1602–1615 *Nucleic Acids Research, 2009, Vol. 37, No. 5* *Published online 16 January 2009*
doi:10.1093/nar/gkn1084

Non-coding RNA annotation of the genome of *Trichoplax adhaerens*

Jana Hertel[1], Danielle de Jong[2], Manja Marz[1], Dominic Rose[1], Hakim Tafer[3], Andrea Tanzer[1,3,4], Bernd Schierwater[2] and Peter F. Stadler[1,3,5,6,*]

[1]Bioinformatics Group, Department of Computer Science, Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraβe 16-18, D-04107 Leipzig, [2]Division of Ecology and Evolution, Institut für Tierökologie und Zellbiologie, Tierärztliche Hochschule Hannover, Bünteweg 17d, D-30559 Hannover, Germany, [3]Department of Theoretical Chemistry, University of Vienna, Währingerstraβe 17, A-1090 Wien, Austria, [4]Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520, USA, [5]RNomics Group, Fraunhofer Institut für Zelltherapie und Immunologie, Deutscher Platz 5e, D-04103 Leipzig, Germany and [6]Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA

Received November 13, 2008; Revised December 22, 2008; Accepted December 23, 2008

ABSTRACT

A detailed annotation of non-protein coding RNAs is typically missing in initial releases of newly sequenced genomes. Here we report on a comprehensive ncRNA annotation of the genome of *Trichoplax adhaerens*, the presumably most basal metazoan whose genome has been published to-date. Since `blast` identified only a small fraction of the best-conserved ncRNAs—in particular rRNAs, tRNAs and some snRNAs—we developed a semi-global dynamic programming tool, `GotohScan`, to increase the sensitivity of the homology search. It successfully identified the full complement of major and minor spliceosomal snRNAs, the genes for RNase P and MRP RNAs, the SRP RNA, as well as several small nucleolar RNAs. We did not find any microRNA candidates homologous to known eumetazoan sequences. Interestingly, most ncRNAs, including the pol-III transcripts, appear as single-copy genes or with very small copy numbers in the *Trichoplax* genome.

INTRODUCTION

The phylum Placozoa consists of only one recognized species—the marine dweller *Trichoplax adhaerens*. Extensive genetic variation between individual placozoan lineages, however, suggests the existence of different species (1). The phylogenetic position of the phylum Placozoa has been the subject of contention dating from the 19th century. Originally, Placozoa were regarded to represent the base of Metazoa, later they were seen as derived (secondarily reduced) with sponges being considered to be the most basal metazoans [see e.g. (2,3,4) for overview and discussion]. Most recently, a basal position among all diploblastic animals has been suggested (5).

Trichoplax lacks tissues, organs and any type of symmetry. It is composed of only a few hundred to a few thousand cells. This organizm has a simple upper and lower epithelium, which surround a network of fiber cells, and as such has an irregular, three-layered, sandwich-type organization. Only five different cell-types have so far been described; upper and lower epithelial cells, glands cells, fibre cells and recently discovered type of small cells that are arranged a relatively evenly spaced pattern within the marginal zone, where upper and lower epithelia meet (6). It is therefore among the simplest multi-cellular organizm. With 106 Mb, the nuclear genome of *T. adhaerens*, which has recently been completely sequenced (7), is among the smallest animal genomes.

So far, the non-coding RNA complement of Placozoa has not been studied. The genome-wide annotation of non-coding RNAs has turned out to be a more complex and demanding problem than one might think. While a few exceptional classes of RNA genes, first and foremost rRNAs and tRNAs are readily found and annotated by `blast` and the widely used tRNA detector `tRNAscanSE` (8), most other ncRNAs are comparably poorly conserved and hard to find within complete genomes. This is in particular true whenever the sensitivity of comparative approaches are limited by large evolutionary distances to the closest well-annotated genomes. The placozoan *T. adhaerens* is a prime example for this situation: in the concatenated set of 104 slowly evolving single-copy nuclear protein-coding genes used for phylogenetic analysis in (7), for instance, the distances from *Trichoplax* to *Amphimedon*, *Nematostella* and *Human* are 0.44, 0.34 and 0.32 substitutions per site, respectively.

*To whom correspondence should be addressed. Tel: +49 341 97 16686; Fax: +49 341 97 16679; Email: jana@bioinf.uni-leipzig.de

As a consequence, only highly conserved DNA is alignable at all, and homology-based gene finding becomes a non-trivial task.

In this contribution, we primarily report on a careful annotation of those Trichoplax ncRNA genes that have well-described homologs in other animals. In addition, we describe computational surveys for novel ncRNA candidates. For a subset of the annotated ncRNAs we verify expression to demonstrate that the predicted homologs are functional genes.

MATERIALS AND METHODS

Sequence data and databases

The `Triad1` assembly of the genome of *T. adhaerens* (7) was downloaded from the website of the Joint Genome Institute (http://genome.jgi-psf.org/Triad1/). For comparison, we used the the `Nemve1` (http://genome.jgi-psf.org/Nemve1/) assembly of *Nematostella vectensis* (9), as well as the available shotgun traces of *Hydra magnapapillata*, *Amphimedon queenslandica*, *Porites lobata*, *Acropora millepora* and *Acropora palmata* (downloaded from the NCBI trace archive).

Known ncRNA sequences were extracted from the `Rfam` (10) and `NonCode` (11) databases. In addition we used the collection of metazoan snRNAs from (12). [The snRNAs found in the current study were made available to (12)].

Software

Homology searches were performed using NCBI `blastall` 2.2.6 (13), `infernal` (14), `fragrep` (15) and the novel `GotohScan` method described below in detail. Alignments were edited in the `emacs` editor using `ralee` mode (16). RNA secondary structures were computed using the `Vienna RNA Package` (17), in particular the programs `RNAfold` for individual structures, `RNAalifold` (18,19) for consensus structures of aligned RNA sequences and `RNAcofold` (20) for interaction structures. We used `RNAmicro` (21) in the updated version (1.3) (http://www.bioinf.uni-leipzig.de/~jana/software/RNAmicro.html) to identify microRNA candidates from multiple alignments. The analysis of putative snoRNAs was performed using `snoReport` (22), targets for box H/ACA snoRNAs were performed using a preliminary version of `snoplex` (H.Tafer *et al.*, in preparation). The genome-wide screens for conserved secondary structure elements were performed using `RNAz` (23) as described below.

RNAz screens

We used `multiz` (24) to produce a three-way alignment of *Trichoplax*, *Nematostella* and *Hydra*. Only the blocks that contained *Trichoplax* and at least one of the two cnidarian species were used for further analysis. In addition, we prepared a six-way alignment using `NcDNAlign` (25) that include the genomic data of the six basal metazoa listed in the previous paragraph. The *Trichoplax* sequence was used as reference and only alignment blocks containing at least three species were processed further.

These two sets of input alignments were passed to the `RNAz` pipeline and processed in the same way: alignments longer than 120 nt are cut into 120 slices in 40 nt steps. In a series of filtering steps sequences were removed from the individual alignments or alignment slices if they are (i) shorter than 50 nt or (ii) contain more than 25% gap characters or (iii) have a base composition outside the definition range of `RNAz`. All pre-processing steps were performed using the script `rnazWindows.pl` of the current release of the `RNAz` package. Overlapping slices with a positive ncRNA classification probability of $P > 0.5$ were combined using `rnazCluster.pl` to a single annotation element, which we refer to as *locus*. In order to estimate the false discovery rate (FDR) of the screen we repeated the entire procedure with shuffled input alignments using `rnazRandomizeAln.pl`.

GotohScan

`Blast` failed to identify many of the ncRNAs that are reasonably expected to be present in the *Trichoplax* genome, for example homologs of the U4atac snRNA, the U3 snoRNA, or RNA component of RNase MRP. These could not be detected by means of `blast`, even with relaxed parameter settings. We therefore decided to use a computationally more costly but more sensitive full dynamic programming approach. Instead of using a local (Smith-Waterman) implementation such as `ssearch` (26) or its partition function version (27), we suggest that a 'semi-global' alignment approach is more natural for the homology search problems at hand. In a semi-global alignment, the best match of the '*complete*' query sequence to the genomic DNA is sought. Due to the relatively long insertion and deletions, the use of an affine gap cost model becomes necessary. This problem is solved by the following straightforward modification of Gotoh's dynamics programming algorithm (28).

Denote the query sequence by $Q = q_1, q_2, \ldots, q_m$ and the genomic 'subject' sequence by $P = p_1, p_2, \ldots, p_n$. Note that the problem is not symmetric since deletions of the ends of P do not incur costs, while deletions of the ends of Q are fully penalized. As usual, denote by S_{ij} the optimal alignment of the prefixes $Q[1 \ldots i]$ and $P[1 \ldots j]$, respectively. The values of D_{ij} and F_{ij} are the optimal scores of alignments of $Q[1 \ldots i]$ and $P[1 \ldots j]$ with the constraint that the alignment is an insertion or a deletion, respectively. The recursions read

$$\begin{aligned} D_{ij} &= \max\{S_{i-1,j} + \gamma_o, D_{i-1,j} + \gamma_e\} \\ F_{ij} &= \max\{S_{i,j-1} + \gamma_o, F_{i,j-1} + \gamma_e\} \\ S_{ij} &= \max\{D_{ij}, F_{ij}, S_{i-1,j-1} + \sigma(p_i, q_j)\} \end{aligned} \quad \mathbf{1}$$

with the initializations

$$\begin{aligned} &S_{00} = 0, \\ &D_{0j} = -\infty,\ S_{0j} = F_{0,j} = \gamma_o + (j-1)\gamma_e, \\ &F_{i0} = -\infty,\ S_{i0} = D_{i,0} = \gamma_o + (i-1)\gamma_e. \end{aligned}$$

In this full version, the algorithm requires $\mathcal{O}(n \times m)$ time and memory, where n is the length of the genome and m is the length of the query sequence. While the time

requirement is uncritical on off-the-shelf PCs even for large genomes, it is necessary to reduce the memory consumption. It is sufficient to compute, for every position k in the genome the score of the best alignment of the query that has its last match in k. For this purpose, we only need to store the values of the current column S_{ij} and $D_{i-1,j}$ and of the previous column $S_{i-1,j}$ and $D_{i-1,j}$, i.e. these two quadratic arrays can be replaced by linear arrays of length m. From the F array only the current value F_{ij} and the previous value $F_{i,j-1}$ need to be stored. The alignments themselves need to be computed only for a very small subset of endpoints k of the forward recursion, namely those with nearly optimal score. For each endpoint, the alignment can be obtained by standard backtracing in $\mathcal{O}(m^2)$ time and space.

