

Non-coding RNA Annotation of the Genome of *Trichoplax adhaerens*

Jana Hertel
Danielle de Jong
Manja Marz
Dominic Rose
Hakim Tafer, et al.

SFI WORKING PAPER: 2009-04-011

SFI Working Papers contain accounts of scientific work of the author(s) and do not necessarily represent the views of the Santa Fe Institute. We accept papers intended for publication in peer-reviewed journals or proceedings volumes, but not papers that have already appeared in print. Except for papers by our external faculty, papers must be based on work done at SFI, inspired by an invited visit to or collaboration at SFI, or funded by an SFI grant.

©NOTICE: This working paper is included by permission of the contributing author(s) as a means to ensure timely distribution of the scholarly and technical work on a non-commercial basis. Copyright and all rights therein are maintained by the author(s). It is understood that all persons copying this information will adhere to the terms and constraints invoked by each author's copyright. These works may be reposted only with the explicit permission of the copyright holder.

www.santafe.edu



SANTA FE INSTITUTE

Non-Coding RNA Annotation of the Genome of *Trichoplax adhaerens*

Jana Hertel^a, Danielle de Jong^b, Manja Marz^a,
Dominic Rose^a, Hakim Tafer^c, Andrea Tanzer^{a,c,d},
Bernd Schierwater^b, Peter F. Stadler^{a,e,c,f,*}

^aBioinformatics Group, Dept. of Computer Science, Interdisciplinary Center for
Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig,
Germany

{*manja, dominic, jana, andrea, studla*}@*bioinf.uni-leipzig.de*

^bDivision of Ecology and Evolution, Institut für Tierökologie und Zellbiologie,
Tierärztliche Hochschule Hannover, Bünteweg 17d, D-30559 Hannover, Germany
{*danielle.dejong, bernd.schierwater*}@*ecolevol.de*

^cDepartment of Theoretical Chemistry, University of Vienna, Währingerstraße 17,
A-1090 Wien, Austria

{*at, htafer, studla*}@*tbi.univie.ac.at*, *at@tbi.univie.ac.at*

^dDepartment of Ecology and Evolutionary Biology, Yale University, New Haven,
CT 06520, USA

^eRNomics Group, Fraunhofer Institut für Zelltherapie und Immunologie,
Deutscher Platz 5e, D-04103 Leipzig, Germany

^fSanta Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA

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.

Key words: *Trichoplax adhaerens*, non-coding RNA, genome annotation,
homology search, spliceosomal RNAs, microRNAs

1 Introduction

The phylum Placozoa consists of only one recognised species – the marine dweller *Trichoplax adhaerens*. Extensive genetic variation between individual placozoan lineages, however, suggests the existence of different species [77]. 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. [14, 72] for overview and discussion). Most recently, a basal position among all diploblastic animals has been suggested [64].

Trichoplax lacks tissues, organs and any type of symmetry. It is composed of only a few hundred to a few thousand cells. This organism has a simple upper and lower epithelium, which surround a network of fiber cells, and as such has an irregular, three-layered, sandwich-type organisation. 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 [34]. It is therefore among the simplest multicellular organism. With 106Mb, the nuclear genome of *Trichoplax adhaerens*, which has recently been completely sequenced [66], 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** [37], most other ncRNAs are relatively 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 *Trichoplax adhaerens* is a prime example for this situation.

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.

2 Homology-Based ncRNA Annotation

2.1 tRNAs

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

Table 1

Summary of tRNA genes arranged by anti-codon.

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

2nd	3rd	A	C	G	T
A	A	Val Leu+(ψ)	Leu [†]	ψ	Leu
	C		Val		Val
	G		Leu		Leu
	T		Met ²	Ile	Ile [†]
C	A	Arg	Trp	Cys ²	SeC
	C		Gly	Gly	Gly ²
	G		Arg ²		Arg [†]
	T		Arg	Ser	Arg [†]
G	A	Ser +(3 ψ)	Ser		Ser
	C	Ala	Ala		Ala
	G	Pro	Pro		Pro
	T	Thr	Thr		Thr
T	A			Thy [†]	
	C		Glu	Asp	Glu ²
	G		Gln	His	Gln
	T		Lys	Asn	Lys

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.

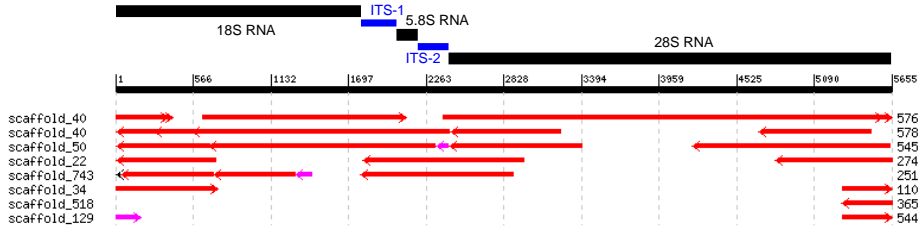


Fig. 1. *Trichoplax* pre-rRNA cluster reconstructed from previously published sequences **L10828**, **Z22783**, **AY652578** (SSU), **AY303975**, **AY652583** (LSU), **U65478** (internal spacers and 5.8S) and **Triad1** genomic sequence. Blast hits of the pre-rRNA to the **Triad1** genome assembly are shown below as in the JGI genome browser.

2.2 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 [50]. *Trichoplax* is no exception, see Fig. 1. The rRNA sequences have already received considerable attention in a phylogenetic context, see [78, 53, 77, 12]. 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 remain uncertain. Fig. 1 summarizes the **blastn** matches of the pre-rRNA to the *Trichoplax* genome.


The 5S rRNA sequence of *Trichoplax* has long been known [76]. 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).

2.3 Spliceosomal snRNAs

Splicing on mRNAs is a common feature to almost all eukaryotic organisms. The spliceosome consists of more than a hundred protein components and five small RNAs that perform crucial catalytic functions, see [51, 75] 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 spliceo-

Table 2

Proximal sequence element (PSE) and location of snRNAs in *Trichoplax adhaerens*. The sequence-logo was generated using **aln2pattern** [45].

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

some, which contains U11, U12, U4atac, U5 and U6atac snRNAs[74]. 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. 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.

Their secondary structures, Fig. 2, closely conform to the metazoan consensus [39], with slightly shorter stems II of U11 snRNA and IV of U12 snRNA. The U12 contains an 5nt insert indicated in red in Fig. 2.

