BMC Genomics

BioMed Central

Research article

Homology-based annotation of non-coding RNAs in the genomes of *Schistosoma mansoni* and *Schistosoma japonicum*

Claudia S Copeland[1,2], Manja Marz[1], Dominic Rose[1], Jana Hertel[1], Paul J Brindley[2], Clara Bermudez Santana[1,8], Stephanie Kehr[1], Camille Stephan-Otto Attolini[3] and Peter F Stadler*[1,4,5,6,7]

Address: [1]Bioinformatics Group, Department of Computer Science and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany, [2]Department of Microbiology, Immunology & Tropical Medicine, George Washington University Medical Center, 2300 I Street, NW, Washington, DC 20037, USA, [3]Memorial Sloan-Kettering Cancer Center, Computational Biology Department, 1275 York Avenue, Box # 460, New York, NY 10065, USA, [4]Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany, [5]Fraunhofer Institute for Cell Therapy and Immunology, Perlickstraße 1, D-04103 Leipzig, Germany, [6]Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM 87501, USA, [7]Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria and [8]Department of Biology, National University of Colombia, Carrera 45 No. 26-85, Bogotá, D.C., Colombia

Email: Claudia S Copeland - cclaudia@bioinf.uni-leipzig.de; Manja Marz - manja@bioinf.uni-leipzig.de; Dominic Rose - dominic@bioinf.uni-leipzig.de; Jana Hertel - jana@bioinf.uni-leipzig.de; Paul J Brindley - mtmpjb@gwumc.edu; Clara Bermudez Santana - clara@bioinf.uni-leipzig.de; Stephanie Kehr - steffi@bioinf.uni-leipzig.de; Camille Stephan-Otto Attolini - camille@bioinf.uni-leipzig.de; Peter F Stadler* - studla@bioinf.uni-leipzig.de

* Corresponding author

Published: 8 October 2009

BMC Genomics 2009, **10**:464 doi:10.1186/1471-2164-10-464

Received: 27 May 2009
Accepted: 8 October 2009

This article is available from: http://www.biomedcentral.com/1471-2164/10/464

Abstract

Background: Schistosomes are trematode parasites of the phylum Platyhelminthes. They are considered the most important of the human helminth parasites in terms of morbidity and mortality. Draft genome sequences are now available for *Schistosoma mansoni* and *Schistosoma japonicum*. Non-coding RNA (ncRNA) plays a crucial role in gene expression regulation, cellular function and defense, homeostasis, and pathogenesis. The genome-wide annotation of ncRNAs is a non-trivial task unless well-annotated genomes of closely related species are already available.

Results: A homology search for structured ncRNA in the genome of *S. mansoni* resulted in 23 types of ncRNAs with conserved primary and secondary structure. Among these, we identified rRNA, snRNA, SL RNA, SRP, tRNAs and RNase P, and also possibly MRP and 7SK RNAs. In addition, we confirmed five miRNAs that have recently been reported in *S. japonicum* and found two additional homologs of known miRNAs. The tRNA complement of *S. mansoni* is comparable to that of the free-living planarian *Schmidtea mediterranea*, although for some amino acids differences of more than a factor of two are observed: Leu, Ser, and His are overrepresented, while Cys, Meth, and Ile are underrepresented in *S. mansoni*. On the other hand, the number of tRNAs in the genome of *S. japonicum* is reduced by more than a factor of four. Both schistosomes have a complete set of minor spliceosomal snRNAs. Several ncRNAs that are expected to exist in the *S. mansoni* genome were not found, among them the telomerase RNA, vault RNAs, and Y RNAs.

Conclusion: The ncRNA sequences and structures presented here represent the most complete dataset of ncRNA from any lophotrochozoan reported so far. This data set provides an important reference for further analysis of the genomes of schistosomes and indeed eukaryotic genomes at large.

Background

Non-coding RNA (ncRNA) plays a crucial role in gene expression regulation, cellular function and defense, and disease. Indeed, in higher eukaryotes, most of the genomic DNA sequence encodes non-protein-coding transcripts [1]. In contrast to protein-coding mRNAs, ncRNAs do not form a homogeneous class. The best-characterized subclasses form stable basepairing patterns (secondary structures) that are crucial for their function. This group includes the well-known tRNAs, catalytically active RNAs such as rRNA, snRNAs, RNase P RNA, and other ribozymes, and regulatory RNAs such as microRNAs and spliceosomal RNAs that direct protein complexes to specific RNA targets. Much less is known about long mRNA-like ncRNAs, which are typically poorly conserved at the level of both sequence and structure.

Most non-vertebrate genome projects have put little emphasis on a comprehensive annotation of ncRNAs. Indeed, most non-coding RNAs, with the notable exception of tRNAs and rRNAs, are difficult or impossible to detect with BLAST in phylogenetically distant organisms. Hence, ncRNA annotation is not part of generic genome annotation pipelines. Dedicated computational searches for particular ncRNAs, for example, RNase P and MRP [2,3], 7SK RNAs [4,5], or telomerase RNA [6,7], are veritable research projects in their own right. Despite best efforts, ncRNAs across the animal phylogeny remain to a large extent uncharted territory.