`GotohScan` is not the only implementation of a semi-global alignments. Alternative approaches use a scoring based on block alignments (29) or employ Hidden Markov Models (30). We constructed our own version since this allowed us to optimize the performance for large genomes on the available off-the-shelf PC hardware and to estimate E-values directly from the observed score distribution. To this end, the current **C** implementation of `GotohScan` stores a histogram of all the scores for each query sequence over all database sequences. The locally maximal scores for each query are computed via a simple divide and conquer implementation that starts with the global maximum and continues with the next maxima to the left and right that are at least m (length of the query sequence) nucleotides away from the global maximum. A priority queue is utilized to hold a fixed number of these top-scoring positions. It is initialized only after the first database sequence (typically the longest chromosome or scaffold) while the following high-scoring positions are inserted according to the alignment score. This minimizes the effort for backtracing candidate alignments. Figure 1a gives some example of score histograms.

Empirically, we found that the score histogram, with respect to one query sequence against all database sequences, closely follows a Gamma distribution

$$f(s;k,\theta)=\frac{1}{\theta\Gamma(k)}\left(\frac{s}{\theta}\right)^{k-1}e^{-s/\theta}, \qquad \mathbf{2}$$

see Figure 1b. Thus, we fitted a Gamma distribution to the histogram of alignment scores and used it to calculate E-values for each of the elements in the priority queue.

The `GotohScan` program uses only the high-density portion of the score histogram to estimate the characteristic quantities $\ln\langle s\rangle$ and $\langle \ln s\rangle$:

$$\ln\langle s\rangle=\ln\left(\sum_{i=a}^{b}\frac{1}{N}Sdistr[i]\right)\text{ and }\quad \langle \ln s\rangle=\frac{1}{N}\left(\sum_{i=a}^{b}\ln(Sdistr[i])\right) \qquad \mathbf{3}$$

with a and b as limits of the high-density portion of the score distribution and N the number of alignments in this range. $Sdistr[i]$ is the number of alignments with score i. From these we estimated the scale and shape parameters θ and k by least-square fitting of $\log f(s;k,\theta)$ against the logarithm of the score histogram, restricting the fitting interval to $[a:b]$ of the score distribution. The calculation of E-values then proceeds by using the asymptotic expansion (31) of the incomplete Gamma function:

$$\log E=(k-1)(\log s-\log\theta)-\log\Gamma(k)+U_k(s/\theta)-\frac{x}{\theta}, \qquad \mathbf{4}$$

where $U_k(z)=\log[1+(k-1)/z+(k-1)(k-2)/z^2+\ldots]\rightarrow 0$ for large arguments.

In the last step, the E-values for all high-scoring positions, stored in the priority queue, are calculated and only those with an E-value lower than a given threshold are returned.

Target prediction

The targets of the novel box H/ACA snoRNA candidate are computed using the novel run-time efficient `snoplex` program (H.Tafer, *et al*. in preparation). This tool implements a dynamic programming algorithm to compute the binding energy of the snoRNA sequence to its target together with the energy of the snoRNA structure itself. In order to assess putative binding sites, `snoplex` furthermore considers the initial energy of the snoRNA structure, the energy that is necessary to open the target site and the duplex energy which is also depended on the surrounding snoRNA structure. Given a snoRNA sequence, `snoplex` scans the target RNA sequence and returns the set of thermodynamically most stable interaction structures.

Experimental verification of expression

Our experimental approach is based on (32). Approximately 400 cultured *Trichoplax* animals were collected (Grell strain; Haplotype 1) and small RNAs purified with the mirPremier microRNA Isolation Kit (Sigma), following the protocol for mammalian cell cultures. In the unlikely event that genomic DNA contamination was present in the purified small RNA samples, digestion with DNaseI (Fermentas) was performed following the manufacturer's protocol. A poly-(A) tail of approximately 20 nl was added to the small RNAs using the Poly(A) Tailing Kit (Ambion). Following this, reverse transcription was performed using SuperScript II Reverse Transcriptase (Invitrogen) and a modified poly-d(T) primer (5′-`AAGCAGTGGTATCAACGCAGAGT`$(\mathtt{T})_3$VN). Amplification of small RNAs was accomplished with the use of a universal reverse primer (5′-`AAGCAGTGG TATCAACGCAGAGT`) and forward primers specific to the predicted small RNA of interest. Putative products were cloned into pGEM-T vector (Promega) and positive clones sequenced using the services of Macrogen (Korea). A full list of primers and protocols can be supplied upon request.

RESULTS

tRNAs

The *Trichoplax* genome contains 49 canonical tRNA genes, a single selenocysteine-tRNA gene and one tRNA pseudogene recognizable by `tRNAscan-SE`, Table 1.

Interestingly, the *Trichoplax* genome is essentially devoid of tRNA-like sequences. In addition, a `blast` search revealed a small cluster of four sequences derived from tRNA-Ser(AGA) located just downstream of the functional tRNA on scaffold 3, and a single degraded pseudogene probably derived from tRNA-Leu(TAG) on scaffold 13. These are indicated in parentheses in Table 1.

Ribosomal RNAs

In eukaryotes, rRNAs (except 5S) are processed from a polycistronic 'rRNA operon' which consists of SSU (18S), 5.8S, and LSU (28S) RNAs, two 'internal spacers' ITS-1 and ITS-2, and two 'external spacers', reviewed in (33). *Trichoplax* is no exception, see Figure 2. The rRNA sequences have already received considerable attention in a phylogenetic context, see (1,34–36). The pre-rRNA sequence appears in several copies throughout the genome. Somewhat disappointingly, the `Triad1` assembly contains none of them in complete and uninterrupted form. The consensus sequence of the pre-rRNA can be easily constructed starting from the previously published sequences and the five fairly complete genomic loci [on scaffolds 22, 40 (two), 50 and 734] together with a partial copy on scaffold 34. Only the exact ends of the external transcribed spacers (ETSs) remain uncertain. Figure 2 summarizes the `blastn` matches of the pre-rRNA to the *Trichoplax* genome.

The 5S rRNA sequence of *Trichoplax* has long been known (37). The current genome assembly contains nine 5S RNA genes, one of which is a degraded pseudogene. Interestingly, there are three anti-parallel pairs (two head-to-head and one tail-to-tail which contains the pseudogene).

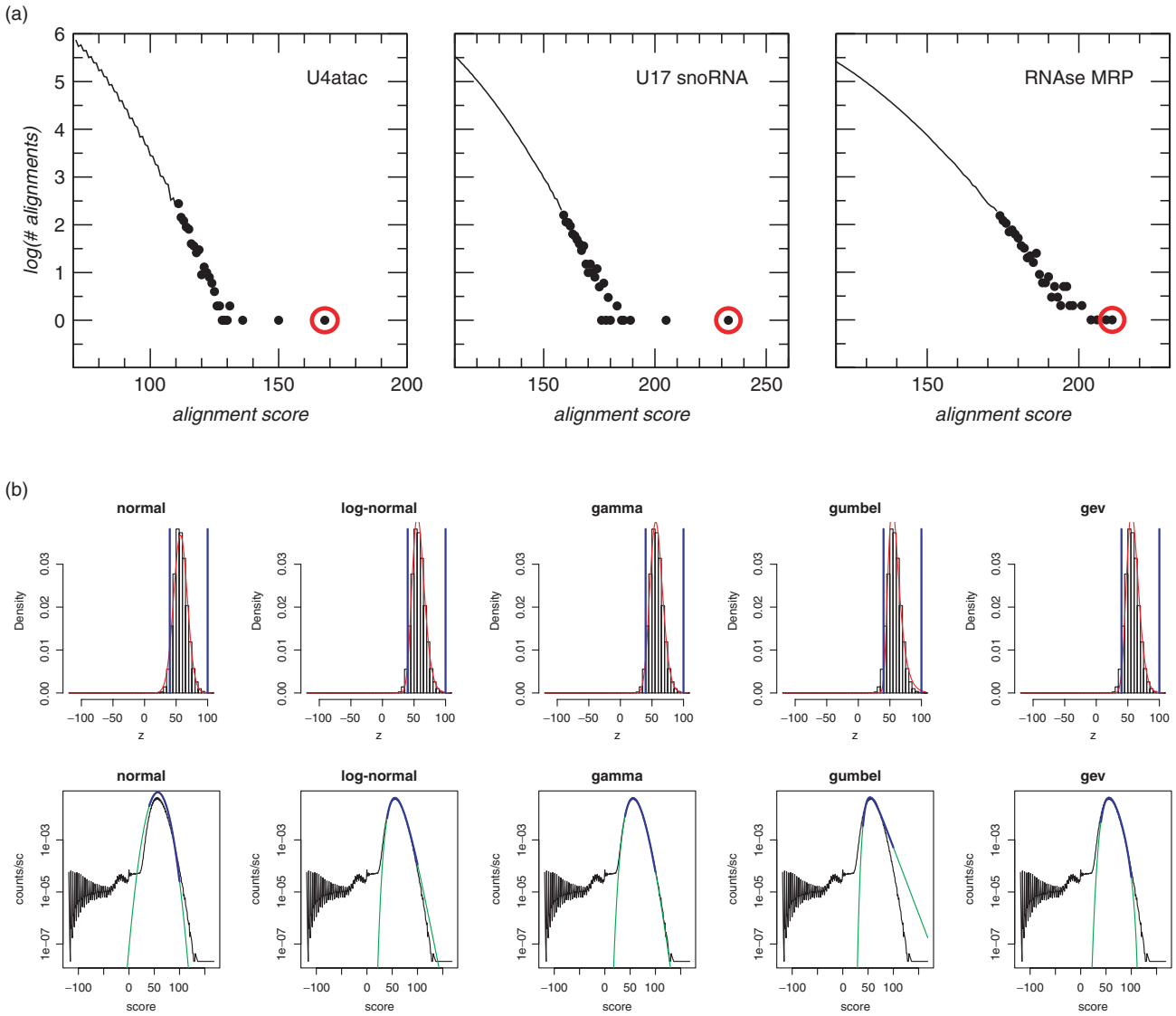

Figure 1. (**a**) Histogram of score distribution for U4atac, U17 and RNase MRP. (**b**) Fitting the `GotohScan` score distribution of U4atac to known density functions.

