definition and search algorithm. *Nucl. Acids Res.*, 29(22):4724–4735, 2001.

[265] B. E. H. Maden. The numerous modified nucleotides in eukaryotic ribosomal RNA. *Prog. Nucl. Acid Res. Mol. Biol.*, 39:241–303, 1990.

[266] B. Maidak, J. Cole, T. Lilburn, C. Parker Jr., P. Saxman, R. Farris, G. Garrity, G. Olsen, T. Schmidt, and J. Tiedje. The RDP-II (ribosomal database project). *Nucl. Acids Res.*, 29:173–174, 2001.

[267] M. H. Malim, J. Hauber, S. Y. Le, J. V. Maizel, and B. Cullen. The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. *Nature*, 338:254–257, 1989.

[268] J. M. Mallatt, J. R. Garey, and J. W. Shultz. Ecdysozoan phylogeny and bayesian inference: first use of nearly complete 28S and 18S rRNA gene sequences to classify the arthropods and their kin. *Mol. Phylogenet. Evol.*, 31:178–191, 2004.

[269] L. M. Márquez, D. J. Miller, J. B. MacKenzie, and M. J. H. van Oppen. Pseudogenes contribute to the extreme diversity of nuclear ribosomal

DNA in the hard coral *Acropora*. *Mol. Biol. Evol.*, 20:1077–1086, 2003.

[270] Y. Martineau, C. Le Bec, L. Monbrun, V. Allo, I. M. Chiu, O. Danos, H. Moine, H. Prats, and A. C. Prats. Internal ribosome entry site structural motifs conserved among mammalian fibroblast growth factor 1 alternatively spliced mRNAs. *Mol. Cell. Biol*, 24:7622–35, 2004.

[271] D. H. Mathews and D. H. Turner. Dynalign: An algorithm for finding secondary structures common to two RNA sequences. *J. Mol. Biol.*, 317:191–203, 2002.

[272] M. B. Mathews. Structure, function, and evolution of adenovirus virus-associated RNAs. *Curr. Top. Microbiol. Immunol.*, 199:173–187, 1995.

[273] J. S. Mattick. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays*, 25:930–939, 2003.

[274] J. S. Mattick. RNA regulation: a new genetics? *Nature Genetics*, 5:316–323, 2004.

[275] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[276] J. P. McCutcheon and S. R. Eddy. Computational identification of non-coding RNAs in *Saccharomyces cerevisiae* by comparative genomics. *Nucl. Acids Res.*, 31:4119–4128, 2003.

[277] N.-C. Meisner, J. Hackermüller, V. Uhl, A. Aszódi, M. Jaritz, and M. Auer. mRNA openers and closers: A methodology to modulate AU-rich element controlled mRNA stability by a molecular switch in mRNA conformation. *Chembiochem.*, 5:1432–1447, 2004.

[278] R. Mendez and J. D. Richter. Translational control by CPEB: a means to the end. *Nat. Rev. Mol. Cell Biol.*, 2(7):521–529, 2001.

[279] E. Merino and C. Yanofsky. Regulation by termination-antitermination: a genomic approach. In A. L. Sonenshein, J. A. Hoch, and R. Losick, editors, Bacillus subtilis *and its closest relatives: From Genes to Cells*, pages 323–336. ASM Press, Washington D.C., 2002.

[280] A. A. Michels, A. Fraldi, Q. Li, T. E. Adamson, F. Bonnet, V. T. Nguyen, S. C. Sedore, J. P. Price, D. H. Price, L. Lania, and O. Bensaude. Binding of the 7SK snRNA turns the HEXIM1 protein into a P-TEFb (CDK9/cyclin T) inhibitor. *EMBO J.*, 23:2608–2619, 2004.

[281] F. Mignone, C. Gissi, S. Liuni, and G. Pesole. Untranslated regions of mRNAs. *Genome Biology*, 3(3), February 2002. reviews0004.1-0004.10.

[282] N. Mise, Y. Goto, N. Nakajima, and N. Takagi. Molecular cloning of antisense transcripts of the mouse *Xist* gene. *Biochem. Biophys. Res. Commun.*, 258:537–541, 1999.

[283] R. K. Mishra and G. L. Eliceiri. Three small nucleolar RNAs that are involved in ribosomal RNA precursor processing. *Proc. Natl. Acad. Sci. USA*, 94:4972–4977, 1997.

[284] J. R. Mitchell, J. Cheng, and C. K. A box H/ACA small nucleolar RNA-like domain at the human telomerase 3'end. *Mol. Cell Biol.*, 19:567–576, 1999.

[285] T. Miyata, T. Yasunaga, and T. Nishida. Nucleotide sequence divergence and functional constraints in mRNA evolution. *Genetics*, 77(12):7328–7332, 1980.

[286] K. Mochizuki, N. A. Fine, T. Fujisawa, and M. A. Gorovsky. Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in tetrahymena. *Cell*, 110:689–699, 2002.

[287] J. Møller-Jensen, T. Franch, and K. Gerdes. Temporal translation control by metastable RNA structure. *J. Biol. Chem.*, 276:35707–35713, 2001.