In contrast to many other invertebrates, *Trichoplax* snRNAs feature a clearly recognizable proximal sequence element (PSE) see [27, 39], which is easily detected by **MEME** [6, 7], see Tab. 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 nucleotides from the consensus.

2.4 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. [56, 84, 83].

RNase P RNA is typically easy to find in genomic DNA, at least within meta-

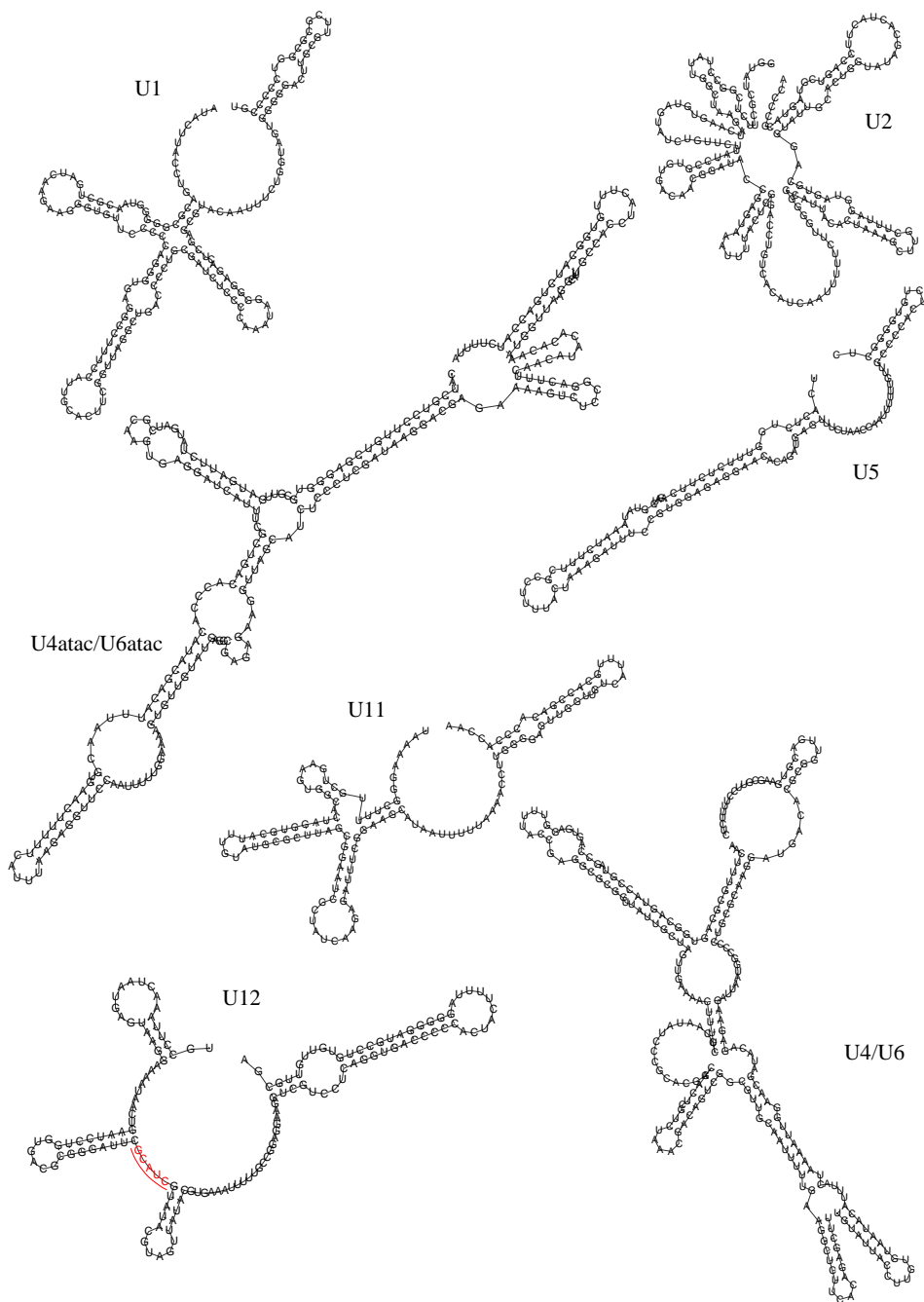


Fig. 2. RNA secondary structures of major spliceosomal (U1, U2, U4, U5, U6) and minor spliceosomal (U11, U12, U4atac, U5, U6atac) snRNAs. For U4/U6 and U4atac/U6atac the interaction structures computed by means of **RNAcofold** are shown. The 5nt insert (relative to other metazoa) is highlighted in the U12.

zoa. The RNase MRP RNA, which is also expected to be present throughout metazoa, is typically much less conserved. Despite substantial efforts [56], RNase MRP RNA homologs have escaped discovery in many bilaterian clades. Not surprisingly, therefore, the *Trichoplax* RNase P RNA was easily identified

[illegible][illegible][illegible]

by **blastn** using the **Rfam** sequences as query. The RNase P sequence is easily verified using **infern**al and the corresponding **Rfam** model.