Spliceosomal snRNAs

Splicing on mRNAs is a common feature to almost all eukaryotic organizms. The spliceosome consists of more than a hundred protein components and five small RNAs that perform crucial catalytic functions, see (38,39) for reviews. The major spliceosome, containing U1, U2, U4, U5 and U6 snRNAs, splices more than 98% of protein coding genes in metazoans, plants and fungi. A small number of protein-coding mRNAs are processed by the minor spliceosome, which contains U11, U12, U4atac, U5 and U6atac snRNAs (40). Previously, nothing was known about placozoan snRNAs. With the exception of the U4atac, the snRNAs were easily found by `blastn`. The U4atac was found by `GotohScan` only with an *E*-value of $9e^{-19}$. The expression of the U4atac was also verified experimentally. With the exception of two U6 genes, each snRNA is encoded by a single gene in the *Trichoplax* genome.

Table 1. Summary of tRNA genes arranged by anti-codon

		First base			
Second base	Third base	A	C	G	T
A	A		Leu[a]	ψ	Leu
	C	Val	Val		Val
	G	Leu + (ψ)	Leu		Leu
	T		Met[2]	Ile	Ile[a]
C	A		Trp	Cys[2]	**SeC**
	C		Gly	Gly	Gly[2]
	G	Arg	Arg[2]		Arg[a]
	T		Arg	Ser	Arg[a]
G	A	Ser + (3ψ)	Ser		Ser
	C	Ala	Ala		Ala
	G	Pro	Pro		Pro
	T	Thr	Thr		Thr
T	A			Thy[a]	
	C		Glu	Asp	Glu[2]
	G		Gln	His	Gln
	T		Lys	Asn	Lys

[a]indicates tRNAs with introns. The multiplicity of genes with more than one copy is indicated by a superscript. **SeC** indicated the selenocysteine tRNA.

Their secondary structures, Figure 3, closely conform to the metazoan consensus (12), with slightly shorter stems II of U11 snRNA and IV of U12 snRNA. The U12 contains an 5 nt insert indicated in red in Figure 3.

In contrast to many other invertebrates, *Trichoplax* snRNAs feature a clearly recognizable proximal sequence element (PSE) see (12,41), which is easily detected by `MEME` (42,43), see Table 2. In line with other species, the PSE element is shared between the pol-II and pol-III transcribed snRNAs. On average the PSE elements differ by 3 nt from the consensus.

RNase P, RNase MRP, SRP RNA

The ribonucleoprotein complexes RNase P and RNase MRP are involved in tRNA and rRNA processing, respectively. Their RNA subunits, which play an essential role in their enzymatic activities, are structurally and evolutionarily related, see e.g. (44,45,46).

RNase P RNA is typically easy to find in genomic DNA, at least within metazoa. The RNase MRP RNA, which is also expected to be present throughout metazoa, is typically much less conserved. Despite substantial efforts (44), RNase MRP RNA homologs have escaped discovery in many bilaterian clades. Not surprisingly, therefore, the *Trichoplax* RNase P RNA was easily identified by `blastn` using the `Rfam` sequences as query. The RNase P sequence is easily verified using `infernal` and the corresponding `Rfam` model.

With standard parameters, `blastn` does not find an MRP RNA homologue. A dedicated, much less stringent, `blastn` search returns two nearly identical candidates. `GotohScan`, on the other hand, easily detects the same two loci. The *E*-value for these two candidates was $3e^{-5}$ and $3e^{-4}$, respectively. The `infernal`-based automatic test for homology to an MRP RNA covariance model provided through the `Rfam` website remained unsuccessful. A manually created alignment containing both metazoan and fungal RNase MRP sequences shows, however, that the *Trichoplax* MRP candidates share the crucial features with both of them, leaving little doubt that we have indeed identified the true MRP sequence. Figure 4 shows the homology-based secondary structure model.

The signal recognition particle (SRP) binds to the signal peptide emerging from the exit tunnel of the ribosome and targets the signal peptide-bearing proteins to the

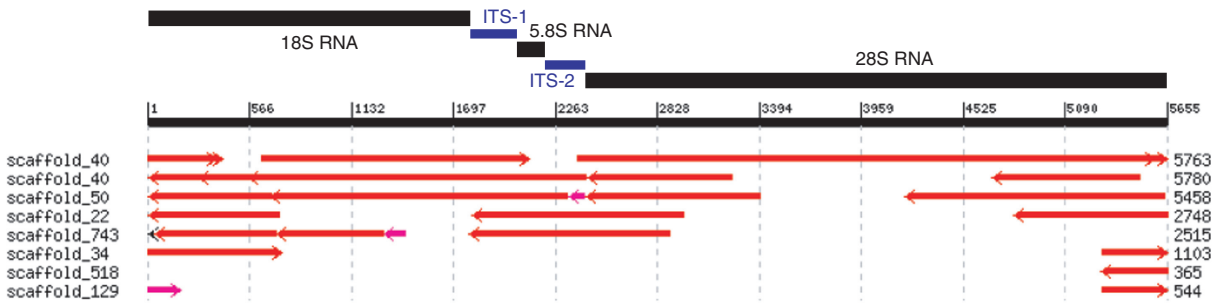

Figure 2. Trichoplax pre-rRNA cluster reconstructed from previously published sequences ***L10828, Z22783, AY652578*** (SSU), ***AY303975, AY652583*** (LSU), ***U65478*** (internal spacers and 5.8S) and `Triad1` genomic sequence. Blast hits of the pre-rRNA to the `Triad1` genome assembly are shown below as in the JGI genome browser.

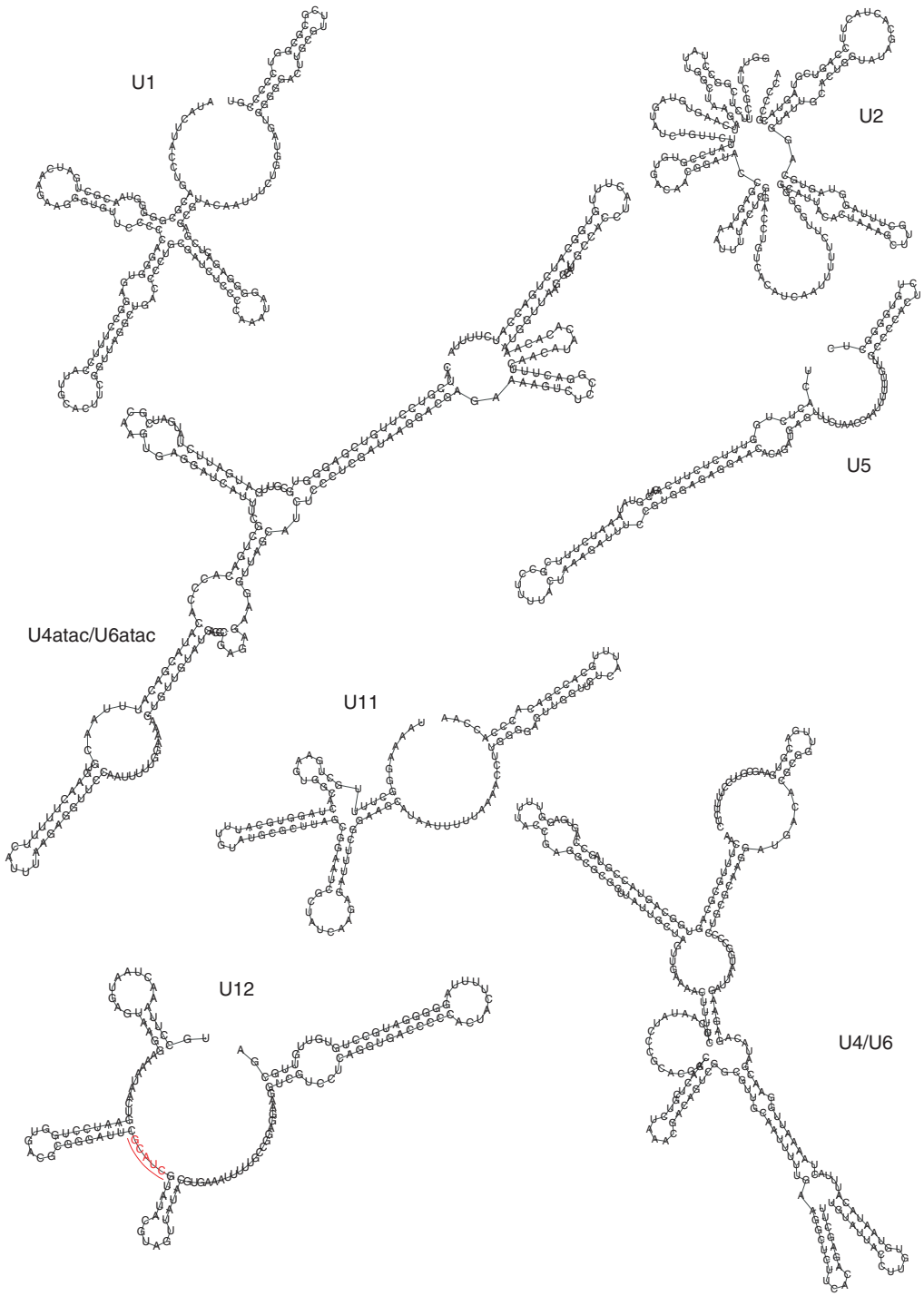

Figure 3. RNA secondary structures of major spliceosomal (U1, U2, U4, U5, U6) and minor spliceosomal (U11, U12, U4atac, U5, U6atac) snRNAs. For U4/U6 and U4atac/U6atac the interaction structures computed by means of `RNAcofold` are shown. The 5 nt insert (relative to other metazoa) is highlighted in the U12.