[288] K. Montzka Wassarman and G. Storz. 6S RNA regulates *E. coli* RNA polymerase activity. *Cell*, 101:613–623, 2000.

[289] P. B. Moore and T. A. Steitz. The involvement of RNA in ribosome function. *Nature*, 418:229–235, 2002.

[290] C. Morey and P. Avner. Employment opportunities for non-coding RNAs. *FEBS Letters*, 567:27–34, 2004.

[291] B. Morgenstern. DIALIGN2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, 15:211–218, 1999.

[292] J. P. Morrissey and D. Tollervey. Birth of the snoRNPs: the evolution of RNase MRP and and the eukaryotic pre-rRNA-processing system. *Trends Biol. Sci.*, 20:78–82, 1995.

[293] A. Mosig, K. Sameith, and P. F. Stadler. *fragrep*: Efficient search for fragmented patterns in genomic sequences. *Preprint*, 2004. submitted.

[294] Z. Mourelatos, J. Dostie, S. Paushkin, A. Sharma, B. Charroux, L. Abel, J. Rappsilber, M. Mann, and G. Dreyfuss. miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev.*, 16:720–728, 2002.

[295] E. Myslinski, A. Krol, and P. Carbon. Characterization of snRNA and snRNA-type genes in the pufferfish *Fugu rubripes*. *Gene*, 330:149–158, 2004.

[296] M. K. Nag, T. T. Thai, E. A. Ruff, N. Selvamurugan, M. Kunnimalaiyaan, and G. L. Eliceiri. Genes for E1, E2, and E3 small nucleolar RNAs. *Proc. Natl. Acad. Sci. USA*, 90:9001–9005, 1993.

[297] J. H. A. Nagel, A. P. Gultyaev, K. Gerdes, and C. W. A. Pleij. Metastable structures and refolding kinetics in *hok* mRNA of plasmid R1. *RNA*, 5:1408–1419, 1999.

[298] A. Nahvi, J. E. Barrick, and R. R. Breaker. Coenzyme b_{12} riboswitches are widespread genetic control elements in prokaryotes. *Nucl. Acids Res.*, 32:143–150, 2004.

[299] P. Nelson, M. Kiriakidou, A. Sharma, E. Maniataki, and Z. Mourelatos. The microRNA world: small is mighty. *Trends Biochem. Sci.*, 28:534–540, 2003.

[300] T. W. Nilsen. Evolutionary origin of SL-addition *trans*-splicing: still an enigma. *Trends Genet.*, 17:678–680, 2001.

[301] T. W. Nilsen. The spliceosome: the most complex molecular machine in

the cell? *Bioessays*, 25:1147–1149, 2003.

[302] I. Nitta, Y. Kamada, H. Noda, T. Ueda, and K. Watanabe. Reconstitution of peptide bond formation with *Escherichia coli* 23S ribosomal RNA domains. *Science*, 281:666–669, 1998.

[303] E. Nudler and A. S. Mironov. The riboswitch control of bacterial metabolism. *Trends Biochem. Sci.*, 29(1):11–17, 2004.

[304] C. A. O'Brien, K. Margelot, and S. L. Wolin. Xenopus Ro ribonucleoproteins: Members of an evolutionarily conserved class of cytoplasmic ribonucleoproteins. *Proc. Natl. Acad. Sci. USA*, 90:7250–7254, 1993.

[305] K. Oguchi, K. Tamura, and H. Takahashi. Characterization of *Oryza sativa* telomerase reverse transcriptase and possible role of its phosphorylation in the control of telomerase activity. *Gene*, 342:57–66, 2004.

[306] M. Ohno and I. Mattaj. Meiosis: MeiRNA hits the spot. *Curr. Biol.*, 28:R66–R69, 1999.

[307] Y. Oleynikov and R. H. Singer. RNA localization: different zipcodes, same postman? *Trends in cell biology*, 8:381–383, 1998.

[308] G. J. Olsen and C. R. Woese. Ribosomal RNA: A key to phylogeny. *FASEB J.*, 7:113–123, 1993.

[309] A. Omer, T. Lowe, A. Russel, H. Ebhardt, S. Eddy, and P. Dennis. Homologs of small nucleolar RNAs in Archaea. *Science*, 288:517–522, 2000.

[310] S. Omoto, M. Ito, Y. Tsutsumi, Y. Ichikawa, H. Okuyama, E. Andi Brisibe, N. K. Saksena, and Y. Fuji. HIV-1 nef suppression by virally encoded microRNA. *Retrovirology*, 1:44, 2004. Epub.

[311] J. Otsuka and N. Sugaya. Advanced formulation of base pair changes in the stem regions of ribosomal RNAs; its application to mitochondrial rRNAs for resolving the phylogeny of animals. *J. Theor. Biol.*, 222:447–460, 2003.

[312] K. C. Pang, S. Stephen, P. G. Engström, K. Tajul-Arifin, W. Chen, C. Wahlestedt, B. Lenhard, Y. Hayashizaki, and J. S. Mattick. RNAdb — comprehensive mammalian noncoding RNA database. *Nucl. Acids Res.*, 33:D125–D130, 2005. Database issue.