The signal recognition particle (SRP) binds to the signal peptide emerging

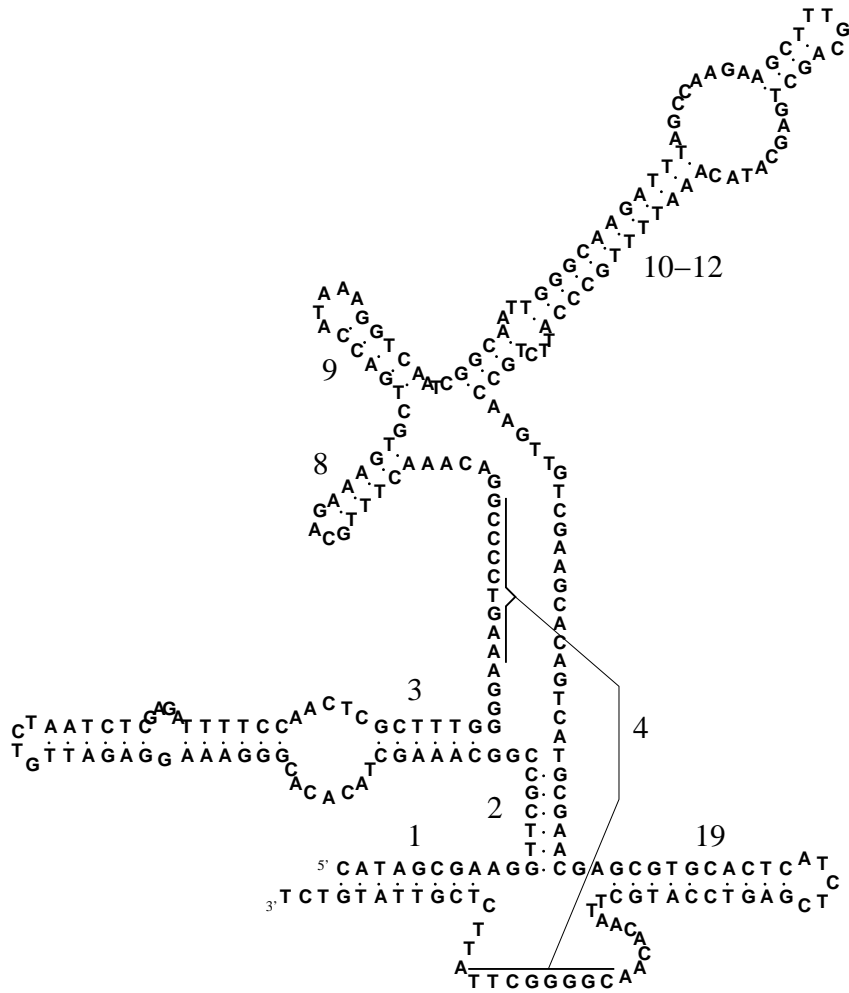


Fig. 4. Secondary structure of *Trichoplax adhaerens* RNase MRP RNA inferred from the multiple alignment of metazoan RNase MRP RNAs provided in the Electronic Supplement.

from the exit site of the ribosome and targets the signal peptide-bearing proteins to the prokaryotic plasma membrane or the eukaryotic endoplasmic reticulum membrane [48]. 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 [1]. The *Trichoplax* SRP RNA is shown in Figure 3.

2.5 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,

Table 3

Small nucleolar RNAs in *Trichoplax*.

Target sites homologous to the ones in human rRNAs are indicated by an asterisk.

Name	Class	target	conservation	Note
U3	C/D	18S 5-22* 18S 1129-1140*	eukaryotes	verified
U18	C/D	28S A740 *	eukaryotes	
U36	C/D	18S A615 *	eukaryotes	
U76	C/D	28S A1549 *	vertebrates	
U106	C/D	28S A2227?	vertebrates	
U17	H/ACA	†	eukaryotes	uncertain
U71 ?	H/ACA	?	vertebrates	
sc.3857:103-213(-)	H/ACA	28S U1370 U1884	novel	

†The U17 snoRNA probably targets the 5'externally transcribed spaces (5'ETS), the exact target is still unknown, however [18, 3].

reviewed e.g. in [5].

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 an uracil residue, it is required in the early steps of rRNA maturation, in particular for the cleavage of the 5' external transcribed spacer (5'ETS) and 18S rRNA maturation, see e.g. [19, 38, 13]. 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, fig. 3. 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 [3], no *Trichoplax* homolog was found using **blast**. Not surprisingly, no other snoRNA homologs were detected by means of **blast**.

The U17 gene was readily identified by **GotohScan**, however. We therefore conducted a survey of the *Trichoplax* genome for homologs 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**¹ [36].

The initial search, for which we used very non-stringent score cut-offs, pro-

¹ <http://www-snorna.biotoul.fr/>

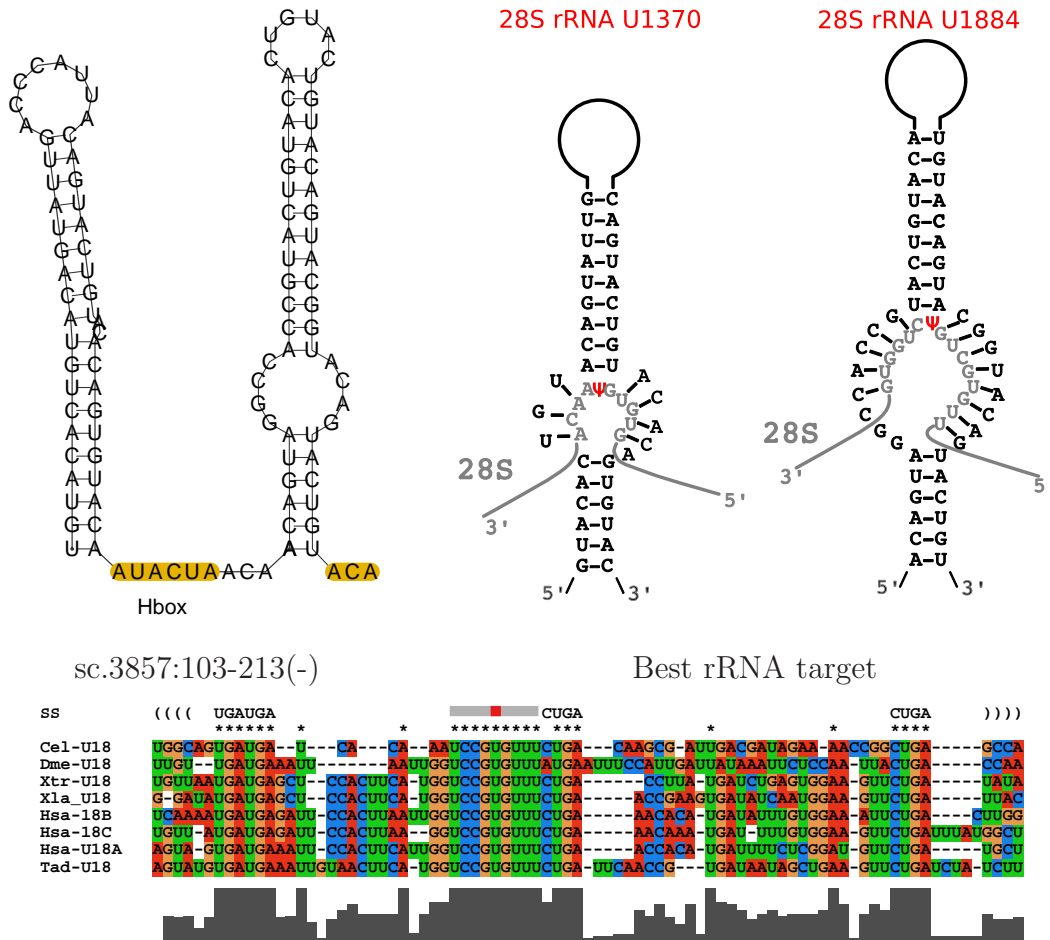


Fig. 5. **Top:** Secondary structure model of a novel H/ACA snoRNA (l.h.s.) and the best *snoPLEX* prediction of its targets sites in the rRNA operon (r.h.s.). **Be-**
low: Alignment of U18 snoRNA sequences from several Metazoa. Boxes and the
conserved target binding site are indicated.

duced 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* [28] 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 Fig. 5, are provided in the Electronic Supplement.