prokaryotic plasma membrane or the eukaryotic endoplasmic reticulum membrane (47). Its RNA component, called 7SL or SRP RNA, is well conserved and hence easy to identify by `blast` comparison starting from the SRP RNA sequences compiled in the SRPDB (48). The *Trichoplax* SRP RNA is shown in Figure 5.

Small nucleolar RNAs

The two classes of snoRNAs, box H/ACA snoRNAs and box C/D snoRNAs, are mutually unrelated in both their function (directing two different chemical modifications of single residues in their target RNA) and their structure, reviewed e.g. in (49).

Table 2. PSE and location of snRNAs in *T. adhaerens* [The sequence logo was generated using `aln2pattern` (15)]

snRNA	Location	Sequence
U1	−58	.........G...GG.
U2	−55	A......G.G...A..
U4	−57	.........A......
U5	−57	A........G...GC.
U6.1	−62	..T.....AG......
U6.2	−62	..T.....AG......
U4atac	−59	.........AG...C.
U6atac	−63	........AA......
U11	−59	A.......CA...C.G
U12	−60	.......G.G.T.C..
Sequence logo		
Consensus	−59	CCCATAATTRAAGNNA

The U3 snoRNA belongs to the box C/D snoRNA class by virtue of its structural characteristics. It is, however, exceptional in several respects. It contains additional well-conserved sequence motifs which appear to be exclusive to U3 snoRNAs. Instead of directing a modification of single rRNA residues, it is required in the early steps of rRNA maturation, in particular for the cleavage of the 5′ETS and 18S rRNA maturation, see e.g. (50,51,52). Taken together, these features may explain that the U3 snoRNA sequence is much better conserved than all other snoRNAs; in fact, it is the only one that can be found directly by a `blast` search. The candidate sequence was easily verified by `infernal`-alignment to the corresponding `Rfam` model, Figure 5. Its expression was verified experimentally.

The box H/ACA U17 is also involved in the nucleolytic processing of pre-rRNA. Although it has been reported to be the best-conserved box H/ACA snoRNA and ubiquitous among eukaryotes (53), no *Trichoplax* homologue was found using `blast`. Not surprisingly, no other snoRNA homologs were detected by means of `blast`.

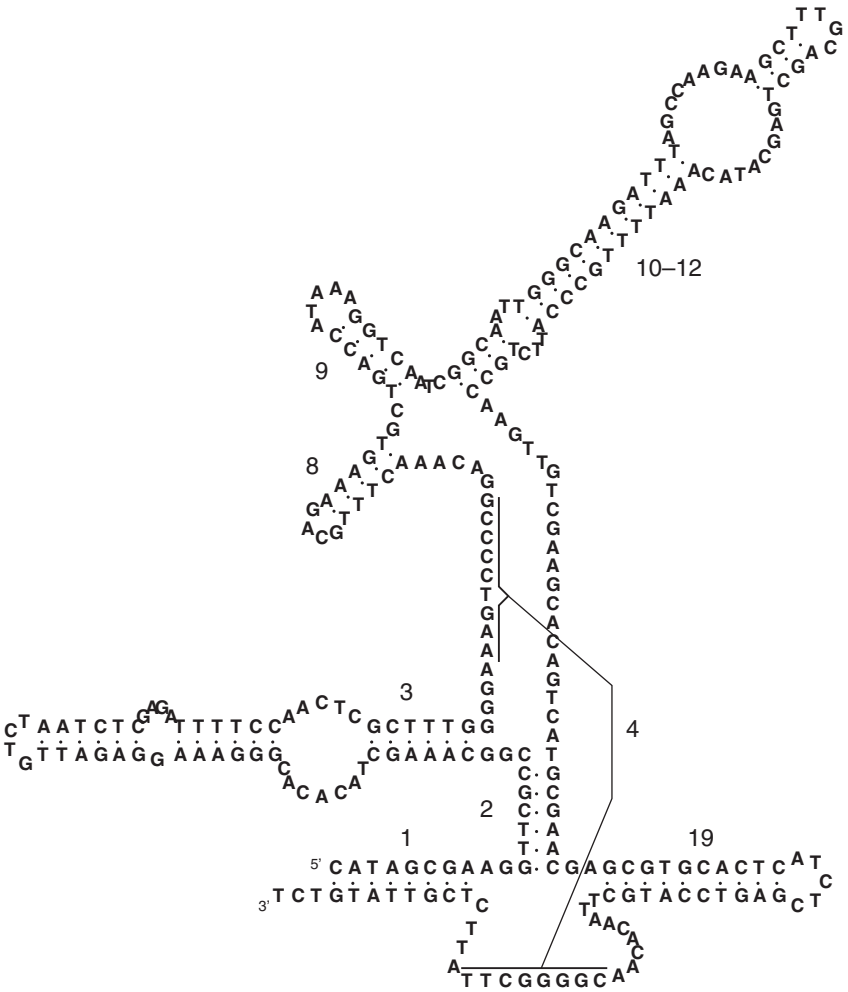

Figure 4. Secondary structure of *T. adhaerens* RNase MRP RNA inferred from the multiple alignment of metazoan RNase MRP RNAs provided in the Electronic Supplement.

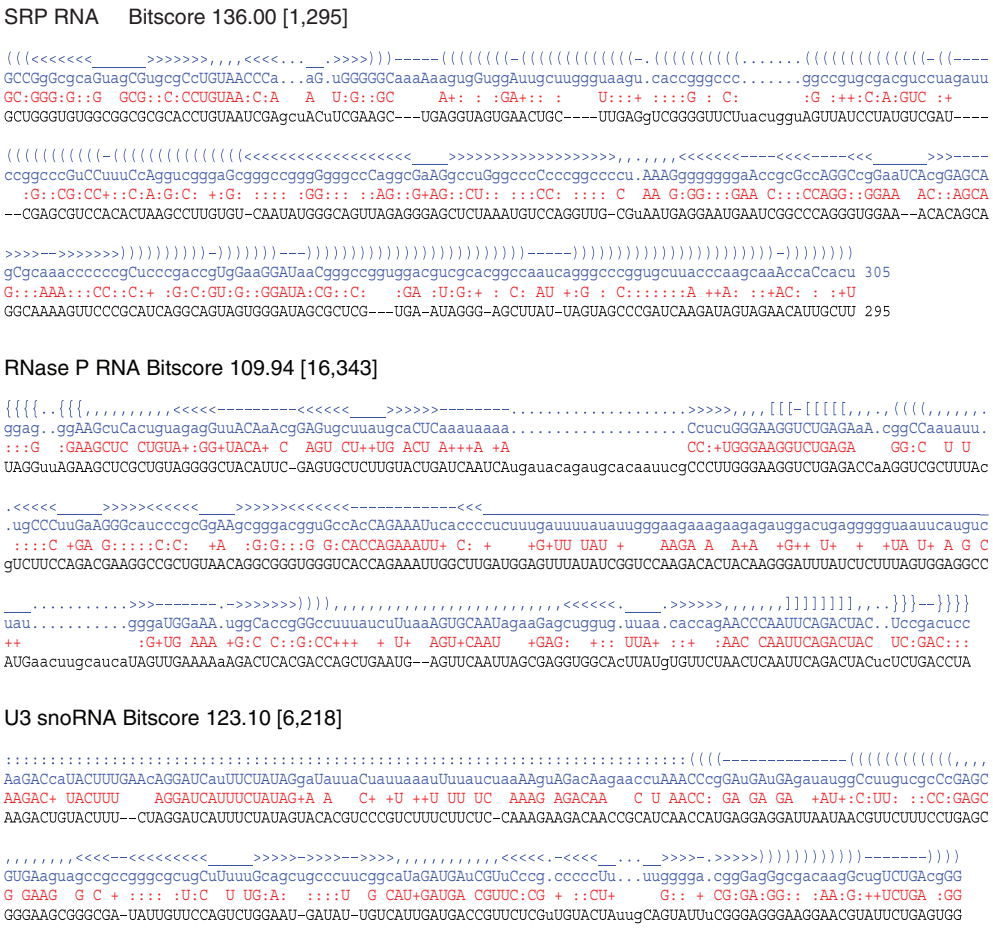

Figure 5. Structural alignments of the *Trichoplax* RNase P RNA, SRP RNA and U3 snoRNA sequences with the corresponding `Rfam` alignments as computed by `infernal`. The first line contains the structure annotation of the two sequences that are aligned in the second and fourth line. Line three in-between describes the sequence conservation and the type of substitution.