The main difficulty with ncRNA annotation is poor sequence conservation and indel patterns that often correspond to large additional "expansion domains". In many cases, the secondary structure is much better conserved than the primary sequence, providing a means of confirming candidate ncRNAs even in cases where sequence conservation is confined to a few characteristic motifs. Secondary structure conservation can also be utilized to detect homologs of some ncRNAs based on characteristic combinations of sequence and structure motifs using special software tools designed for this purpose.

In [8] we described a protocol for a more detailed homology-based ncRNA annotation than what can be achieved with currently available automatic pipelines. Here, we apply this scheme to the genome of *S. mansoni*, and by comparison with the newly sequenced *S. japonicum* genome, identify ncRNAs in both of these clinically important schistosomes.

Schistosomes belong to an early-diverging group within the Digenea, but are clearly themselves highly derived [9-11]. The flatworms are a long-branch group, suggesting rapid mutation rates (see [12]).

Schistosome genomes are comparatively large, estimated to be over 350 megabase pairs, and perhaps as high as 400 megabase pairs, for the haploid genome of *S. mansoni* and *S. japonicum* [13-15]. The other major schistosome species parasitizing humans probably have a genome of similar size, based on the similarity in appearance of their karyotypes [16]. These large sizes may be characteristic of platyhelminth genomes in general: the genome of *Schmidtea mediterranea* is even larger, with the current genome sequencing project reporting a size of ~480 million base pairs [17]http://genome.wustl.edu/genomes/view/schmidtea_mediterranea/.

Genome sequencing of the seven autosomes and the pair of sex chromosomes of *S. mansoni* with about 8× coverage has lead to a genome assembly comprising 5,745 scaffolds (> 2 kb) covering 363 Mb [13,14,18]. Similarly, shotgun sequencing of *S. japonicum* with coverage of 5.4× decoded 397 Mb of sequence [15]. These form about 25,000 scaffolds. Albeit both genome projects did not lead to complete finished genomes, we therefore know at least 90-95% of the genomic DNA sequences of *S. japonicum* and *S. mansoni*, respectively.

The protein-coding portion of the *Schistosoma* genomes have received much attention in recent years. Published work includes transcriptome databases for both *S. japonicum* [19] and *S. mansoni* [20], microarray-based expression analysis [21], characterization of promoters [22,23], and physical mapping and annotation of protein-coding genes from both the *S. mansoni* and *S. japonicum* genome projects [18]. Recently, a systematic annotation of protein-coding genes in *S. japonicum* was reported [24]. In contrast to other, better-understood, parasites such as *Plasmodium* [25], however, not much is known about the non-coding RNA complement of schistosomes. Only the spliced leader RNA (SL RNA) of *S. mansoni* [26], the hammer-head ribozymes encoded by the SINE-like retrotransposons Sm-α and Sj-α [27,28], and secondary structure elements in the LTR retrotransposon *Boudicca* [29] have received closer attention. Ribosomal RNA sequences have been available mostly for phylogenetic purposes [30], and tRNAs have been studied to a limited degree [31].

The wealth of available ESTs, in principle, provides a valuable resource for ncRNA detection. Since mostly poly-A ESTs have been generated, it is not surprising that most ESTs have been attributed to protein-coding genes [32]. The large evolutionary distance, with 55% of the genes without homologs outside the genus [13,18], makes it hard or even impossible to reliably distinguish ESTs of putative mRNA-like ncRNAs from non-coding portions of protein-coding transcripts.

In this contribution we therefore focus on a comprehensive overview of the evolutionary conserved non-coding RNAs in the genomes of *S. mansoni* and *S. japonicum*. We discuss representatives of 23 types of ncRNAs that were detected based on both sequence and secondary structure homology.

Results and discussion

Structure and homology-based searches of the schistosome genomes revealed ncRNAs from 23 different RNA categories. Table 1 lists these functional ncRNA categories, the number of predicted genes in each category, and references associated with each RNA type. Supplementary `fasta` files containing the ncRNA genes, `bed` files with the genome annotation, and `stockholm`-format alignment files can be accessed at http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/08-014.

Transfer RNAs

Candidate tRNAs were predicted with `tRNAscan-SE` in the genomes of *S. mansoni*, *S. japonicum* and *S. mediterranea* (a free-living platyhelminth, used for comparison). After removal of transposable element sequences (see below), `tRNAscan-SE` predicted a total of 713 tRNAs for *S. mansoni* and 739 for *S. mediterranea*, while 154 tRNAs were found in the *S. japonicum* sequences. These included tRNAs encoding the standard 20 amino acids of the traditional genetic code, selenocysteine encoding tRNAs (tRNAsec) [33] and possible suppressor tRNAs [34] in all three genomes. The tRNAsec from schistosomes has been characterized, and is similar in both size and structure to tRNAsec from other eukaryotes [35].