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 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 [11], see also [82]. 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*.

2.6 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 hair-

pin 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.

3 *Ab initio* ncRNA Prediction

3.1 *RNAz* Screen

An alternative to direct homology-based annotation is the *ab initio* prediction of ncRNAs. In particular *RNAz* [80] has been proved to yield results in wide variety of species, from screens of the human genome compared against (mostly) mammalia [79, 81], teleost fishes [61], urochordates [42], nematodes [43], flies [62], yeasts [69], and plasmodium [47]. 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 [80].

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 [10] of *Trichoplax adhaerens*, and the cnidaria *Hydra magnipapillata* and *Nematostella vectensis*. We used all alignment blocks containing *Trichoplax* and at least one of the two cnidarians.

A second screen was performed using *NcDNAalign* alignments [60] constructed from *Trichoplax adhaerens*, *Porites lobata*, and shotgun traces from *Amphimedon queenslandica*, *Acropora millepora*, *Acropora palmata*, and *Hydra magnipapillata*. 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 *NcDNAalign* 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

Table 4
RNAz screens of *Trichoplax adhaerens* genome.

	multiz	NcDNalign	known
Aligned DNA (nt)	4837148	135140	—
alignments	35039	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	*
snRNAs	6	4	10
MRP,P,7SL	1	0	3
protein coding	1022	11	96963
repeat elements	66	1	—
total annotated	1211	101	
unannotated with EST	12	0	
without annotation	205	0	

The asterisk (*) indicates that 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.

correspond to annotated ncRNAs, while 1088 (77%) overlap known protein-coding regions or known repetitive elements. 12 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*.

Fig. 6 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 false discovery rate 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, Tab. 4. The data nevertheless provide at least statis-

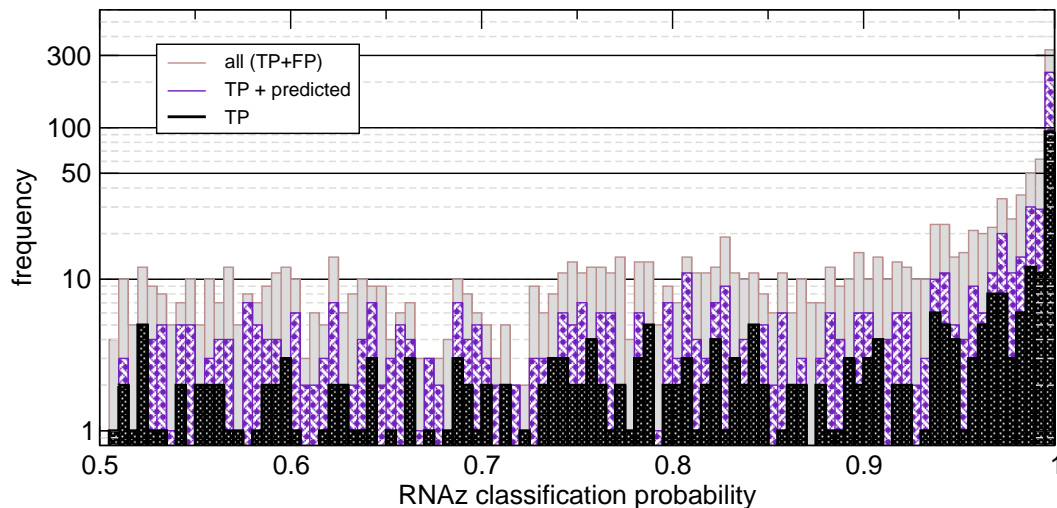


Fig. 6. Distribution of RNAz classification score for known (true positive) (black) all predictions (grey), and only those that are identified as coding or repetitive (maroon). Note the logarithmic scale: there are more than 100 non-annotated predictions with a classification confidence above 99%.

tical evidence for a set of about 100-200 novel structured RNA elements.

The 744 *NcDNa*align were searched with *RNA*micro 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 *RNA*micro evaluates alignment and the corresponding consensus fold, we also checked whether the *Trichoplax* candidate sequences alone fold into a microRNA-like hairpin structure. 64 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 *Nematostella vectensis* or the 8 *Amphimedon queenslandica* microRNAs described in [23]. We thus suggest that these conserved hairpins are not microRNAs. Instead they might belong to a previously undescribed class of hairpin structures.

4 Discussion

We have reported here on a comprehensive computational study of non-protein-coding RNA genes in the genome of the placozoan *Trichoplax 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 *Trichoplax 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 [59], 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. *C. elegans* [70], *Drosophila* [62, 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 [65, 57], we found no trace of these genes in *Trichoplax*. Neither did we find a homolog of one of the 8 sponge microRNAs [23]. Our analysis is thus consistent with the recent report based on short RNA sequencing [23] 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 [29, 65, 57, 52, 35]. 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 *Amphimedon queenslandica* [23] 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 [31], algae [87, 44] and plants [86, 4, 71]. Under the hypothesis of monophyletic diploplasts, which has recently gained substantial support [17, 64], 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*.

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 [41], 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 300nt — it has not been identified so far in any invertebrate species [85]. A similarly fast evolution is observed for the 7SK RNA [25, 24]. Due to their small size and weak sequence constraints, U7 snRNA [40, 15], Y RNAs [46, 55], and vault RNAs [67] 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.

Materials and Methods

4.1 Sequence Data and Databases

The **Triad1** assembly of the genome of *Trichoplax adhaerens* [66] was downloaded from the website of the Joint Genome Institute². For comparison, we used the the **Nemve1**³ assembly of *Nematostella vectensis* [58], as well as the

² <http://genome.jgi-psf.org/Triad1/>

³ <http://genome.jgi-psf.org/Nemve1/>

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** [22] and **NonCode** [26] databases. In addition we used the collection of metazoan snRNAs from [39]. (The snRNAs found in the current study were made available to [39]).