[313] G. Panopoulou, S. Hennig, D. Groth, A. Krause, A. J. Poustka, R. Herwig, M. Vingron, and H. Lehrach. New evidence for genome-wide duplications at the origin of vertebrates using an amphioxus gene set and completed animal genomes. *Genome Res.*, 13:1056–1066, 2003.

[314] A. E. Pasquinelli, A. McCoy, E. Jiménez, S. Emili, G. Ruvkun, M. Q. Martindale, , and Baguñà. Expression of the 22 nucleotide let-7 heterochronic RNA throughout the metazoa: a role in life history evolution? *Evol. Dev.*, 5:372–378, 2003.

[315] A. E. Pasquinelli, B. J. Reinhart, F. Slack, M. Q. Martindale, M. I. Kurodak, B. Maller, D. C. Hayward, E. E. Ball, B. Degnan, P. Müller, J. Spring, A. Srinivasan, M. Fishman, J. Finnerty, J. Corbo, M. Levine, P. Leahy, E. Davidson, and G. Ruvkun. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*,

408:86–89, 2000.

[316] A. A. Patel and J. A. Steitz. Splicing double: insights from the second spliceosome. *Nat. Rev. Mol. Cell Biol.*, 4:960–970, 2003.

[317] M. R. Paule and R. J. White. Survey and summary: transcription by RNA polymerases i and iii. *Nucl. Acids Res.*, 28:1283–1298, 2000.

[318] G. Pavesi, G. Mauri, M. Stefani, and G. Pesole. RNAProfile: an algorithm for finding conserved secondary structure motifs in unaligned RNA sequences. *Nucl. Acids Res.*, 32(10):3258–3269, 2004.

[319] J. S. Pedersen, I. M. Meyer, R. Forsberg, and J. Hein. An evolutionary model for protein-coding regions with conserved RNA structure. *Mol. Biol. Evol.*, 21:1913–1922, 2004.

[320] J. S. Pedersen, I. M. Meyer, R. Forsberg, P. Simmonds, and J. Hein. A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucl. Acids Res.*, 32:4925–4936, 2004.

[321] D. Penny and A. Poole. The nature of the last universal common ancestor. *Curr. Opin. Genet. Dev.*, 9:672–677, 1999.

[322] G. D. Penny, G. F. Kay, S. A. Sheardown, S. Rastan, and N. Brockdorff. The *Xist* gene is required in cis for X chromosome inactivation. *Nature*, 379:131–137, 1996.

[323] G. Pesole, C. Gissi, G. Grillo, F. Licciulli, S. Liuni, and C. Saccone. Analysis of oligonucleotide AUG start codon context in eukariotic mRNAs. *Gene*, 261(1):85–91, December 2000.

[324] G. Pesole, S. Liuni, and M. D'Souza. PatSearch: a pattern matcher software that finds functional elements in nucleotide and protein sequences and assesses their statistical significance. *Bioinformatics*, 16(5):439–450, 2000.

[325] G. Pesole, S. Liuni, G. Grillo, F. Licciulli, F. Mignone, C. Gissi, and C. Saccone. UTRdb and UTRSite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs. update 2002. *Nucl. Acids Res.*, 30(1):335–340, 2002.

[326] G. Pesole, F. Mignone, C. Gissi, G. Grillo, F. Licciulli, and S. Liuni. Structural and functional features of eukaryotic mRNA untranslated regions. *Gene*, 276:73–81, 2001.

[327] K. Peterson and D. J. Eernisse. Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S DNA gene sequences. *Evol. Devel.*, 3:170–205, 2001.

[328] S. Pfeffer, M. Zavolan, F. A. Grasser, M. Chien, J. J. Russo, J. Ju, B. John, A. J. Enright, D. Marks, C. Sander, and T. Tuschl. Identification of virus-encoded microRNAs. *Science*, 304:734–736, 2004.

[329] S. C. Phillips and P. C. Turner. Sequence and expression of a mouse U7 snRNA type II pseudogene. *DNA Seq.*, 1:401–404, 1991.

[330] V. Pirotta. Trans-splicing in drosophila. *Bioessays*, 24:988–991, 2002.

[331] C. Pitulle, M. Garcia-Paris, K. R. Zamudio, and N. R. Pace. Comparative structural analysis of vertebrate ribonuclease P RNA. *Nucl. Acids Res.*, 26:3333–3339, 1998.

[332] N. J. Pokrywka and E. C. Stephenson. Microtubules mediate the localization of bicoid RNA during *Drosophila* oogenesis. *Dev.*, 113:55–66, 1991.

[333] A. Poole, D. Penny, and B.-M. Sjäberg. Methyl-RNA: an evolutionary bridge between RNA and DNA? *Chem. & Biol.*, 7:R207–R216, 2000.

[334] S. S. Potter and W. W. Branford. Evolutionary conservation and tissue-specific processing of *Hoxa 11* antisense transcripts. *Mamm. Genome*, 9:799–806, 1998.