The tRNA complements of the three platyhelminth genomes are compared in detail in Figure 1. The amino

Table 1: Summary of homology-based RNA annotations from the sequenced genomes of *S. mansoni* and *S. japonicum*.

RNA class	Functional Category	*S. man.*	*S. jap.*	Related reference(s)
7SK	Transcription regulation	(1)	0	This study
Hammerhead ribozymes	Self-cleaving	> 38, 000	> 5, 000	[27]
miRNA	Translation control	8	7	[109], this study
potassium channel motif	RNA editing	9	3	[65]
RNase MRP	Mitochondrial replication, rRNA processing	(1)	(1)	This study
RNase P	tRNA processing	1	1	This study
rRNA-operon	Polypeptide synthesis	80-105	50-280	[39], this study
5S rRNA	Polypeptide synthesis	21	1-13	This study
SL RNA	*Trans*-splicing	6-48	1-9	[26], this study
SnoRNA U3	Nucleolar rRNA processing	1	1	This study
SRP	Protein transportation	12	4+1	This study
tRNA	Polypeptide synthesis	663	154	This study
U1	Splicing	3-34	2-6	[44], this study
U2	Splicing	3-15	1-63	[44], this study
U4	Splicing	1-19	1-6	[44], this study
U5	Splicing	2-9	1-24	[44], this study
U6	Splicing	9-55	2-12	[44], this study
U11	Splicing	1	1	This study
U12	Splicing	1-2	0-1	[44], this study
U4atac	Splicing	1	1	This study
U6atac	Splicing	1	1	This study
U7	Histone maturation	0	(2)	This study

Where are range of numbers is given, it remains uncertain whether multiple copies in the genomic DNA are true copies of the gene or assembly artifacts.