4.2 Software

Homology searches were performed using NCBI **blastall** 2.2.6 [2], **infernall** [49], **fragrep** [45], and the novel **GotohScan** method described below in detail. Alignments were edited in the **emacs** editor using **ralee** mode [21]. RNA secondary structures were computed using the **Vienna RNA Package** [33], in particular the programs **RNAfold** for individual structures, **RNAalifold** [32, 8] for consensus structures of aligned RNA sequences and **RNAcofold** [9] for interaction structures. We used **RNAmicro** [30] in the updated version (1.3)⁴ to identify microRNA candidates from multiple alignments. The analysis of putative snoRNAs was performed using **snoReport** [28], targets for box H/ACA snoRNAs were performed using a preliminary version of **snoplex** [73]. The genome-wide screens for conserved secondary structure elements were performed using **RNAz** [80] as described below.

4.3 RNAz Screens

We used **multiz** [10] 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 **NcDNAalign** [60] 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 120nt are cut into 120 slices in 40nt steps. In a series of filtering steps sequences were removed from the individual alignments or alignment slices if they are (a) shorter than 50nt, or (b) contain more than 25% gap characters or (c) have a base composition outside the definition range of **RNAz**. All preprocessing steps were performed using the script **rnazWindows.pl** of the current release of the **RNAz** package.

⁴ <http://www.bioinf.uni-leipzig.de/~jana/software/RNAmicro.html>

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`.

4.4 GotohScan

Since `blast` failed to identify many of the ncRNAs that are reasonably expected to be present in the *Trichoplax* genome, such as homologs of the U4atac, the U3 snoRNA, and RNase MRP RNA, we decided to use a full dynamic programming approach. Instead of using a local (Smith-Waterman) implementation such as `ssearch` [54] or its partition function version [63], 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 straight-forward modification of Gotoh’s dynamics programming algorithm [20].

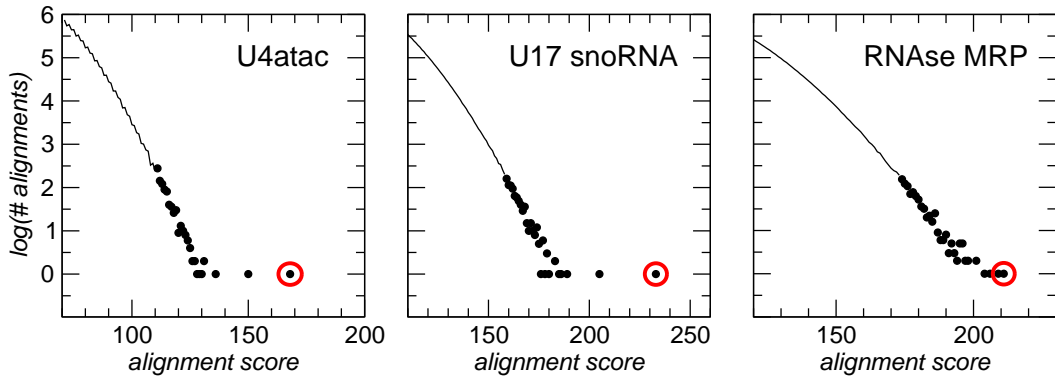
Denote the query sequence by $Q = q_1, q_2, \dots, q_m$ and the genomic “subject” sequence by $P = p_1, p_2, \dots, p_n$. Note that the problem is not symmetric since deletions of the ends of P do not incur costs, while deletions of the ends of Q are fully penalized. As usual, denote by S_{ij} the optimal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

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

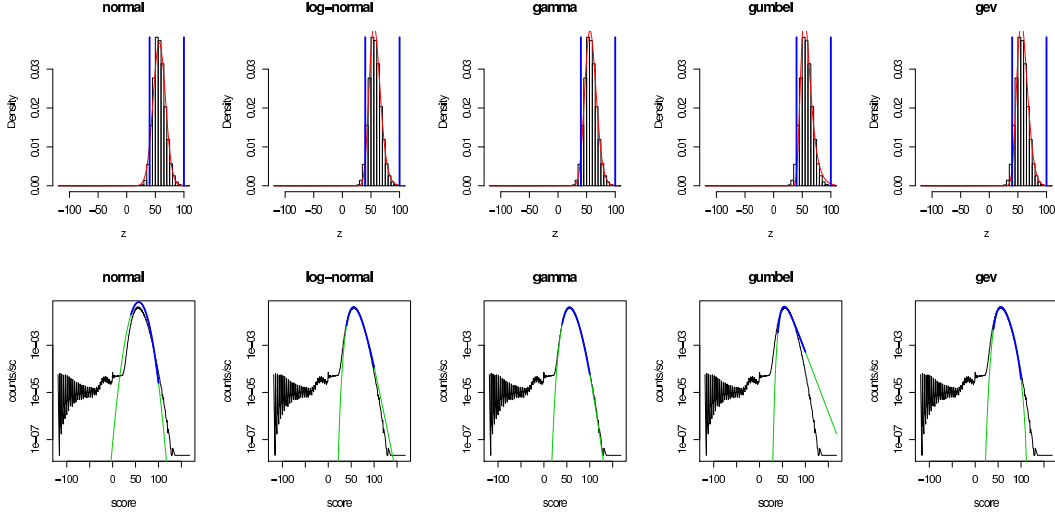
with the initializations

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

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



(a) Histogram of score distribution for U4atac, U17 and RNase MRP



(b) Fitting the GotohScan score distribution of U4atac to known density functions

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.

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. Fig. 7(a) 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

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

see Fig. 7(b). 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 $\langle \ln s \rangle$:

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

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 [16] of the incomplete Gamma function:

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

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

4.5 Target prediction

The targets of the novel box H/ACA snoRNA candidate are computed using the novel run-time efficient **snoplex** program [73]. 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.

4.6 Experimental Verification of Expression

Our experimental approach is based on [59]. 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 nucleotides 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(T)₃₀VN). Amplification of small RNAs was accomplished with the use of a universal reverse primer (5’-AAGCAGTGGTATCAACGCAGAGT) 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.

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/>.