[335] E. M. Precott and N. J. Proudfoot. Transcriptional collision between convergent genes in budding yeast. *Proc. Natl. Acad. Sci. USA*, 99:8796–8801, 2002.

[336] S. J. Prohaska, C. Fried, C. Flamm, G. P. Wagner, and P. F. Stadler. Surveying phylogenetic footprints in large gene clusters: Applications to Hox cluster duplications. *Mol. Phyl. Evol.*, 31:581–604, 2004.

[337] H. Putzer, N. Gendron, and M. Grunberg-Manago. Co-ordinate expression of the two threonyl-tRNA synthetase genes in Bacillus subtilis: Control by transcriptional antitermination involving a conserved regulatory sequence. *EMBO J.*, 11:3117–3127, 1992.

[338] V. Ramakrishnan and P. B. Moore. Atomic structures at last: the ribosome in 2000. *Curr. Opinions Struct. Biol.*, 11:144–154, 2001.

[339] M. Regalia, M. A. Rosenblad, and T. Samuelson. Prediction of signal recognition particle RNA genes. *Nucl. Acids Res.*, 30:3368–3377, 2002.

[340] M. Rehmsmeier, P. Steffen, M. Höchsmann, and R. Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10:1507–1517, 2004.

[341] F. J. Reinhart, B. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horwitz, and G. Ruvkun. The 21-nucleotide RNA *let-7* regulates developmental timing in *Caenorhabditis elegans*. *Nature*, 403:901–906, 2000.

[342] E. Rivas and S. R. Eddy. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, 16(7):583–605, 2000.

[343] E. Rivas and S. R. Eddy. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2:8, 2001.

[344] E. Rivas, R. J. Klein, T. A. Jones, and S. R. Eddy. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr. Biol.*, 11:1369–1373, 2001.

[345] S. Rodin, O. S, and A. Rodin. Transfer RNAs with complementary anticodons: could they reflect early evolution of discriminative genetic code adaptors? *Proc. Natl. Acad. Sci. USA*, 90:4723–4727, 1993.

[346] D. A. Rodionov, A. G. Vitreschak, A. A. Mironov, and M. S. Gelfand. Comparative genomics of the methionine metabolism in Gram-positive bacteria: a variety of regulatory systems. *Nucl. Acids Res.*, 32:3340–3353, 2004.

[347] A. Rodriguez, S. Griffiths-Jones, J. L. Ashurst, and A. Bradley. Iden-

tification of mammalian microRNA host genes and transcription units. *Genome Res*, 14:1902–1910, 2004.

[348] A. P. Rooney. Mechanisms underlying the evolution and maintenance of functionally heterogeneous 18S rRNA genes in apicomplexans. *Mol. Biol. Evol.*, 21:1704–1711, 2004.

[349] M. A. Rosenblad, J. Gorodkin, B. Knudsen, and C. Zwieb. SRPDB: Signal Rrecognition Particle database. *Nucl. Acids Res.*, 31:363–364, 2003.

[350] M. A. Rosenblad and T. Samuelsson. Identification of chloroplast signal recognition particle RNA genes. *Plant Cell Physiol.*, 45:1633–1639, 2004.

[351] M. A. Rosenblad, C. Zwieb, and T. Samuelson. Identification and comparative analysis of components from the signal recognition particle in protozoa and fungi. *BMC Genomics*, 5:# 5, 2004.

[352] R. R. Rueckert. Picornaviridae: The viruses and their replication. In N. Fields, D. Knipe, and P. Howley, editors, *Virology*, volume 1, pages 609–654. Lippincott-Raven Publishers, Philadelphia, New York, third edition, 1996.

[353] A. G. Russell, M. N. Schnare, and M. W. Gray. Pseudouridine-guide RNAs and other Cbf5p-associated RNAs in *Euglena gracilis*. *RNA*, 10:1034–1046, 2004.

[354] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425, 1987.

[355] D. Samarsky and M. Fournier. A comprehensive database for the small nucleolar RNAs from saccharomyces cerevisiae. *Nucleic Acids Res.*, 27:161–164, 1999.

[356] D. A. Samarsky, G. S. Schneider, and M. J. Fournier. An essential domain in *Saccaromyces cerevisiae* U14 snoRNA is absent in vertebrates, but conserved in other yeasts. *Nucl. Acids Res.*, 24:2059–2066, 1996.

[357] D. Sankoff. Simultaneous solution of the RNA folding, alignment, and proto-sequence problems. *SIAM J. Appl. Math.*, 45:810–825, 1985.

[358] N. J. Savill, D. C. Hoyle, and P. G. Higgs. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. *Genetics*, 157:399–411, 2001.

[359] E. C. Scharl and J. A. Steitz. Length suppression in histone messenger RNA 3-end maturation: Processing defects of insertion mutant premessenger RNAs can be compensated by insertions into the U7 small nuclear RNA. *Proc. Natl. Acad. Sci. USA*, 93:14659–14664, 1996.