The U17 gene was readily identified by `GotohScan` with an *E*-value of $1e^{-16}$, however. We therefore conducted a survey of the *Trichoplax* genome for homologues of all 244 known human box C/D snoRNAs (belonging to 107 distinct snoRNA families) and all 94 known human box H/ACA snoRNAs (82 families) extracted from the `snoRNA-LBME-de` (http://www-snorna.biotoul.fr) (54).

The initial search, for which we used very non-stringent score cut-offs, produced a candidate set of 22 H/ACA and 18 C/D snoRNA. Upon manual inspection, most of these sequences neither fold into secondary structures characteristic for snoRNAs nor match a query sequence unambiguously. Thus we used `SnoReport` (22) to check both secondary structures and sequence motifs. Candidates not recognized by `SnoReport` were removed.

In the next step, we manually added the remaining candidate sequences to multiple sequence alignments of individual snoRNA families. These were retrieved from the `Rfam`, constructed from the sequences provided through the `snoRNA-LBME-db` and (in the case of the U71 snoRNA) compiled from sequences deposited in `Genbank`. This stringent filtering step left 3 H/ACA and 4 C/D snoRNA (not including U3) (Table 3). The multiple sequence alignments, see also Figure 6, are provided in the Electronic Supplement.

Table 3. Small nucleolar RNAs in *Trichoplax*

Name	Class	Target	Conservation	Note
U3	C/D	18S 5-22[b] 18S 1129-1140[b]	Eukaryotes	Verified
U18	C/D	28S A740[b]	Eukaryotes	–
U36	C/D	18S A615[b]	Eukaryotes	–
U76	C/D	28S A1549[b]	Vertebrates	–
U106	C/D	28S A2227	Vertebrates	–
U17	H/ACA	[a]	Eukaryotes	
U71?	H/ACA	–	Vertebrates	Uncertain
sc.3857:103-213(-)	H/ACA	28S U1370 U1884	Novel	

[a]The U17 snoRNA probably targets the 5′ETS, the exact target is still unknown, however (55,53).
[b]Target sites homologous to the ones in human rRNAs.

The U71 candidate shows enough sequence identity with the vertebrates sequences to make its homology with the vertebrate U71 snoRNA very likely; its putative target site, however, is not conserved between human and *Trichoplax*, we thus list it as an uncertain candidate.

Two of the H/ACA candidates were found using ACA1 as query but their homology to the human ACA1 snoRNAs cannot be established. Nevertheless, upon

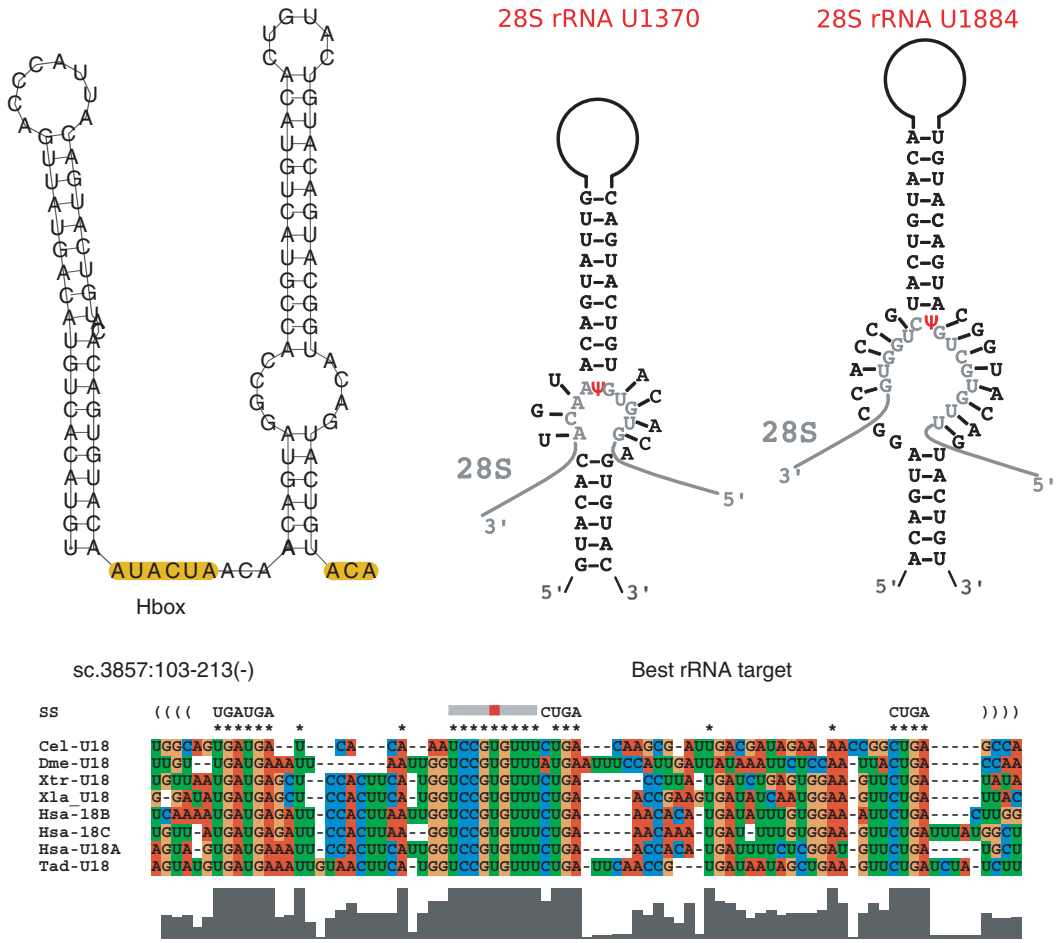

Figure 6. Top: secondary structure model of a novel H/ACA snoRNA (l.h.s.) and the best `snoplex` prediction of its targets sites in the rRNA operon (r.h.s.). Below: alignment of U18 snoRNA sequences from several Metazoa. Boxes and the conserved target binding site are indicated.

inspection, both show all hallmarks of box H/ACA snoRNAs. However, the corresponding primers for the candidate located on scaffold 4365 amplified a sequence fragment located immediately upstream of the predicted snoRNA (see Electronic Supplement for the corresponding alignment). Furthermore, no plausible target site could be identified for this candidate. We therefore did not include it in the list of snoRNAs. The second candidate, 'sc.3857:103-213(-)', on the other hand, exhibits two plausible rRNA targets on 28S rRNA (U1370 and U1884) and most likely constitutes a novel snoRNA. In addition, `snoplex` identifies two additional possible targets in the 18S rRNA (see Electronic Supplement).

This leaves the exceptional U17 snoRNA as the only box H/ACA snoRNA in the *Trichoplax* genome that can be identified unambiguously by computational means. For the three of the four box C/D snoRNA candidates (U18, U36, U76) we find nearly absolute conservation of the target-binding motifs, which are homologous to the corresponding target sites in human. For the U106 snoRNA candidate we can also identify a plausible target site in the 28S rRNA, which, however, is not homologous to that of the human U106 snoRNA.

The putative host genes of the *Trichoplax* snoRNAs are not conserved in human. It is known, however, that snoRNAs can change their genomic location on evolutionary time-scales. For instance, several host gene switches are observed for U17 already within vertebrates (56), see also (57). Furthermore, several human snoRNA host genes are non-coding (e.g. the GAS5 transcript for U76 and the unnamed host gene of U71) or are poorly described ORFs (such as C20orf199 for snoRNA U106), making it virtually impossible to determine whether they are homologous between human and *Trichoplax*.

No microRNAs

Homology based searches for microRNAs remained unsuccessful employing both `blast` and `GotohScan` using the complete set of pre-microRNA hairpins listed in `miRBase` (release 12.0) as query. Both short `blast` hits and weak `GotohScan` signals were analysed. Removing all sequences for which sequence conservation was very poor on the putative mature microRNA sequence and/or the putative precursor did not fold into the characteristic hairpin structure left a single candidate possibly homologous to mir-789. The best-conserved region is located opposite to the annotated mature sequence from *Caenorhabditis* species. Hence, this candidate also remains inconclusive.

Ab initio ncRNA prediction

An alternative to direct homology-based annotation is the *ab initio* prediction of ncRNAs. In particular RNAz (23) has been proved to yield results in wide variety of species, from screens of the human genome compared against (mostly) mammalia (58,59), teleost fishes (60), urochordates (61), nematodes (62), flies (63), yeasts (64) and plasmodium (65). In brief, RNAz is a machine learning tool that determines for a slice of aligned genomic DNA whether it encodes a structured RNA depending on measures of thermodynamics stability and evolutionary conservation (23).

In the case of *Trichoplax*, the use of comparative genomics is limited by the comparably large distance to other sequenced genomes, because most of the genome thus cannot be unambiguously aligned with better understood genomes. We therefore investigated two different genome-wide alignments. In the first screen, we used three species MultiZ-alignments (24) of *T. adhaerens*, and the Cnidaria *Hydra magnipapillata* and *N. vectensis*. We used all alignment blocks containing *Trichoplax* and at least one of the two cnidarians.

A second screen was performed using NcDNAlign alignments (25) constructed from *T. adhaerens*, *P. lobata* and shotgun traces from *A. queenslandica*, *A. millepora*, *A. palmata* and *H. papilata*. This screen was limited to alignment blocks containing *Trichoplax* and at least two other species. As expected, the large evolutionary distances in both screen limit the sensitivity of the comparative approach and preclude the detection of Placozoan-specific ncRNAs.