Acknowledgments

This work was supported in part by the *Deutsche Forschungsgemeinschaft* through the “Graduierten-Kolleg Wissensrepräsentation” at the University of Leipzig, the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung* (project P19411 “Genomdynamik”), the 6th Framework Programme of the European Union (projects SYNLET and EMBIO), grants from Alexander von

References

- [1] M. Alm Rosenblad, G. J., B. Knudsen, C. Zwieb, and T. Samuelsson. SR-PDB (signal recognition particle database). *Nucleic Acids Res.*, 31:D363–364, 2003.
- [2] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, 25:3389–3402, 1997.
- [3] V. Atzorn, P. Fragapane, and T. Kiss. U17/snR30 is a ubiquitous snoRNA with two conserved sequence motifs essential for 18S rRNA production. *Mol Cell Biol.*, 24:17691778, 2004.
- [4] M. J. Axtell, J. A. Snyder, and D. P. Bartel. Common functions for diverse small RNAs of land plants. *Plant Cell*, 19:1750–1769, 2007.
- [5] J.-P. Bachellerie, J. Cavaillé, and A. Hüttenhofer. The expanding snoRNA world. *Biochimie*, 84:775–790, 2002.
- [6] T. L. Bailey and C. Elkan. 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*, pages 28–36, Menlo Park, CA, 1994. AAAI Press.
- [7] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li. MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res.*, 34:W369–W373, 2006.
- [8] S. H. Bernhart, I. L. Hofacker, S. Will, A. R. Gruber, and P. F. Stadler. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 2008.
- [9] S. H. Bernhart, H. Tafer, U. Mückstein, C. Flamm, P. F. Stadler, and I. L. Hofacker. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol. Biol.*, 1:3 [epub], 2006.
- [10] M. Blanchette, W. J. Kent, C. Riemer, L. Elnitski, A. F. A. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, D. Haussler, and W. Miller. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, 14:708–715, 2004.
- [11] A. F. Bompfünewerer, C. Flamm, C. Fried, G. Fritzsche, I. L. Hofacker, J. Lehmann, K. Missal, A. Mosig, B. Müller, S. J. Prohaska, B. M. R. Stadler, P. F. Stadler, A. Tanzer, S. Washietl, and C. Witwer. Evolutionary patterns of non-coding rnas. *Th. Biosci.*, 123:301–369, 2005.
- [12] F. Britto da Silva, V. Muschner, and S. L. Bonatto. Phylogenetic position of placozoa based on large subunit (LSU) and small subunit (SSU) rRNA genes. *Genetics Mol. Biol.*, 30:127–132, 2007.

- [13] A. Cléry, V. Senty-Ségault, F. Leclerc, H. A. Raué, and C. Branlant. Analysis of sequence and structural features that identify the B/C motif of U3 small nucleolar RNA as the recognition site for the Snul3p-Rrp9p protein pair. *Mol. Cellular Biol.*, 27:1191–1206, 2007.
- [14] A. G. Collins, P. Cartwright, C. S. McFadden, and B. Schierwater. Phylogenetic context and basal metazoan model systems. *Integr. Compar. Biol.*, 45:585–594, 2005.
- [15] M. Dávila López and T. Samuelsson. Early evolution of histone mRNA 3' end processing. *RNA*, 14:1–10, 2008.
- [16] P. J. Davis. Gamma function and related function. In M. Abramowitz and I. A. Stegun, editors, *Handbook of Mathematical Functions*, pages 253–266. National Bureau of Standards, Washington, DC, 1964.
- [17] C. W. Dunn, A. Hejno, D. Q. Matus, K. Pang, W. Browne, S. A. Smith, E. Seaver, G. W. Rouse, M. Obst, G. D. Edgecombe, M. V. Sørensen, S. H. D. Haddock, A. Schmidt-Rhaesa, A. Okusu, R. M. Kristensen, W. Wheeler, M. Q. Martindale, and G. Giribet. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*, 452:745–749, 2008.
- [18] C. A. Enright, E. S. Maxwell, G. L. Eliceiri, and B. Sollner-Webb. 5'ETS rRNA processing facilitated by four small RNAs: U14, E3, U17, and U3. *RNA*, 2:1094–1099, 1996.
- [19] S. A. Gerbi, A. V. Borovjagin, M. Ezrokhi, and T. S. Lange. Ribosome biogenesis: role of small nucleolar RNA in maturation of eukaryotic rRNA. *Cold Spring Harbor Symp. Quant. Biol.*, LXVI:575–590, 2001.
- [20] O. Gotoh. An improved algorithm for matching biological sequences. *J. Mol. Biol.*, 162:705–708, 1982.
- [21] S. Griffiths-Jones. RALEE—RNA alignment editor in Emacs. *Bioinformatics*, 21:257–259, 2005.
- [22] 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:D121–D124, 2005.
- [23] A. Grimson, M. Srivastava, B. Fahey, B. J. Woodcroft, H. R. Chiang, N. King, A. M. Degnan, D. S. Rokhsar, and D. P. Bartel. Early origins and evolution of miRNAs and Piwi-interacting RNAs in animals. *Nature*, 455:1193–1197, 2008.
- [24] A. Gruber, C. Kilgus, A. Mosig, I. L. Hofacker, W. Hennig, and P. F. Stadler. Arthropod 7SK rna. *Mol. Biol. Evol.*, 1923–1930:25, 2008.
- [25] A. R. Gruber, D. Koper-Emde, M. Marz, H. Tafer, S. Bernhart, G. Obernosterer, A. Mosig, I. L. Hofacker, P. F. Stadler, and B.-J. Benecke. Invertebrate 7SK snRNAs. *J. Mol. Evol.*, 107–115:66, 2008.
- [26] S. He, C. Liu, G. Skogerbo, H. Zhao, J. Wang, T. Liu, B. Bai, Y. Zhao, and R. Chen. NONCODE v2.0: decoding the non-coding. *Nucleic Acids Res.*, 36:D170–D172, 2008.
- [27] N. Hernandez. Small nuclear RNA genes: a model system to study fundamental mechanisms of transcription. *J. Biol. Chem.*, 276:26733–26736,