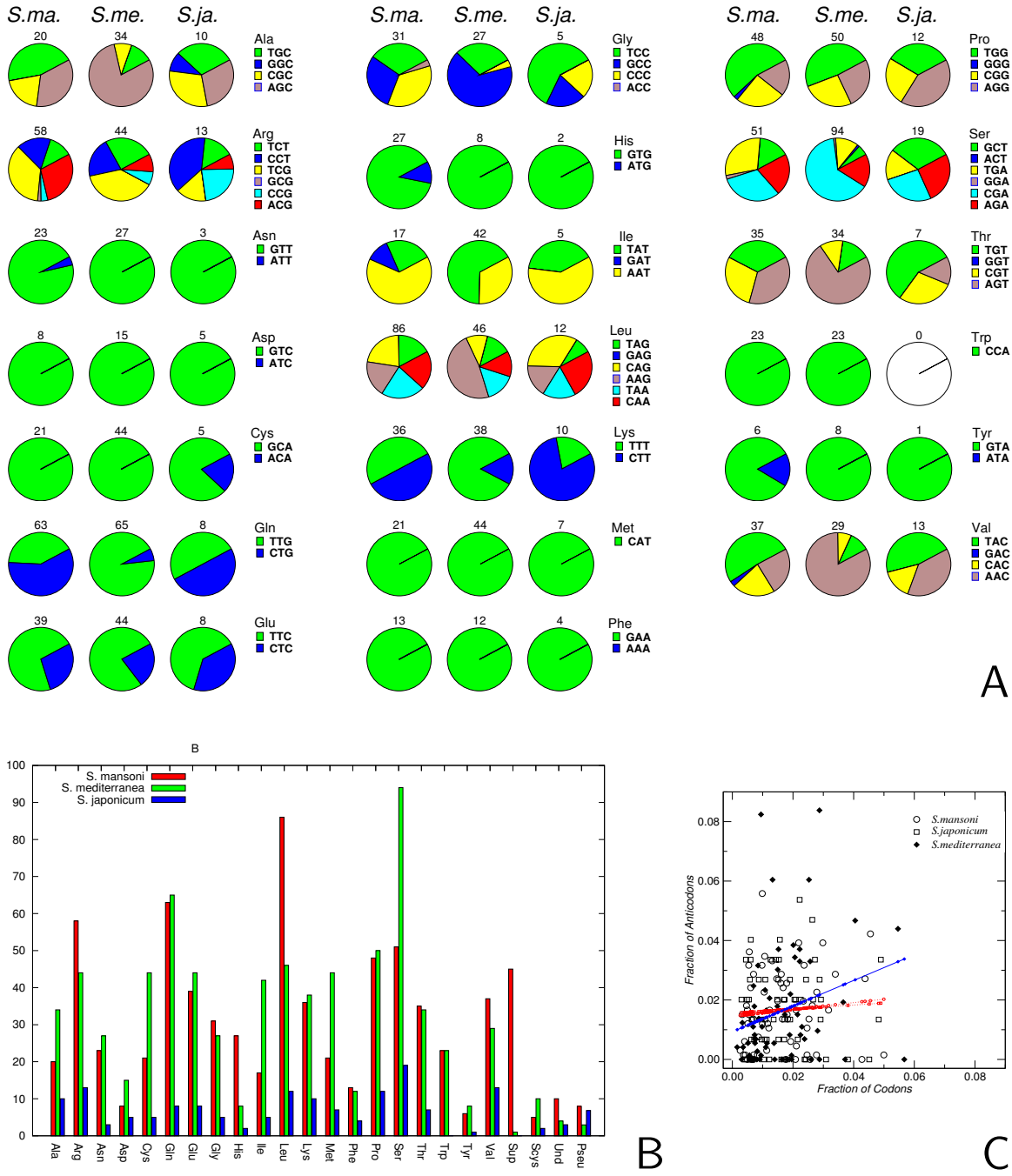

Figure 1

Comparison of the tRNA complement of *Schistosoma mansoni*, *Schistosoma japonicum*, and *Schmidtea mediterranea*. **A**: Comparison of anti-codon distributions for the 20 amino acids. Numbers below each pie-chart are the total number of tRNA genes coding the corresponding amino acid. Left columns: *S. mansoni*; middle columns: *S. mediterranea*; right columns: *S. japonicum*. **B**: Number of tRNAs encoding a particular amino acid. red: *S. mansoni*, blue: *S. japonicum*, green: *S. mediterranea*. Abbreviations: `Sup`: putative suppressor tRNAs (CTA, TTA); `Scys`: Selenocysteine tRNAs (TCA); `Pseu`: predicted pseudogenes; Und: tRNA predictions with uncertain anticodon; likely these are also tRNA pseudogenes. The Gln-tRNA derived repeat family (see text) is not included in these data. **C**: Comparison of codon usage and anti-codon abundance. No significant correlation is observed for the two schistosomes. For *S. mediterranea* there is a weak, but statistically significant, positive correlation: $t \approx 2.0$

acids are represented in approximately equal numbers in *S. mansoni* and *Schmidtea*. Nevertheless, there are several notable deviations. *S. mansoni* contains many more leucine (86 vs. 46) and histidine (27 vs. 8) tRNAs, while serine (51 vs. 94), cysteine (21 vs. 44), methionine (21 vs. 44), and isoleucine (17 vs. 42) are underrepresented. In addition, there are several substantial differences in codon usage. In most cases, *S. mansoni* has a more diverse repertoire of tRNAs: tRNA-Asn-ATT, tRNA-Arg-CGC, tRNA-His-ATG, tRNA-Ile-GAT, tRNA-Pro-GGG, tRNA-Tyr-ATA, tRNA-Val-GAC are missing in *Schmidtea*. Only tRNA-Ser-ACT is present in *Schmidtea* but absent in *Schistosoma*. The tRNA complement of *S. japonicum*, on the other hand, differs strongly from its two relatives. Not only is the number of tRNAs decreased by more than a factor of four, *S. japonicum* also prefers anticodons that are absent or rare in its relatives, such as tRNA-Ala-GGC, tRNA-Cys-ACA, and Lys-CTT. On the other hand, no tRNA-Trp was found. Since the UGG codon is present in many open reading frames we interpret this as a problem with the incompleteness of the genome assembly rather than a genuine gene loss. The reduction in the number of tRNAs is also evident by comparing the number of tRNAs with introns: 27 in *S. mansoni* versus 5 in *S. japonicum*.

It has been shown recently that changes in codon usage, even while coding the same protein sequences, can severely attenuate the virulence of viral pathogens [36] by "de-optimizing" translational efficiency. This observation leads us to speculate that the greater diversity of the tRNA repertoire could be related to the selection pressures of the parasitic life-style of *S. mansoni*. The effect is not straightforward, however, because there is no significant correlation of tRNA copy numbers with the overall codon usage in both *S. mansoni* and *S. japonicum*, Figure 1C. In contrast, a weak but statistically significant correlation can be observed in *Schmidtea mediterranea*. It would be interesting, therefore, to investigate in detail whether there are differences in codon usage of proteins that are highly expressed in different stages of *S. mansoni*'s life cycle, and whether the relative expression levels of tRNAs are under stage-specific regulation.