Both of the differently created alignment sets are screened with RNAz, the corresponding results are compiled in Table 4. The restrictive NcDNalign alignments revealed no novel ncRNAs. Of only 101 loci, 11 were identified as false positives mapping to four different protein-coding gene families, while the remaining hits coincide with ncRNAs that have already been identified by homology-based annotation. With the much more liberal multiz alignments we obtained 3027 RNAz hits comprising 1416 distinct genomic loci that show 'some' sign of evolutionary conserved secondary structure. Of these, 382 loci correspond to annotated ncRNAs, while 1088 (77%) overlap known protein-coding regions or known repetitive elements. Twelve of the remaining loci are supported by ESTs and may constitute novel ncRNAs. The remaining 193 hits contain the U3 and U17 snoRNA genes, which were found by blast and/or GotohScan.

Figure 7 summarizes the distribution of the RNAz classification scores of the MultiZ-based screen. Many of the known ncRNAs appear with moderate classification probability, with a significant enrichment observed only for scores close to one. This reflects the high expected FDR of these data, which are largely based on pairwise alignments. This implies that the initial candidates of this screen need to be post-processed with respect to gene annotation and/or other filtering methods. Indeed, the majority of predictions—even somewhat more than the estimated FDR—are located in the protein-coding regions (Table 4). The data nevertheless provide at least statistical evidence for a set of about 100–200 novel structured RNA elements.

Table 4. RNAz screens of *T. adhaerens* genome

	multiz	NcDNalign	Known
Aligned DNA (nt)	4 837 148	1 35 140	–
alignments	35 039	744	–
RNAz $P > 0.5$	1416	101	–
FDR random	56% (797)	43% (43)	–
RNAz $P > 0.9$	751	79	–
FDR	27% (386)	15% (15)	–
tRNAs	39	35	50 + 1
5S rRNA	6	8	9
rRNA operon	33 + 3	43	–[a]
snRNAs	6	4	10
MRP, P, 7SL	1	0	3
Protein coding	1022	11	96 963
Repeat elements	66	1	–
Total annotated	1211	101	–
Unannotated with EST	12	0	–
Without annotation	205	0	–

[a]The rDNA operons appear as series of multiple RNAz hits. *Known* refers to all ncRNAs that have been reported previously and those that have been identified by homology search in this study.

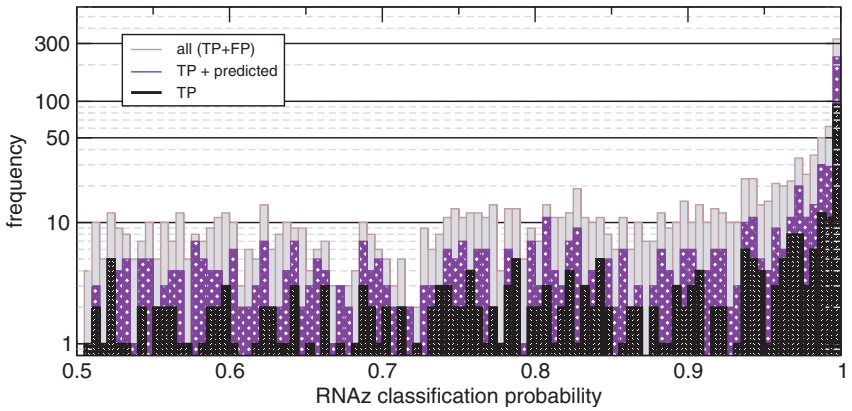

Figure 7. Distribution of RNAz classification score for known (true positive) (black) all predictions (grey), and only those that are identified as coding or repetitive (maroon). Note the logarithmic scale: there are more than 100 non-annotated predictions with a classification confidence above 99%.

The 744 NcDNalign were searched with RNAmicro for possible microRNAs. After removing known ncRNAs, in particular the U5 snRNA and several hits to hairpins in the rRNA operon, exons of annotated protein coding genes and repetitive elements recognized by repeatmasker, we retained 82 candidates. Since RNAmicro evaluates alignment and the corresponding consensus fold, we also checked whether the *Trichoplax* candidate sequences alone fold into a microRNA-like hairpin structure. Sixty-four sequences passed this filter. Most of these sequences appear to be repetitive, mapping to more than three distinct loci in the *Trichoplax* genome, leaving 13 microRNA-like hairpins that are conserved between *Trichoplax* and *Nematostella*. However, none of these candidates resembles any of the 40 in *N. vectensis* or the eight *A. queenslandica* microRNAs described in (66). We thus suggest that these conserved hairpins are not microRNAs. Instead they might belong to a previously undescribed class of hairpin structures.

DISCUSSION

We have reported here on a comprehensive computational study of non-protein-coding RNA genes in the genome of the placozoan *T. adhaerens*. We observed that only a limited set of the best-conserved ncRNAs, in particular tRNAs, rRNAs and a few additional 'housekeeping' RNAs are readily found by means of blastn. We have therefore developed a more sensitive tool, GotohScan, which implements a full semi-global dynamic programming algorithm. Using this method, we were able to detect homologs of several fast-evolving ncRNAs, including a few box C/D and box H/ACA snoRNAs, the RNase MRP RNA, and the full complement of spliceosomal snRNAs.

In addition to the homology-based annotation, we conducted surveys evolutionary conserved RNA secondary structures using RNAz and RNAmicro. Reasoned by the large evolutionary distance between *Trichoplax* and other sequenced genomes, the sensitivity of these screens was rather low, however. Nevertheless a handful of novel ncRNA candidates was found.

Due to the small size and slow growth of *T. adhaerens*, it is hard—if not impossible—to obtain sufficient amounts of RNAs to verify the expression of ncRNA candidates directly by Northern blots. Instead, we used here a PCR-based approach introduced by (32), which requires much smaller quantities of RNA. We did not attempt to validate the entire set of predictions but rather selected a small subset, consisting of a few of the homologs detected by GotohScan and a small collection of novel predictions. Due to the small amount of RNA, the sensitivity is still limited. Nevertheless, we unambiguously identified a few previously undescribed *Trichoplax* ncRNAs, namely: U4atac, as a representative of the minor spliceosome; the U3 snoRNA and a putative novel ncRNA on scaffold 3857.

Our computational annotation of the *Trichoplax* genome reveals much of the expected complement of the ncRNA repertoire. Most ncRNAs are single-copy genes or appear in very small copy numbers. This contrasts the situation in many of the higher metazoa, for which more detailed ncRNA annotations are available [e.g. *Caenorhabditis elegans* (67), *Drosophila* (63,68) and the Rfam-based annotation in mammalian genomes]. In particular, the small copy number of tRNAs and other pol-III transcripts is surprising, since these genes appear in dozens or hundreds of copies in many bilaterian genomes.

The lack of microRNAs is surprising at a first glance. While a few orthologous microRNAs—in particular the mir-100 family—are shared between Cnidaria and Bilateria (69,70), we found no trace of these genes in *Trichoplax*. Neither did we find a homolog of one of the eight sponge microRNAs (66). Our analysis is thus consistent with the recent report based on short RNA sequencing (66) that *Trichoplax* does not have microRNAs. The continuing expansion of the repertoire of microRNA and their targets has been associated with both major body-plan innovations as well as the emergence of phenotypic variation in closely related species (71,69–73). The microRNA precursors of Cnidaria and Bilateria are imperfectly paired hairpin structures about 80 nt in length. In contrast, the precursors of the recently discovered miRNAs of the sponge *A. queenslandica* (66) are not orthologous to any of the Cnidarian/Bilaterian microRNA families and resemble the structurally more diverse and more complex RNAs described in slime-molds (74), algae (75,76) and plants (77–79). Under the hypothesis of monophyletic diploplasts, which has recently gained substantial support (5,80), Placozoa have secondarily lost their ability to produce microRNAs, while sponges have secondarily relaxed the constraints on precursor structures. The complete loss of microRNAs in Placozoa is consistent with the morphological simplicity of *Trichoplax*. Although, argonaute, Dicer and Drosha proteins could be found in *Trichoplax*, no Pasha homolog, which partners with Drosha during miRNA biogenesis, was found. Since, all these core RNAi proteins, except Pasha, are also involved in non-miRNA related processes, it is likely that Pasha has been discarded together with the miRNAs in *Trichoplax* (66).

De novo predictions of evolutionarily conserved RNAs suggest that the *Trichoplax* genome may have preserved some ncRNAs characteristic to basal metazoans, such as the handful of hairpin structures that are conserved between *Trichoplax* and *Nematostella*. We do not know at this point, however, whether these purely computational signals are expressed *in vivo*, and what their function might be.

Our survey also misses several ncRNA classes that we should expect to be present in *Trichoplax*, in particular telomerase RNA, U7 snRNA [which are involved in histone 3′-end processing (81)], the Ro-associated Y-RNAs, the RNA components of the vault complex (the *Trichoplax* genome contains the Major Vault Protein) and possibly also a 7SK RNA. In contrast to microRNAs, however, recent studies have highlighted how difficult it is to identify these particular classes of RNA from genomic DNA: telomerase RNA evolves so rapidly that—despite its size of over 300 nt—it has not been identified so far in any invertebrate species (82).

A similarly fast evolution is observed for the 7SK RNA (83,84). Due to their small size and weak sequence constraints, U7 snRNA (85,86), Y RNAs (87,88) and vault RNAs [P. F. Stadlex *et al*. (submitted for publication)] are also largely unknown beyond deuterostomes (in some cases Drosophilids or *C. elegans*, where homologs were discovered independently). Our failure to find these genes thus most likely points at the limitations of the currently available homology search methodology rather than at the absence of these RNA classes in the *Trichoplax* genome.

SUPPLEMENTAL INFORMATION

An Electronic Supplement provides a complete set of coordinates of all described putative RNA elements, alignments of snoRNAs, RNase MRP and genomic locations of the snoRNA targets. The data can be accessed in machine readable formats at http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/08-024/.