[360] P. Schattner. Searching for RNA genes using base-composition statistics. *Nucl. Acids Res.*, 30(9):2076–2082, 2002.

[361] C. Schlötterer and D. Tautz. Chromosomal homogeneity of drosophila ribosomal DNA arrays suggests intrachromosomal exchanges drive concerted evolution. *Curr. Biol.*, 4:777–783, 1994.

[362] M. Schöninger and A. von Haeseler. Towards assigning helical regions in alignments of ribosomal RNA and testing the appropriateness of evolutionary models. *J. Mol. Evol.*, 49:691–698, 1999.

[363] L. Schramm and N. Hernandez. Recruitment of RNA polymerase III to its target promoters. *Genes Dev.*, 16:2593–2620, 2002.

[364] E. A. Schultes and D. P. Bartel. One sequence, two ribozymes: Implications for the emergence of new ribozyme folds. *Science*, 289:448–452, 2000.

[365] E. A. Schultes, P. T. Hraber, and T. H. LaBean. Estimating the contributions of selection and self-organization in RNA secondary structure. *J. Mol. Evol.*, 49:76–83, 1999.

[366] D. Schümperli and R. S. Pillai. The special Sm core structure of the U7 snRNP: far-reaching significance of a small nuclear ribonucleoprotein. *Cell. Mol. Life Sci.*, 61:2560–2570, 2004.

[367] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: A case study in RNA secondary structures. *Proc. Roy. Soc. Lond. B*, 255:279–284, 1994.

[368] H. Seitz, H. Royo, S.-P. Lin, N. Youngson, A. C. Ferguson-Smith, and J. Cavaillé. Imprinted small RNA genes. *Biol Chem*, 385:905–911, 2004.

[369] N. Selvamurugan and G. L. Eliceiri. The gene for human E2 small nucleolar RNA resides in an intron of a laminin-binding protein gene. *Genomics*, 30:400–401, 1995.

[370] S. A. Shabalina, A. Y. Ogurtsov, I. B. Rogozin, E. V. Koonin, and D. J. Lipman. Comparative analysis of orthologous eukaryotic mRNAs: potential hidden functional signals. *Nucl. Acids Res.*, 32(5):1774–1782, 2004.

[371] B. A. Shapiro and K. Zhang. Comparing multiple RNA secondary structures using tree comparisons. *CABIOS*, 6:309–318, 1990.

[372] S. M. Sharkady and K. P. Williams. A third lineage with two-piece tmRNA. *Nucl. Acids Res.*, 32:1–8, 2004.

[373] J. Shendure and G. M. Church. Computational discovery of sense-antisense transcription in the human and mouse genome. *Genome Biol.*, 3:1–14, 2002.

[374] S. Siebert and R. Backofen. MARNA: A server for multiple alignment of RNAs. In H.-W. Mewes, V. Heun, D. Frishman, and S. Kramer, editors, *Proceedings of the German Conference on Bioinformatics. GCB 2003*, volume 1, pages 135–140, München, D, 2003. belleville Verlag Michael Farin.

[375] S. K. Silverman. Rube goldberg goes (ribo)nuclear? molecular switches and sensors made from RNA. *RNA*, 9:377–383, 2003.

[376] L. Simpson, O. H. Thiemann, N. J. Savill, J. D. Alfonzo, and D. A. Maslov. Evolution of RNA editing in trypanosome mitochondria. *Proc. Natl. Acad. Sci. USA*, 97:6986–6993, 2000.

[377] D. Soldati and D. Schümperli. Structural and functional characterization of mouse U7 small nuclear RNA active in 3' processing of histone pre-mRNA. *Mol. Cell Biol.*, 8:1518–1524, 1988.

[378] G. A. Soukup and R. R. Breaker. Engineering precision RNA molecular switches. *Proc. Natl. Acad. Sci. USA*, 96:3584–3589, 1999.