- 2001.
- [28] J. Hertel, I. L. Hofacker, and P. F. Stadler. **snoReport**: Computational identification of snoRNAs with unknown targets. *Bioinformatics*, 24:158–164, 2008.
 - [29] J. Hertel, M. Lindemeyer, K. Missal, C. Fried, A. Tanzer, C. Flamm, I. L. Hofacker, P. F. Stadler, and The Students of Bioinformatics Computer Labs 2004 and 2005. The expansion of the metazoan microRNA repertoire. *BMC Genomics*, 7:15 [epub], 2006.
 - [30] J. Hertel and P. F. Stadler. Hairpins in a haystack: Recognizing microRNA precursors in comparative genomics data. *Bioinformatics*, 22:e197–e202, 2006.
 - [31] A. Hinas, J. Reimegård, E. G. Wagner, W. Nellen, V. Ambros, and F. Söderbom. The small RNA repertoire of *Dictyostelium discoideum* and its regulation by components of the RNAi pathway. *Nucleic Acids Res.*, 6714-6726:35, 2007.
 - [32] I. L. Hofacker, M. Fekete, and P. F. Stadler. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, 319:1059–1066, 2002.
 - [33] 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.
 - [34] W. Jakob, S. Sagasser, S. Dellaporta, P. Holland, K. Kuhn, and B. Schierwater. The Trox-2 Hox/ParaHox gene of *Trichoplax* (placozoa) marks an epithelial boundary. *Dev Genes Evol.*, 214:170–175, 2004.
 - [35] C. T. Lee, T. Risom, and W. M. Strauss. 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, 2007.
 - [36] L. Lestrade and M. J. Weber. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucl. Acids Res.*, 34:D158–D162, 2006.
 - [37] T. M. 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.
 - [38] N. Marmier-Gourrier, A. Cléry, V. Senty-Ségault, B. Charpentier, F. Schlotter, F. Leclerc, R. Fournier, and C. Branlant. A structural, phylogenetic, and functional study of 15.5-kD/Snu13 protein binding on U3 small nucleolar RNA. *RNA*, 9:821–838, 2003.
 - [39] M. Marz, T. Kirsten, and P. F. Stadler. Evolution of spliceosomal snrna genes in metazoan animals. *J. Mol. Evol.*, 2008. in press.
 - [40] M. Marz, A. Mosig, B. M. R. Stadler, and P. F. Stadler. U7 snRNAs: A computational survey. *Geno. Prot. Bioinf.*, 5:187–195, 2007.
 - [41] W. F. Marzluff. Metazoan replication-dependent histone mRNAs: a distinct set of RNA polymerase II transcripts. *Curr. Opin. Cell. Biol.*, 17:274–280, 2005.
 - [42] K. Missal, D. Rose, and P. F. Stadler. Non-coding RNAs in *Ciona intesti-*

- nal. Bioinformatics*, 21 S2:i77–i78, 2005. Proceedings ECCB/JBI’05, Madrid.
- [43] K. Missal, X. Zhu, D. Rose, W. Deng, G. Skogerbø, R. Chen, and P. F. Stadler. Prediction of structured non-coding RNAs in the genome of the nematode *Caenorhabditis elegans*. *J. Exp. Zool.: Mol. Dev. Evol.*, 306B:379–392, 2006.
 - [44] A. Molnár, F. Schwach, D. Studholme, E. C. Thuenemann, and D. C. Baulcombe. miRNAs control gene expression in the single-cell alga *Chlamydomonas reinhardtii*. *Nature*, 447:1126–1129, 2007.
 - [45] A. Mosig, J. L. Chen, and P. F. Stadler. Homology search with fragmented nucleic acid sequence patterns. In R. Giancarlo and S. Hannenhalli, editors, *Algorithms in Bioinformatics (WABI 2007)*, volume 4645 of *Lecture Notes in Computer Science*, pages 335–345, Berlin, Heidelberg, 2007. Springer Verlag.
 - [46] A. Mosig, M. Guofeng, B. M. R. Stadler, and P. F. Stadler. Evolution of the vertebrate Y RNA cluster. *Th. Biosci.*, 126:9–14, 2007.
 - [47] T. Mourier, C. Carret, S. Kyes, Z. Christodoulou, P. P. Gardner, D. C. Jeffares, R. Pinches, B. Barrell, M. Berriman, S. Griffiths-Jones, A. Ivens, C. Newbold, and A. Pain. Genome-wide discovery and verification of novel structured RNAs in *Plasmodium falciparum*. *Genome Res.*, 18:281–292, 2008.
 - [48] K. Nagai, C. Oubridge, A. Kuglstatter, E. Menichelli, C. Isel, and L. Jovine. Structure, function and evolution of the signal recognition particle. *EMBO J.*, 22:3479–3485, 2003.
 - [49] E. P. Nawrocki and S. R. Eddy. Query-dependent banding for faster RNA similarity searches. *PLoS Comp. Biol.*, 3:e56, 2007. doi:10.1371/journal.pcbi.0030056.
 - [50] R. N. Nazar. Ribosomal RNA processing and ribosome biogenesis in eukaryotes. *IUBMB Life*, 56:457–465, 2004.
 - [51] T. W. Nilsen. The spliceosome: the most complex macromolecular machine in the cell? *Bioessays*, 25:1147–1149, 2003.
 - [52] R. Niwa and F. J. Slack. The evolution of animal microRNA function. *Curr. Op. Gen. Devel.*, 17:145–150, 2007.
 - [53] D. M. Odorico and D. J. Miller. 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, 1997.
 - [54] W. R. Pearson. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics*, 11:635–650, 1991.
 - [55] J. Perreault, J.-P. Perreault, and G. Boire. The Ro associated Y RNAs in metazoans: evolution and diversification. *Mol. Biol. Evol.*, 24:1678–1689, 2007.
 - [56] P. Piccinelli, M. A. Rosenblad, and T. Samuelsson. Identification and analysis fo ribonuclease P and MRP RNA in a broad range of eukaryotes.