The most striking result of the `tRNAscan-SE` analysis was the initial finding of 1,135 glutamine tRNAs (Gln-tRNAs) in *S. mansoni* in contrast to the 8 Gln-tRNAs in *S. japonicum* and 65 in *S. mediterranea*. Nearly all of these (1,098 in *S. mansoni*) were tRNA-Gln-TTG. In addition, an extreme number of 1,824 tRNA-pseudogenes in *S. mansoni* (vs. 951 in *S. japonicum* and 19 in *S. mediterranea*) was predicted. Of these, 1,270 were also homologous to tRNA-Gln-TTG. These two groups of tRNA-Gln-TTG-derived genes (those predicted to be pseudogenes and those predicted to be functional tRNAs) totaled 2,368. These high numbers suggest a tRNA-derived mobile genetic element. We therefore ran the 2,368 *S. mansoni* tRNA-Gln-TTG genes through the `RepeatMasker` program [37]. Almost all of them (2,342) were classified as SINE elements. Further `BLAST` analysis revealed that these elements are similar to members of the Sm-α family of *S. mansoni* SINE elements [38]. Removal of these SINE-like elements yielded a total of 63 predicted glutamine-encoding tRNAs in *S. mansoni*. About 650 of 951 pseudogenes in *S. japonicum* derived from tRNA-Pro-CGG.

Homology-based analysis yielded similar, though somewhat less sensitive, results to those of `tRNAscan-SE`. For instance, a `BLAST` search in *S. mansoni* with `Rfam's` tRNA consensus yielded 617 predicted tRNAs compared to the 663 predictions made by `tRNAscan-SE`.

Ribosomal RNAs

As usual in eukaryotes, the 18S, 5.8S, and 28S genes are produced by RNA polymerase I from a tandemly repeated polycistronic transcript, the ribosomal RNA operon. The *S. mansoni* genome contains about 90-100 copies [39,40] which are nearly identical at sequence level, because they are subject to concerted evolution [41]. The repetitive structure of the rRNA operons causes substantial problems for genome assembly software [42]. In order to obtain a conservative estimate of the copy number, we retained only partial operon sequences that contained at least two of the three adjacent rRNA genes. We found 48 loci containing parts of 18S, 5.8S, and 28S genes, 32 loci covering 18S and 5.8S rRNA, and 57 loci covering 5.8S and 28S rRNAs [see Additional file 1 - Figures S1 and S2]. Adding the copy numbers, we have not fewer than 80 copies (based on linked 18S rRNAs) and no more than 137 copies (based on linked 5.8S rRNA). The latter is probably an overestimate due to the possibility that the 5.8S rRNA may be contained in two scaffolds. The copy number of rRNA operons is thus consistent with the estimate of 90-100 from hybridization analysis [39]. An analogous analysis of the current *S. japonicum* assembly yields less accurate results. Due to the many short fragments, we obtained 90 copies; the true number may lie between 50 and 280, however.

The 5S rRNA is a polymerase III transcript that has not been studied in schistosomes so far. We found 21 copies of the 118 nt long 5S rRNA in *S. mansoni*, compared with 13 copies in *S. japonicum*. Four of the 21 copies are located within a 3,000 nt cluster on *Scaffold010519*.

Spliceosomal RNAs and Spliced Leader RNA

Spliceosomes, the molecular machines responsible for most splicing reactions in eukaryotic cells, are ribonucleoprotein complexes similar to ribosomes [43]. The major spliceosome, which cleaves GT-AG introns, includes the five snRNAs U1, U2, U4, U5, and U6. In the

S. mansoni genome, all of them are multicopy genes. By homology search we found 34 U1, 15 U2, 19 U4, 9 U5, and 55 U6 sequences in the genome assembly. Interpreting all sequences that are identical in short flanking regions as the same, we would retain only 3 U1, 3 U2, 1 U4, 2 U5, and nine U6 genes [44]. The true copy number in the *S. mansoni* genome is most likely somewhere between these upper and lower bounds. For *S. japonicum*, the corresponding numbers are U1: 2-6, U2: 1-63 U2, U4: 1-6 U4, U5: 1-24, and U6: 2-12. Due to the more fragmented genome assembly we expect the true numbers to be closer to the lower bounds. Secondary structures for these candidates are similar to those of typical snRNAs, Figure 2.

A second, much less frequent, minor spliceosome is responsible for the processing of atypical AT-AC introns. It shares only the U5 snRNA with the major spliceosome. The other four RNA components are replaced by variants called U11, U12, U4atac, and U6atac [45]. The minor-spliceosomal snRNAs are typically much less conserved than the RNA components of the major spliceosome [44]. It was not surprising, therefore, that these RNAs were detectable only by means of `GotohScan`[8] but not with the much less sensitive `BLAST` searches. Although U4atac and U6atac are quite diverged compared to known homologs, they can be recognized unambiguously based on both secondary structure and conserved sequence motifs. Furthermore, the U4atac and U6atac sequences can interact to form the functionally necessary duplex structure shown in Figure 2. As in many other species, there is only a single copy of each of the minor spliceosomal snRNAs in both of the schistosome genomes, Tab. 1. An analysis of promoter sequences showed that the putative snRNA promoter motifs in *S. mansoni* are highly derived. Only one of the two U12 genes exhibited a clearly visible snRNA-like promoter organization.