[379] M. Sprinzl, C. Horn, M. Brown, A. Ioudovitch, and S. Steinberg. Compilation of tRNA sequences and sequences of tRNA genes. *Nucl. Acids Res.*, 26:148–153, 1998.
[380] M. Sprinzl, C. Steegborn, F. Hübel, and S. Steinberg. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res*, 24:68–72, 1996.
[381] M. Sprinzl and K. S. Vassilenko. Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic Acids Res*, 33 Database Issue:139–140, 2005.
[382] T. A. Steitz and P. B. Moore. RNA, the first macromolecular catalyst: the ribosome is a ribozyme. *Trends Biochem. Sci.*, 28:411–418, 2003.
[383] G. Storz, J. A. Opdyke, and A. Zhang. Controlling mRNA stability and translation with small noncoding RNAs. *Cur. Op. Microbiol.*, 7:140–144, 2004.
[384] K. Stuart, T. E. Allen, S. Heidmann, and S. S. D. RNA editing in kinetoplastid protozoa. *Microbiol. Mol. Biol. Rev.*, 61:105–120, 1997.
[385] N. Sudarsan, J. E. Barrick, and R. R. Breaker. Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA*, 9:644–647, 2003.
[386] B. A. Sullenger. Riboswitches—to kill or save the messenger. *N. Engl. J. Med.*, 351:2759–2760, 2004.
[387] K. Sumiyama, S. Q. Irvine, and F. H. Ruddle. The role of gene duplication in the evolution and function of the vertebrate Dlx/distal-less bigene clusters. *J. Struct. Funct. Genomics*, 3:151–159, 2003.
[388] Y. Sun, S. Koo, N. White, E. Peralta, C. Esau, N. M. Dean, and R. J. Perera. Development of a micro-array to detect human and mouse microRNAs and characterization of expression in human organs. *Nucleic Acids Res*, 32:doi:10.1093/nar/gnh186, 2004.
[389] M. Suzuki and Y. Hayashizaki. Mouse-centric comparative transcriptomics of protein coding and non-coding RNAs. *BioEssays*, 26:833–843, 2004.
[390] M. Szymański, M. Barciszewska, J. Barciszewski, and V. Erdmann. 5S ribosomal RNA database Y2K. *Nucl. Acids Res.*, 28:166–167, 2000.
[391] M. Szymański, M. Z. Barciszewska, M. Żywicki, and J. Barciszewski. Noncoding RNA transcripts. *J. Appl. Genet.*, 44:1–19, 2003.
[392] E. Talla, V. Anthouard, C. Bouchier, L. Frangeul, and B. Dujon. The complete mitochondrial genome of the yeast *Kluyveromyces thermotolerans*. *FEBS letter*, 579:30–40, 2005.
[393] T.-H. Tang, J.-P. Bachellerie, T. Rozhdestvensky, M.-L. Bortolin, H. Huber, M. Drungowski, T. Elge, J. Brosius, and A. Hüttenhofer. Identification of 86 candidates for small non-messenger RNAs from the archaeon *Archaeoglobus fulgidus*. *Proc. Natl. Acad. Sci. USA*, 99:7536–7541, 2002.
[394] A. Tanzer, C. T. Amemiya, C.-B. Kim, and P. F. Stadler. Evolution of microRNAs located within *Hox* gene clusters. *J. Exp. Zool.: Mol. Dev. Evol.*, 304B:75–85, 2005. doi: 10.1002/jez.b.21021.
[395] A. Tanzer and P. F. Stadler. Molecular evolution of a microRNA cluster.

J. Mol. Biol., 339:327–335, 2004.

[396] W. Y. Tarn, T. A. Yario, and J. A. Steitz. U12 snRNAs in vertebrates: Evolutionary conservation of 5' sequences implicated in splicing of pre-mRNAs containing a minor class of introns. *RNA*, 1:644–656, 1995.

[397] M. J. Telford and P. W. H. Holland. Evolution of 28S ribosomal DNA in chaetognaths: Duplicate genes and molecular phylogeny. *J. Mol. Evol.*, 44:135–144, 1997.

[398] M. P. Terns and R. M. Terns. Small nucleolar RNAs: Versatile *trans*-acting molecules of ancient evolutionary origin. *Gene Expr.*, 10:17–39, 2002.

[399] S. W. Teunissen, M. J. Kruithof, A. D. Farris, J. B. Harley, W. J. Venrooij, and G. J. Pruijn. Conserved features of Y RNAs: a comparison of experimentally derived secondary structures. *Nucl. Acids Res.*, 28:610–619, 2000.

[400] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucl. Acids Res.*, 22(22):4673–4680, 1994.

[401] M. Thompson, R. A. Haeusler, P. D. Good, and D. R. Engelke. Nucleolar clustering of dispersed tRNA genes. *Science*, 302:1399–1401, 2003.

[402] C. Thurner, C. Witwer, I. Hofacker, and P. F. Stadler. Conserved RNA secondary structures in Flaviviridae genomes. *J. Gen. Virol.*, 85:1113–1124, 2004.

[403] E. Tran, J. Brown, and S. E. Maxwell. Evolutionary origins of the RNA-guided nucleotide-modification complexes: from the primitive translation apparatus? *Trends Biochem. Sci.*, 29:343–350, 2004.

[404] C. Tschudi and E. Ullu. Unconventional rules of small nuclear RNA transcription and cap modification in trypanosomatids. *Gene Expr.*, 10:3–16, 2002.

[405] A. Tuplin, D. J. Evans, and P. Simmonds. Detailed mapping of RNA secondary structures in core and NS5B-encoding region sequence of hepatitis C virus by RNase cleavage and novel bioinformatic prediction methods. *J. Gen. Virol.*, 85:3037–3047, 2004.

[406] A. Tuplin, J. Wood, D. J. Evans, A. H. Patel, and P. Simmonds. Thermodynamic and phylogenetic prediction of RNA secondary structures in the coding region of hepatitis C virus. *RNA*, 8:824–841, 2002.

[407] I. A. Turner, C. M. Norman, M. J. Churcher, and N. A. J. Roles of the U5 snRNP in spliceosome dynamics and catalysis. *Biochem. Soc. Trans.*, 32:928–931, 2004.