- Nucleic Acids Res.*, 33:4485–4495, 2005.
- [57] S. E. Prochnik, D. S. Rokhsar, and A. A. Aboobaker. Evidence for a microRNA expansion in the bilaterian ancestor. *Dev Genes Evol.*, 217:73–77, 2007.
 - [58] N. H. Putnam, M. Srivastava, U. Hellsten, B. Dirks, J. Chapman, A. Salamov, A. Terry, H. Shapiro, E. Lindquist, V. Kapitonov, J. Jurka, G. Genikhovich, I. Grigoriev, S. M. Lucas, R. E. Steele, J. R. Finnerty, U. Technau, M. Q. Martindale, and D. Rokhsar. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*, 317:86–94, 2007.
 - [59] S. Ro, C. Park, J. Jin, K. M. Sanders, and W. Yan. A PCR-based method for detection and quantification of small RNAs. *Biochem Biophys Res Comm*, 351:756–763, 2006.
 - [60] D. Rose, J. Hertel, K. Reiche, P. F. Stadler, and J. Hackermüller. **McDNAAlign**: Plausible multiple alignments of non-protein-coding genomic sequences. *Genomics*, 92:65–74, 2008.
 - [61] D. Rose, J. Jöris, J. Hackermüller, K. Reiche, Q. Li, and P. F. Stadler. Duplicated RNA genes in teleost fish genomes. *J. Bioinf. Comp. Biol.*, 2008. in press.
 - [62] D. R. Rose, J. Hackermüller, S. Washietl, S. Findeiß, K. Reiche, J. Hertel, P. F. Stadler, and S. J. Prohaska. Computational RNomics of drosophilids. *BMC Genomics*, 8:406, 2007.
 - [63] U. Roshan, S. Chikkagoudar, and D. R. Livesay. Searching for evolutionary distant RNA homologs within genomic sequences using partition function posterior probabilities. *BMC Bioinformatics*, 9:61, 2008.
 - [64] B. Schierwater, M. Eitel, W. Jakob, H.-J. Osigus, H. Hadrys, S. Dellaporta, S.-O. Kolokotronis, and R. DeSalle. Concatenated molecular and morphological analysis sheds light on early metazoan evolution and fuels a modern “urmetazoon” hypothesis. *PLoS Biol.*, 2008.
 - [65] L. F. Sempere, C. N. Cole, M. A. McPeck, and K. J. Peterson. The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *J Exp Zoolog B Mol Dev Evol.*, 306B:575–588, 2006.
 - [66] M. S. Srivastava, E. Begovic, J. Chapman, N. H. Putnam, U. Hellsten, T. Kawashima, A. Kuo, T. Mitros, M. L. Carpenter, A. Y. Signorovitch, M. A. Moreno, K. Kamm, H. Shapiro, I. V. Grigoriev, L. W. Buss, B. Schierwater, S. L. Dellaporta, and D. S. Rokhsar. The Trichoplax genome and the nature of placozoans. *Nature*, 454:955–960, 2008.
 - [67] P. F. Stadler, J. J.-L. Chen, J. Hackermüller, S. Hoffmann, F. Horn, P. Khaitovich, A. K. Kretzschmar, A. Mosig, S. J. Prohaska, X. Qi, K. Schutt, and K. Ullmann. Evolution of vault rnas. 2008. submitted.
 - [68] A. Stark, P. Kheradpour, L. Parts, J. Brennecke, E. Hodges, G. J. Hannon, and M. Kellis. Systematic discovery and characterization of fly microRNAs using 12 drosophila genomes. *Genome Res*, 17:1865–1879, 2007.
 - [69] S. Steigle, W. Huber, C. Fried, P. F. Stadler, and K. Nieselt. Compar-

- ative analysis of structured RNAs in *S. cerevisiae* indicates a multitude of different functions. *BMC Biology*, 5v:25, 2007.
- [70] S. L. Stricklin, S. Griffiths-Jones, and S. R. Eddy. *C. elegans* noncoding RNA genes. *WormBook*, doi/10.1895/wormbook.1.7.1, 2005. http://www.wormbook.org/chapters/www_noncodingRNA/noncodingRNA.html.
 - [71] R. Sunkar and G. Jagadeeswaran. *In silico* identification of conserved microRNAs in large number of diverse plant species. *BMC Plant Biol.*, 8:37, 2008.
 - [72] T. Syed and B. Schierwater. *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, 2002.
 - [73] H. Tafer, J. Hertel, I. L. Hofacker, and P. F. Stadler. **snoplex**: Efficient search for h/aca snorna targets. 2008. in preparation.
 - [74] 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.
 - [75] S. Valadkhan. The spliceosome: caught in a web of shifting interactions. *Curr. Op. Struct. Biol.*, 17:310–315, 2007.
 - [76] K. M. Val'ekho-Roman, V. K. Bobrova, A. V. Troitskiĭ, A. B. Tsetlin, and I. L. Okshteĭn. [new data on *Trichoplax*: the nucleotide sequence of 5S rRNA]. *Dokl Akad Nauk SSSR*, 311:500–503, 1990.
 - [77] O. Voigt, A. G. Collins, V. B. Pearse, J. S. Pearse, A. Ender, H. Hadrys, and B. Schierwater. Placozoa – no longer a phylum of one. *Curr. Biol.*, 14:R944–R945, 2004.
 - [78] P. O. Wainright, G. Hinkle, M. L. Sogin, and S. K. Stickel. The monophyletic origins of the metazoa; an unexpected evolutionary link with fungi. *Science*, 260:340–342, 1993.
 - [79] S. Washietl, I. L. Hofacker, M. Lukasser, A. Hüttenhofer, and P. F. Stadler. Mapping of conserved RNA secondary structures predicts thousands of functional non-coding RNAs in the human genome. *Nature Biotech.*, 23:1383–1390, 2005.
 - [80] S. Washietl, I. L. Hofacker, and P. F. Stadler. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA*, 102:2454–2459, 2005.
 - [81] S. Washietl, J. S. Pedersen, J. O. Korbel, A. Gruber, J. Hackermüller, J. Hertel, M. Lindemeyer, K. Reiche, C. Stocsits, A. Tanzer, C. Ucla, C. Wyss, S. E. Antonarakis, F. Denoeud, J. Lagarde, J. Drenkow, P. Kapranov, T. R. Gingeras, R. Guigó, M. Snyder, M. B. Gerstein, A. Reymond, I. L. Hofacker, and P. F. Stadler. Structured RNAs in the ENCODE selected regions of the human genome. *Gen. Res.*, 17:852–864, 2007.
 - [82] M. J. Weber. Mammalian small nucleolar RNAs are mobile genetic elements. *PLoS Genet.*, 2:e205, 2006.
 - [83] D. K. Willkomm and R. K. Hartmann. An important piece of the RNase P jigsaw solved. *Trends Biochem Sci.*, 32:247–250, 2007.

- [84] M. D. Woodhams, P. F. Stadler, D. Penny, and L. J. Collins. RNase MRP and the RNA processing cascade in the eukaryotic ancestor. *BMC Evol. Biol.*, 7:S13, 2007.
- [85] M. Xie, A. Mosig, X. Qi, Y. Li, P. F. Stadler, and J. J.-L. Chen. Size variation and structural conservation of vertebrate telomerase RNA. *J. Biol. Chem.*, 283:2049–2059, 2008.
- [86] B. Zhang, X. Pan, C. H. Cannon, G. P. Cobb, and T. A. Anderson. Conservation and divergence of plant microRNA genes. *Plant J.*, 46:243–259, 2006.
- [87] T. Zhao, G. Li, S. Mi, S. Li, G. J. Hannon, X. J. Wang, and Y. Qi. A complex system of small RNAs in the unicellular green alga *Chlamydomonas reinhardtii*. *Genes Dev.*, 21:1190–1203, 2007.