The Spliced Leader (SL) RNA is one of the very few previously characterized ncRNAs from *S. mansoni* [26]. The 90 nt SL RNA, which was found in a 595 nt tandemly repeated fragment (accession number M34074), contains the 36 nt leader sequence at its 5' end which is transferred in the *trans*-splicing reaction to the 5' termini of mature mRNAs. Using `blastn`, we identified 54 SL RNA genes. These candidates, along with 100 nt flanking sequence, were aligned using `ClustalX`, revealing 6 sequences with aberrant flanking regions, which we suspect to be pseudogenic. The remaining sequences are 43 identical copies and 5 distinct sequence variants. A secondary structure analysis corroborates the model of [26], according to which the *S. mansoni* SL RNA has only two loops, with an

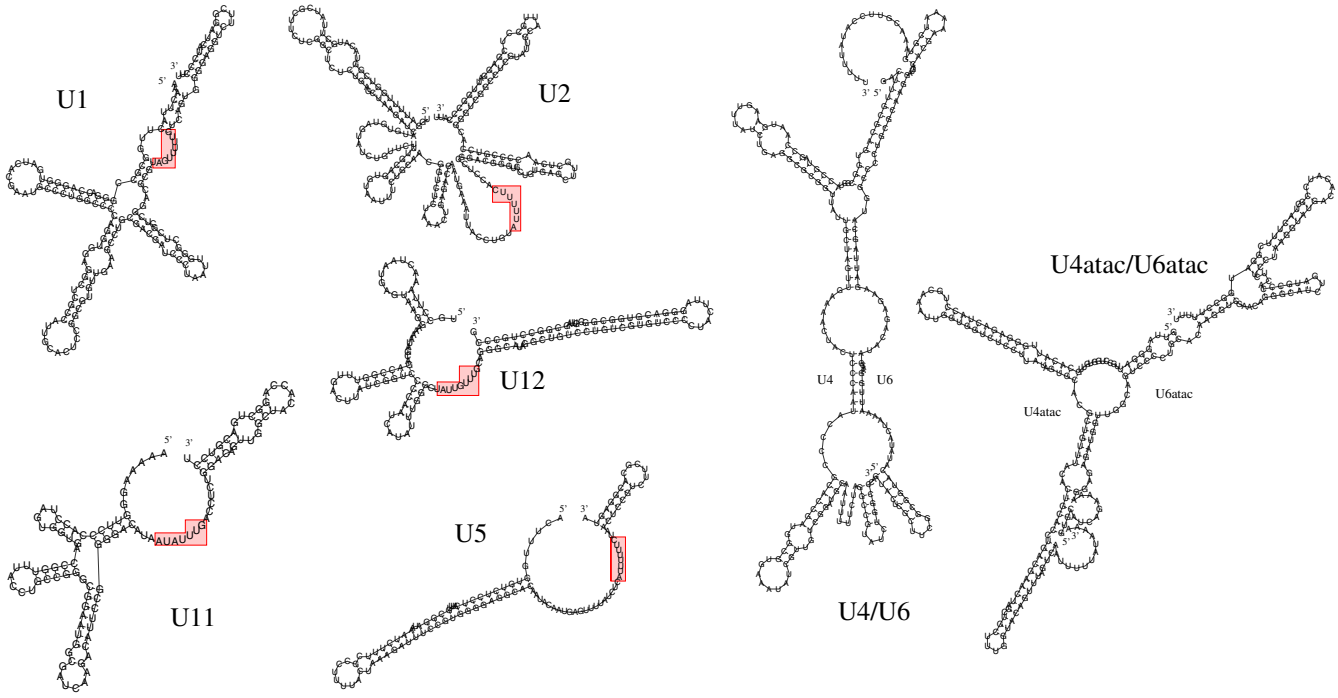

Figure 2
Secondary structures of the nine snRNAs and the interaction complexes of U4/U6 and U4atac/U6atac, respectively, in *S. mansoni*. Structure prediction was performed by `RNAfold, RNAalifold` and for U4/U6 and U4atac/U6atac by `RNAcofold` from the RNA Vienna Package [96,108]. Boxes indicate Sm binding sites. Additional details on sequences, structures, and alignments are available at the supplementary material.

unpaired Sm binding site [see Additional File 1 - Figure S3]. This coincides with the SL RNA structure of Rotifera [46], but is in contrast to the SL RNAs in most other groups of eukaryotes, which exhibit single or triple stem-loop structures [47]. A `blastn`-search against *S. mansoni* EST data confirms that the 5' part of the SL is indeed *trans*-spliced to mRNAs. Several nearly identical SL RNA homologs are found in *S. japonicum*.

SRP RNA and Ribonuclease P RNA

Signal recognition particle (SRP) RNA, also known as 7SL RNA, is part of the signal recognition particle, a ribonucleoprotein that directs packaged proteins to their appropriate locations in the endoplasmic reticulum. Although one of the protein subunits of this ribonucleoprotein was cloned in 1995 [48], little is known about the other subunits or the RNA component in *S. mansoni*. We found eight probable candidates for the SRP RNA, with one almost canonical sequence [see Additional file 1 - Figure S4], and four possible candidates with point mutations which may influence their function.