[408] K. T. Tycowski, A. Aab, and J. A. Steitz. Guide RNAs with 5' caps and novel box C/D snoRNA-like domains for modification of snRNAs in metazoa. *Curr. Biol.*, 14:1985–1995, 2004.

[409] K. T. Tycowski and J. A. Steitz. Non-coding snoRNA host genes in Drosophila: expression strategies for modification guide snoRNAs. *Eur. J. Cell. Biol.*, 80:119–125, 2001.

[410] S. Uliel, X.-h. Liang, R. Unger, and S. Michaeli. Small nucleolar RNAs that guide modification in trypanosomatids: repertoire, targets, genome organization, and unique functions. *Int. J. Parasit.*, 34:445–454, 2004.

[411] P. J. Unrau and D. P. Bartel. RNA-catalysed nucleotide synthesis. *Nature*, 395:260–263, 1998.

[412] C. Ushida, A. Yoshida, Y. Miyakawa, Y. Ara, and A. Muto. Distribution of the MCS4 RNA genes in mycoplasmas belonging to the *Mycoplasma mycoides* cluster. *Gene*, 314:149–155, 2003.

[413] S. Valadkhan and J. L. Manley. Splicing-related catalysis by protein-free snRNAs. *Nature*, 413:701–707, 2001.

[414] S. Valadkhan and J. L. Manley. Characterization of the catalytic activity of U2 and U6 snRNAs. *RNA*, 9:892–904, 2003.

[415] Y. Van de Peer, S. L. Baldauf, W. F. Doolittle, and A. Meyer. An updated and comprehensive rRNA phylogeny of (crown) eukaryotes based on rate-calibrated evolutionary distances. *J. Mol. Evol.*, 51:565–576, 2000.

[416] Y. Van de Peer, P. De Rijk, J. Wuyts, T. Winkelmans, and R. De Wachter. The european small subunit ribosomal RNA database. *Nucl. Acids Res.*, 28:175–176, 2000.

[417] D. J. Van Horn, D. Eisenberg, C. A. O'Brien, and S. L. Wolin. Caenorhabditis elegans embryos contain only one major species of Ro RNP. *RNA*, 1:293–303, 1995.

[418] A. van Zon, M. Mossink, M. Schoester, G. Scheffer, R. Scheper, P. Sonneveld, and E. Wiemer. Multiple human vault RNAs. Expression and association with the vault complex. *J. Biol. Chem.*, 276:37715–37721, 2001.

[419] S. K. Vasu and L. H. Rome. *Dictyostelium* vaults: Disruption of the major proteins reveals growth and morphological defects and uncovers a new associated protein. *J. Biol. Chem.*, 270:16588–16594, 1995.

[420] M. C. Vella, K. Reinert, and F. J. Slack. Architecture of a validated MicroRNA::Target interaction. *Chem Biol*, 11:1619–1623, 2004.

[421] P. Vitali, H. Royo, H. Seitz, J.-P. Bachellerie, A. Hüttenhofer, and J. Cavaillé. Identification of 13 novel human modification guide RNAs. *Nucl. Acids Res.*, 31:6543–6551, 2003.

[422] A. G. Vitreschak, D. A. Rodionov, A. A. Mironov, and M. S. Gelfand. Regulation of the vitamine B_{12} metabolism and transport in bacteria by a conserved RNA structural element. *RNA*, 9:1084–1097, 2003.

[423] A. G. Vitreschak, D. A. Rodionov, A. A. Mironov, and M. S. Gelfand. Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Gen.*, 20(1):44–50, 2004.

[424] J. Vogel, V. Bartels, T. H. Tang, G. Churakov, J. G. Slagter-Jäger, A. Hüttenhofer, and G. H. E. Wagner. RNomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucl. Acids Res.*, 31:6435–6443, 2003.

[425] E. G. H. Wagner and K. Flärdh. Antisense RNAs everywhere? *Trends*

Genet., 18:223–226, 2002.

[426] S. Washietl and I. L. Hofacker. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol.*, 342:19–30, 2004.

[427] S. Washietl, I. L. Hofacker, and P. F. Stadler. Fast and reliable detection of noncoding RNAs. *Proc. Natl. Acad. Sci.*, 102, 2005. doi: 10.1073/pnas.0409169102.

[428] D. A. Wassarman and J. A. Steitz. Structural analyses of the 7SK ribonucleoprotein (RNP), the most abundant human small RNP of unknown function. *Mol. Cell. Biol.*, 11:3432–3445, 1991.

[429] K. Wassarman, F. Repoila, C. Rosenow, G. Storz, and S. Gottesman. Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev.*, 15:1637–1651, 2001.

[430] M. J. Weber. New human and mouse microRNA genes found by homology search. *FEBS J*, 272:59–73, 2005.

[431] A. M. Weiner and R. A. Denison. Either gene amplification or gene conversion may maintain the homogeneity of the multigene family encoding human U1 small nuclear RNA. *Cold Spring Harb. Symp. Quant. Biol.*, 47:1141–1149, 1983.