FUNDING

Deutsche ForschungsGemeinschaft (through the 'Graduierten-Kolleg Wissensrepräsentation' University of Leipzig); Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung* (project P19411 'Genomdynamik'); sixth Framework Programme of the European Union (projects SYNLET and EMBIO); Alexander von Humboldt Foundation and John Templeton Foundations, and an Alexander von Humboldt Research Fellowship (D.dJ.). Funding for open access charge: Deutsche Forschungsgemeinschaft (DFG).

Conflict of interest statement. None declared.

REFERENCES

1. Voigt,O., Collins,A.G., Pearse,V.B., Pearse,J.S., Ender,A., Hadrys,H. and Schierwater,B. (2004) Placozoa—no longer a phylum of one. *Curr. Biol.*, **14**, R944–R945.
2. Syed,T. and Schierwater,B. (2002) *Trichoplax adhaerens*: discovered as a missing link, forgotten as a hydrozoan, re-discovered as a key to metazoan evolution. *Vie et Milieu*, **52**, 177–187.
3. Collins,A.G., Cartwright,P., McFadden,C.S. and Schierwater,B. (2005) Phylogenetic context and basal metazoan model systems. *Integr. Compar. Biol.*, **45**, 585–594.
4. Miller,D.J. and Ball,E.E. (2008) Animal evolution: *Trichoplax*, trees and taxonomic turmoil. *Curr. Biol.*, **18**, R1003–R1005.
5. Schierwater,B., Eitel,M., Jakob,W., Osigus,H.-J., Hadrys,H., Dellaporta,S., Kolokotronis,S.-O. and DeSalle,R. (2008) Concatenated molecular and morphological analysis sheds light on early metazoan evolution and fuels a modern 'Urmetazoon' hypothesis. *PLoS Biol.*, **7**.
6. Jakob,W., Sagasser,S., Dellaporta,S., Holland,P., Kuhn,K. and Schierwater,B. (2004) The Trox-2 Hox/ParaHox gene of *Trichoplax* (Placozoa) marks an epithelial boundary. *Dev. Genes Evol.*, **214**, 170–175.
7. Srivastava,M., Begovic,E., Chapman,J., Putnam,N.H., Hellsten,U., Kawashima,T., Kuo,A., Mitros,T., Carpenter,M.L., Signorovitch,A.Y. *et al*. (2008) The Trichoplax genome and the nature of placozoans. *Nature*, **454**, 955–960.
8. Lowe,T.M. and Eddy,S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, **25**, 955–964.
9. Putnam,N.H., Srivastava,M., Hellsten,U., Dirks,B., Chapman,J., Salamov,A., Terry,A., Shapiro,H., Lindquist,E., Kapitonov,V. *et al*. (2007) Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*, **317**, 86–94.
10. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
11. He,S., Liu,C., Skogerbø,G., Zhao,H., Wang,J., Liu,T., Bai,B., Zhao,Y. and Chen,R. (2008) NONCODE v2.0: decoding the non-coding. *Nucleic Acids Res.*, **36**, D170–D172.
12. Marz,M., Kirsten,T. and Stadler,P.F. (2008) Evolution of spliceosomal snRNA genes in metazoan animals. *J. Mol. Evol.*; Nov 22, doi: 10.1007/s00239-008-9149-6 [Epub ahead of print].
13. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
14. Nawrocki,E.P. and Eddy,S.R. (2007) Query-dependent banding for faster RNA similarity searches. *PLoS Comp. Biol.*, **3**, e56.
15. Mosig,A., Chen,J.L. and Stadler,P.F. (2007) Homology search with fragmented nucleic acid sequence patterns. In Giancarlo,R. and Hannenhalli,S. (eds), *Algorithms in Bioinformatics (WABI 2007)*. Vol. 4645 of *Lecture Notes in Computer Science*, Springer Verlag, Berlin, pp. 335–345.
16. Griffiths-Jones,S. (2005) RALEE—RNA alignment editor in Emacs. *Bioinformatics*, **21**, 257–259.
17. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,S.L., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, **125**, 167–188.
18. Hofacker,I.L., Fekete,M. and Stadler,P.F. (2002) Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, **319**, 1059–1066.
19. Bernhart,S.H., Hofacker,I.L., Will,S., Gruber,A.R. and Stadler,P.F. (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474.
20. Bernhart,S.H., Tafer,H., Mückstein,U., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2006) Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol. Biol.*, **1**, 3.
21. Hertel,J. and Stadler,P.F. (2006) Hairpins in a haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics*, **22**, e197–e202.
22. Hertel,J., Hofacker,I.L. and Stadler,P.F. (2008) snoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics*, **24**, 158–164.
23. Washietl,S., Hofacker,I.L. and Stadler,P.F. (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 2454–2459.
24. Blanchette,M., Kent,W.J., Riemer,C., Elnitski,L., Smit,A.,F.A., Roskin,K.M., Baertsch,R., Rosenbloom,K., Clawson,H., Green,E.D. *et al*. (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.
25. Rose,D., Hertel,J., Reiche,K., Stadler,P.F. and Hackermüller,J. (2008) NcDNAlign: plausible multiple alignments of non-protein-coding genomic sequences. *Genomics*, **92**, 65–74.
26. Pearson,W.R. (1991) Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, **11**, 635–650.
27. Roshan,U., Chikkagoudar,S. and Livesay,D.R. (2008) Searching for evolutionary distant RNA homologs within genomic sequences using partition function posterior probabilities. *BMC Bioinformatics*, **9**, 61.
28. Gotoh,O. (1982) An improved algorithm for matching biological sequences. *J. Mol. Biol.*, **162**, 705–708.
29. Basu,K., Sriraam,N. and Richard,R.J. (2007) A pattern matching approach for the estimation of alignment between any two given DNA sequences. *J. Med. Syst.*, **31**, 247–253.
30. Kann,M.G., Sheetlin,S.L., Park,Y., Bryant,S.H. and Spouge,J.L. (2007) The identification of complete domains within protein sequences using accurate e-values for semi-global alignment. *Nucleic Acids Res.*, **35**, 4678–4685.
31. Davis,P.J. (1964) Gamma function and related function. In: Abramowitz,M. and Stegun,I.A. (eds), *Handbook of Mathematical Functions*, National Bureau of Standards, Washington, DC, pp. 253–266.