The RNA component of Ribonuclease P (RNase P) is the catalytically active part of this enzyme that is required for the processing of tRNA precursors [49,50]. We found one classic RNase P RNA in the *S. mansoni* genome using both `GotohScan` and `rnabob` with the eukaryotic ("nuclear") `Rfam` consensus sequence for RNase P as search sequence.

MicroRNAs

MicroRNAs are small RNAs that are processed from hairpin-like precursors, see e.g. [51]. They are involved in post-transcriptional regulation of mRNA molecules. So far, no microRNAs have been verified experimentally in *S. mansoni*. The presence of four protein-coding genes encoding crucial components of the microRNA processing machinery (Dicer, Argonaut, Drosha, and Pasha/DGCR8) [52,53], and the presence of Argonaut-like genes in both *S. japonicum* [54] and *S. mansoni* (detected by `tblastn` in EST data, see Supplemental Data online), strongly suggests that schistosomes have a functional microRNA system. Indeed, most recently five miRNAs were found by direct cloning in *S. japonicum* that are also conserved in *S. mansoni* [55]: *let-7*, *mir-71*, *bantam*, *mir-125*, and a single schistosome-specific microRNA. These sequences, including the precursor hairpins, are well conserved in *S. japonicum*. On the other hand, the microRNA precursor sequences of both schistosomes are quite diverged from the consensus of the homologous genes in Bilateria.

Using bioinformatics (see methods) we were able to find only one further miRNA candidate in *S. mansoni*, *mir-124*, that is also conserved in *S. japonicum*. In insects, this miRNA is clustered with *mir-287*. The distance of both miRNAs is approximately 8 kb in Drosophilids. We found an uncertain *mir-287* candidate in *S. mansoni*, however, on a different scaffold than *mir-124*. Although this sequence nicely folds into a single stem-loop structure, it is conserved only antisense to the annotated mature sequence in insects (see, Figure 3). This *S. mansoni mir-287* candidate does not seem to be conserved in *S. japonicum*.

In [56], 71 microRNAs are described for the distantly related trematode *Schmidtea mediterranea*, and additional ones are announced in a recent study focussing on piRNAs [57]. The overwhelming majority, 54, were reported to be members of 29 widely conserved metazoan microRNA families, although in some cases even the mature miRNA sequence is quite diverged. Therefore, we regard several family assignments as tentative at best. Of those 29 miRNAs, we found *mir-124* only. However, the schistosome sequences are more related to the other bilaterian *mir-124* homologs than to those of *S. mediterranea*. Out of the remaining 54 miRNAs that were annotated in *S. mediterranea* we found that *mir-749* is also conserved in the two schistosome species. Here, the sequences show a common consensus sequence *and* secondary structure in their precursors (see Figure 3).

The small number of recognizable microRNAs in schistosomes is in strong contrast to the extensive microRNA complement in *S. mediterranea*, indicating massive loss of microRNAs relative to the planarian ancestor. This may be a consequence of the parasitic lifestyle of the schistosomes.

Small Nucleolar RNAs

Small nucleolar RNAs play essential roles in the processing and modification of rRNAs in the nucleolus [58,59]. Both major classes, the box H/ACA and the box C/D snoRNAs are relatively poorly conserved at the sequence level and hence are difficult to detect in genomic sequences. This has also been observed in a recent ncRNA annotation project of the *Trichoplax adhaerens* genome [8]. The best-conserved snoRNA is the atypical U3 snoRNA, which is essential for processing of the 18S rRNA transcript into mature 18S rRNA [60]. In the current assembly of the *S. mansoni* genome we found six U3 loci, but they are also identical in the flanking sequences, suggesting that in fact there is only a single U3 gene. No unambiguous homologue was detected for any of the other known snoRNAs.

A *de novo* search for snoRNAs (see methods for details) resulted in 2,610 promising candidates (1,654 box C/D and 956 box H/ACA), see Supplemental Data online. All these predictions exhibit highly conserved sequence boxes as well as the typical secondary features of box C/D and box H/ACA snoRNAs, respectively.