[432] L. B. Weinstein and J. A. Steitz. Guided tours: from precursor snoRNA to functional snoRNP. *Curr. Op. Cell Biol.*, 11:378–384, 1999.

[433] A. Werner, K. Preston-Fayers, L. Dehmelt, and P. Nalbant. Regulation of the NPT gene by a naturally occurring antisense transcript. *Cell Biochem. Biophys.*, 36:241–252, 2002.

[434] E. Westhof and C. Massire. Evolution of RNA architecture. *Science*, 306:62–63, 2004.

[435] R. J. White. *RNA Polymerase III Transcription*. Springer-Verlag, New York, NY, 1998.

[436] G. S. Wilkie, K. S. Dickson, and N. G. Gray. Regulation of mRNA translation by 5'- and 3'-UTR-binding factors. *TRENDS in Biochemical Sciences*, 28(4):182–188, 2003.

[437] K. P. Williams. Descent of a split DNA. *Nucl. Acids Res.*, 30:2025–2030, 2002.

[438] W. C. Winkler and R. R. Breaker. Genetic control by metabolite-binding riboswitches. *Chembiochem*, 4(10):1024–1032, 2003.

[439] C. Witwer, S. Rauscher, I. Hofacker, and P. Stadler. Conserved RNA secondary structures in picornaviridae genomes. *Nucl. Acids Res.*, 29:5079–5089, 2001.

[440] A. P. Wolffe. The role of transcription factors, chromatin structure and DNA replication in 5S RNA gene regulation. *J. Cell Sci.*, 107:2055–2063, 1994.

[441] V. Wood, R. Gwilliam, M. A. Rajandream, and *et al.* (132 co-authors). The genome sequence of *Schizosaccharomyces pombe*. *Nature*, 415:871–880, 2002.

[442] C.-H. H. Wu and J. G. Gall. U7 small nuclear RNA in C snurposomes of

the Xenopus germinal vesicle. *Proc. Natl. Acad. Sci. USA*, 90:6257–6259, 1993.

[443] J. Wuyts, P. De Rijk, Y. Van de Peer, T. Winkelmans, and R. De Wachter. The european large subunit ribosomal RNA database. *Nucl. Acids Res.*, 29:175–177, 2001.

[444] M. C. Yao, P. Fuller, and X. Xi. Programmed DNA deletion as an RNA-guided system of genome defense. *Science*, 300:1517–1518, 2003.

[445] A. J. Ye and D. P. Romero. Phylogenetic relationships amongst tetrahymenine ciliates inferred by a comparison of telomerase RNAs. *Int. J. Syst. Evol. Microbiol.*, 52:2297–2302, 2002.

[446] S. Yekta, I.-h. Shih, and D. P. Bartel. MircoRNA-directed cleavage of *HoxB8* mRNA. *Science*, 304:594–596, 2004.

[447] R. Yelin, D. Dahary, R. Sorek, E. Y. Levanon, O. Goldstein, A. Shoshan, A. Diber, S. Biton, Y. Tamir, R. Khosravi, S. Nemzer, E. Pinner, S. Walach, J. Bernstein, K. Savitsky, and G. Rotman. Widespread occurrence of antisense transcription in the human genome. *Nat. Biotechnol.*, 21:379–386, 2003.

[448] J. H. Yik, R. Chen, R. Nishimura, J. L. Jennings, A. J. Link, and Q. Zhou. Inhibition of P-TEFb (CDK9/Cyclin T) kinase and RNA polymerase II transcription by the coordinated actions of HEXIM1 and 7SK snRNA. *Mol. Cell*, 12:971–982, 2003.

[449] S.-Y. Ying and S.-L. Lin. Intron-derived microRNAs — fine tuning of gene functions. *Gene*, 342:25–28, 2004.

[450] S.-Y. Ying and S.-L. Lin. Intronic microRNAs. *Biochem Biophys Res Commun*, 326:515–520, 2005.

[451] Y.-T. Yu, W.-Y. Tarn, T. A. Yario, and J. A. Steitz. More Sm snRNAs from vertebrate cells. *Exp. Cell Res.*, 229:276–281, 1996.

[452] D. C. Zappulla and T. R. Cech. Yeast telomerase RNA: a flexible scaffold for protein subunits. *Proc. Natl. Acad. Sci. USA*, 101:10024–10029, 2004.

[453] Z. Zhang and M. Gerstein. Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. *J. Biol.*, 2:11; 4 pp., 2003.

[454] S. Zimmerly, G. Hausner, and X.-c. Wu. Phylogenetic relationships among group II intron ORFs. *Nucl. Acids Res.*, 29:1238–1250, 2001.

[455] C. Zwieb and J. Eichler. Getting on target: the archaeal signal recognition particle. *Archaea*, 1:27–34, 2002.

[456] C. Zwieb, R. W. van Nues, M. A. Rosenblad, J. D. Brown, and T. Samuelson. A nomenclature for all signal recognition particle RNAs. *RNA*, 11:7–13, 2005.

[457] C. Zwieb and J. Wower. tmRDB (tmRNA database). *Nucleic Acids Res.*, 28:169–170, 2000.