32. Ro,S., Park,C., Jin,J., Sanders,K.M. and Yan,W. (2006) A PCR-based method for detection and quantification of small RNASs. *Biochem. Biophys. Res. Comm.*, **351**, 756–763.
33. Nazar,R.N. (2004) Ribosomal RNA processing and ribosome biogenesis in eukaryotes. *IUBMB Life*, **56**, 457–465.
34. Wainright,P.O., Hinkle,G., Sogin,M.L. and Stickel,S.K. (1993) The monophyletic origins of the metazoa; an unexpected evolutionary link with fungi. *Science*, **260**, 340–342.
35. Odorico,D.M. and Miller,D.J. (1997) Internal and external relationships of the Cnidaria: implications of primary and predicted secondary structure of the 5′-end of the 23S-like rDNA. *Proc. R. Soc. Lond. B Biol. Sci.*, **264**, 77–82.
36. da Silva,F.B., Muschner,V. and Bonatto,S.L. (2007) Phylogenetic position of Placozoa based on large subunit (LSU) and small subunit (SSU) rRNA genes. *Genet. Mol. Biol.*, **30**, 127–132.
37. Val'ekho-Roman,K.M., Bobrova,V.K., Troitskiĭ,AV., Tsetlin,A.B. and Okshteĭn,I.L. (1990) New data on Trichoplax: the nucleotide sequence of 5S rRNA. *Dokl Akad Nauk SSSR*, **311**, 500–503.
38. Nilsen,T.W. (2003) The spliceosome: the most complex macromolecular machine in the cell? *Bioessays*, **25**, 1147–1149.
39. Valadkhan,S. (2007) The spliceosome: caught in a web of shifting interactions. *Curr. Opin. Struct. Biol.*, **17**, 310–315.
40. Tarn,W.Y., Yario,T.A. and Steitz,J.A. (1995) U12 snRNAs in vertebrates: Evolutionary conservation of 5′ sequences implicated in splicing of premRNAs containing a minor class of introns. *RNA*, **1**, 644–656.
41. Hernandez,N. (2001) Small nuclear RNA genes: a model system to study fundamental mechanisms of transcription. *J. Biol. Chem.*, **276**, 26733–26736.
42. Bailey,T.L. and Elkan,C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, AAAI Press, Menlo Park, CA, pp. 28–36.
43. Bailey,T.L., Williams,N., Misleh,C. and Li,W.W. (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, **34**, W369–W373.
44. Piccinelli,P., Rosenblad,M.A. and Samuelsson,T. (2005) Identification and analysis fo ribonuclease P and MRP RNA in a broad range of eukaryotes. *Nucleic Acids Res.*, **33**, 4485–4495.
45. Woodhams,M.D., Stadler,P.F., Penny,D. and Collins,L.J. (2007) RNAse MRP and the RNA processing cascade in the eukaryotic ancestor. *BMC Evol. Biol.*, **7**, S13.
46. Willkomm,D.K. and Hartmann,R.K. (2007) An important piece of the RNase P jigsaw solved. *Trends Biochem. Sci.*, **32**, 247–250.
47. Nagai,K., Oubridge,C., Kuglstatter,A., Menichelli,E., Isel,C. and Jovine,L. (2003) Structure, function and evolution of the signal recognition particle. *EMBO J.*, **22**, 3479–3485.
48. M Alm Rosenblad, Gorodkin,J., Knudsen,B., Zwieb,C. and Samuelsson,T. (2003) SRPDB (signal recognition particle database). *Nucleic Acids Res.*, **31**, D363–D364.
49. Bachellerie,J.-P., Cavaillé,J. and A Hüttenhofer. (2002) The expanding snoRNA world. *Biochimie*, **84**, 775–790.
50. Gerbi,S.A., Borovjagin,A.V., Ezrokhi,M. and Lange,T.S. (2001) Ribosome biogenesis: role of small nucleolar RNA in maturation of eukaryotic rRNA. *Cold Spring Harbor Symp. Quant. Biol.*, **LXVI**, 575–590.
51. Marmier-Gourrier,N., Cléry,A., Senty-Ségault,V., Charpentier,B., Schlot-ter,F., Leclerc,F., Fournier,R. and Branlant,C. (2003) A structural, phylogenetic and functional study of 15.5-kD/Snu13 protein binding on U3 small nucleolar RNA. *RNA*, **9**, 821–838.
52. Cléry,A., Senty-Ségault,V., Leclerc,F., Raué,H.A. and Branlant,C. (2007) Analysis of sequence and structural features that identify the B/C motif of U3 small nucleolar RNA as the recognition site for the Snu13p-Rrp9p protein pair. *Mol. Cellular Biol.*, **27**, 1191–1206.
53. Atzorn,V., Fragapane,P. and Kiss,T. (2004) U17/snR30 is a ubiquitous snoRNA with two conserved sequence motifs essential for 18S rRNA production. *Mol. Cell Biol.*, **24**, 1769–1778.
54. Lestrade,L. and Weber,M.J. (2006) snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nuclcie Acids Res.*, **34**, D158–D162.
55. Enright,C.A., Maxwell,E.S., Eliceiri,G.L. and Sollner-Webb,B. (1996) 5′ ETS rRNA processing facilitated by four small RNAs: U14, E3, U17 and U3. *RNA*, **2**, 1094–1099.
56. Bompfünewerer,A.F., Flamm,C., Fried,C., Fritzsch,G., Hofacker,I.L., Lehmann,J., Missal,K., Mosig,A., B Müller, Prohaska,S.J. *et al.* (2005) Evolutionary patterns of non-coding RNAs. *Theorc. Biosci.*, **123**, 301–369.
57. Weber,M.J. (2006) Mammalian small nucleolar RNAs are mobile genetic elements. *PLoS Genet.*, **2**, e205.
58. Washietl,S., Hofacker,I.L., Lukasser,M., Hüttenhofer,A. and Stadler,P.F. (2005) Mapping of conserved RNA secondary structures predicts thousands of functional non-coding RNAs in the human genome. *Nature Biotech.*, **23**, 1383–1390.
59. Washietl,S., Pedersen,J.S., Korbel,J.O., Gruber,A., J Hackermüller, Hertel,J., Lindemeyer,M., Reiche,K., Stocsits,C., Tanzer,A. *et al.* (2007) Structured RNAs in the ENCODE selected regions of the human genome. *Gen. Res.*, **17**, 852–864.
60. Rose,D., J Jöris, J Hackermüller, Reiche,K., Li,Q. and Stadler,P.F. (2008) Duplicated RNA genes in teleost fish genomes. *J. Bioinf. Comp. Biol.*, **6**, 1157–1175.
61. Missal,K., Rose,D. and Stadler,P.F. (2005) Non-coding RNAs in *Ciona intestinalis*. *Bioinformatics*, **21**(S2), i77–i78, [Proceedings ECCB/JBI'05, Madrid].
62. Missal,K., Zhu,X., Rose,D.,WDeng, G Skogerbø, Chen,R. and Stadler,P.F. (2006) Prediction of structured non-coding RNAs in the genome of the nematode Caenorhabitis elegans. *J. Exp. Zool. Mol. Dev. Evol.*, **306B**, 379–392.
63. Rose,R.D., Hackermüller,J., Washietl,S., Findeiß,S., Reiche,K., Hertel,J., Stadler,P.F. and Prohaska,S.J. (2007) Computational RNomics of drosophilids. *BMC Genomics*, **8**, 406.
64. Steigele,S., Huber,W., Fried,C., Stadler,P.F. and Nieselt,K. (2007) Comparative analysis of structured RNAs in S. cerevisiae indicates a multitude of different functions. *BMC Biol.*, **5v**, 25.
65. Mourier,T., Carret,C., Kyes,S., Christodoulou,Z., Gardner,P.P., Jeffares,D.C., Pinches,R., Barrell,B., Berriman,M., Griffiths-Jones,S. *et al.* (2008) Genome-wide discovery and verification of novel structured RNAs in Plasmodium falciparum. *Genome Res.*, **18**, 281–292.
66. Grimson,A., Srivastava,M., Fahey,B., Woodcroft,B.J., Chiang,H.R., King,N., Degnan,A.M., Rokhsar,D.S. and Bartel,D.P. (2008) Early origins and evolution of miRNAs and Piwi-interacting RNAs in animals. *Nature*, **455**, 1193–1197.
67. Stricklin,S.L., Griffiths-Jones,S. and Eddy,S.R. (2005) C. elegans noncoding RNA genes. *WormBook*, http://www.wormbook.org/chapters/www_noncodingRNA/noncodingRNA.html. Last acessed date, October 8, 2008.
68. Stark,A., Kheradpour,P., Parts,L., Brennecke,J., Hodges,E., Hannon,G.J. and Kellis,M. (2007) Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes. *Genome Res.*, **17**, 1865–1879.
69. Sempere,L.F., Cole,C.N., McPeek,M.A. and Peterson,K.J. The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *J. Exp. Zool. B. Mol. Dev. Evol.*, **306B**, 575–588.
70. Prochnik,S.E., Rokhsar,D.S. and Aboobaker,A.A. (2007) Evidence for a microRNA expansion in the bilaterian ancestor. *Dev. Genes Evol.*, **217**, 73–77.
71. Hertel,J., Lindemeyer,M., Missal,K., Fried,C., Tanzer,A., Flamm,C., Hofacker,I.L., Stadler,P.F. and The Students of Bioinformatics Computer Labs 2004 and 2005. (2006) The expansion of the metazoan microRNA repertoire. *BMC Genomics*, **7**, 15.
72. Niwa,R. and Slack,F.J. (2007) The evolution of animal microRNA function. *Curr. Opin. Gen. Devel.*, **17**, 145–150.
73. Lee,C.T., Risom,T. and Strauss,W.M. (2007) Evolutionary conservation of microRNA regulatory circuits: an examination of microRNA gene complexity and conserved microRNA-target interactions through metazoan phylogeny. *DNA Cell Biol.*, **26**, 209–218.
74. Hinas,A., Reimegard,J.,, Wagner,E.G., Nellen,W., Ambros,V. and Söderbom,F. (2007) The small RNA repertoire of *Dictyostelium discoideum* and its regulation by components of the RNAi pathway. *Nucleic Acids Res.*, **35**, 6714–6726.
75. Zhao,T., Li,G., Mi,S., Li,S., Hannon,G.J., Wang,X.J. and Qi,Y. (2007) A complex system of small RNAs in the unicellular green alga *Chlamydomonas reinhardtii*. *Genes Dev.*, **21**, 1190–1203.

76. A Molnár, Schwach,F., Studholme,D.J., Thuenemann,E.C. and Baulcombe,D.C. (2007) miRNAs control gene expression in the single-cell alga *Chlamydomonas reinhardtii*. *Nature*, **447**, 1126–1129.
77. Zhang,B., Pan,X., Cannon,C.H., Cobb,G.P. and Anderson,T.A. (2006) Conservation and divergence of plant microRNA genes. *Plant J.*, **46**, 243–259.
78. Axtell,M.J., Snyder,J.A. and Bartel,D.P. (2007) Common functions for diverse small RNAs of land plants. *Plant Cell*, **19**, 1750–1769.
79. Sunkar,R. and Jagadeeswaran,G. (2008) *In silico* identification of conserved microRNAs in large number of diverse plant species. *BMC Plant Biol.*, **8**, 37.
80. Dunn,C.W., Hejno,A., Matus,D.Q., Pang,K., Browne,W., Smith,S.A., Seaver,E., Rouse,G.W., Obst,M., Edgecombe,G.D. *et al.* (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, **452**, 745–749.
81. Marzluff,W.F. (2005) Metazoan replication-dependent histone mRNAs: a distinct set of RNA polymerase II transcripts. *Curr. Opin. Cell. Biol.*, **17**, 274–280.
82. Xie,M., Mosig,A., Qi,X., Li,Y., Stadler,P.F. and Chen,J.J.-L. (2008) Size variation and structural conservation of vertebrate telomerase RNA. *J. Biol. Chem.*, **283**, 2049–2059.
83. Gruber,A.R., Koper-Emde,D., Marz,M., Tafer,H., Bernhart,S., Obernosterer,G., Mosig,A., Hofacker,I.L., Stadler,P.F. and Benecke,B.-J. (2008) Invertebrate 7SK snRNAs. *J. Mol. Evol.*, **66**, 107–115.
84. Gruber,A., Kilgus,C., Mosig,A., Hofacker,I.L., Hennig,W. and Stadler,P.F. (2008) Arthropod 7SK RNA. *Mol. Biol. Evol.*, **25**, 1923–1930.
85. Marz,M., Mosig,A., Stadler,B.M.R. and Stadler,P.F. (2007) U7 snRNAs: A computational survey. *Geno. Prot. Bioinf.*, **5**, 187–195.
86. López,M.D. and Samuelsson,T. (2008) Early evolution of histone mRNA 3′ end processing. *RNA*, **14**, 1–10.
87. Mosig,A., Guofeng,M., Stadler,B.M.R and Stadler,P.F. (2007) Evolution of the vertebrate Y RNA cluster. *Theorc. Biosci.*, **126**, 9–14.
88. Perreault,J., Perreault,J.-P. and Boire,G. The Ro associated Y RNAs in metazoans: evolution and diversification. *Mol. Biol. Evol.*, **24**, 1678–1689.