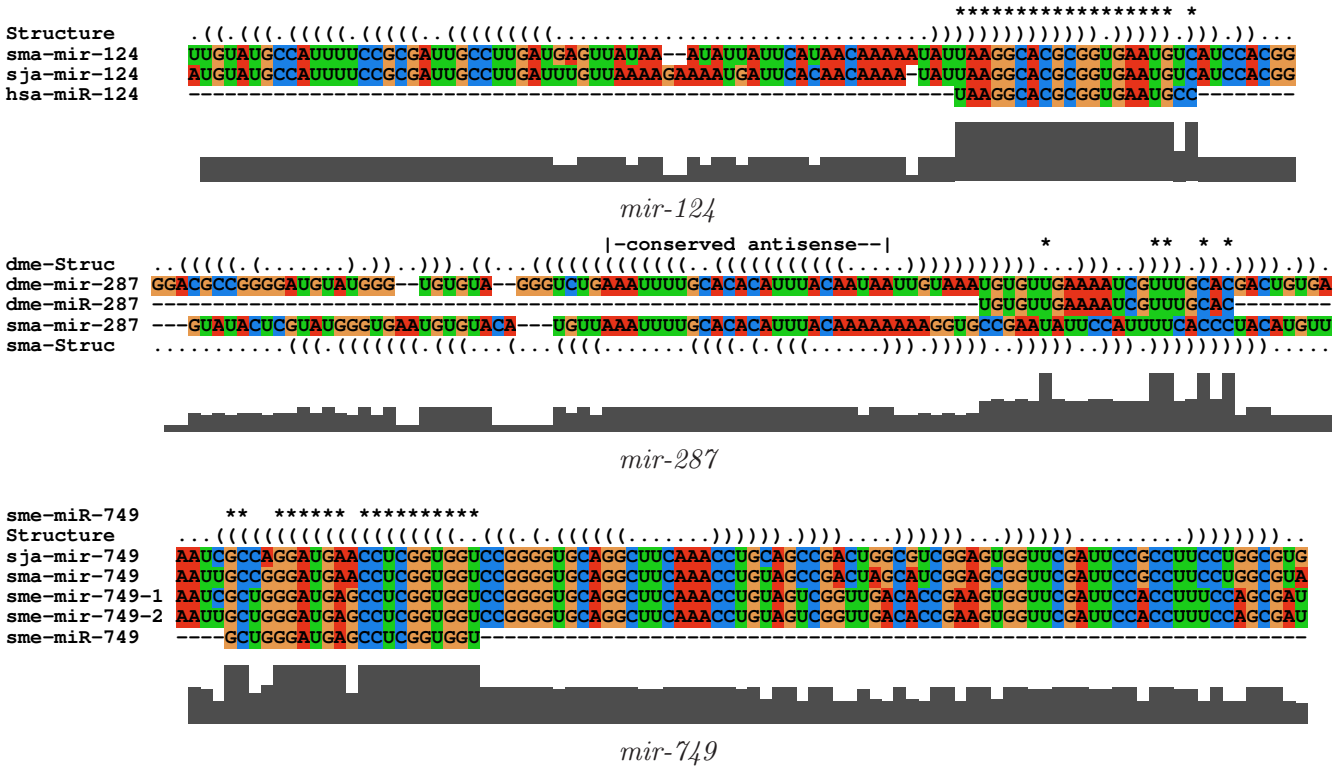

Figure 3

Multiple sequence alignments of the *pre*-miRNAs that were computationally found in *S. mansoni*. For *mir-124* and *mir-749* the sequences share a common consensus structure. The uncertain *mir-287* candidate clusters together with *mir-124* in insect genomes. However, though it also exhibits a single stem-loop structure, it is different from that of insects. Here the sequence is only conserved at the antisense region of the annotated mature miRNA.

A comparison of the predicted snoRNAs with the entries in the Rfam[61] and NONCODE[62] databases returned only 47 hits that match to several other RNAs like tRNAs, parts of the rRNA operon, snRNAs, mRNAlike genes and a few of our candidates map to the hammerhead ribozyme. These sequences are likely false positives and have been removed from the candidate list. The number of predicted candidates is much larger than the number of snoRNAs reported in other organisms; for instance [59] lists 456 for the human genome. Although we most likely do not yet know the full snoRNA complement of eukaryotic genomes, we have to expect that a large fraction of prediction will turn out to be false positives.

We therefore analysed the conservation of the candidates in *S. japonicum* and focussed on the snoRNA candidates with targets in the 18S, 28S and/or 5.8S ribosomal RNA. While targets are predicted for more than half of the candidates, see Table 2, the numbers are drastically reduced when conservation of the candidates in *S. japonicum* is required. Note, furthermore, that the fraction of conserved candidates is strongly enriched among those with ribosomal RNA targets, indicating that these sets are likely to contain a sizeable fraction of true positives. This filtering step leaves us with 227 box C/D and 352 box H/ACA snoRNA candidates. While still high, these numbers fall into the expected range for a metazaon snoRNA complement.

Table 2: Conservation and target prediction of snoRNA candidates.

snoReporttargets	Box C/D (snoscan)			Box H/ACA (RNAsnoop)		
	≥ 2	1	0	≥ 2	1	0
predicted in *S. mansoni*	926	110	613	284	495	177
conserved in *S. japonicum*	200	27	83	149	203	62

Only ribosomal RNAs were searched for putative target sites.

We remark, finally, that five of the snoRNA candidates (three box C/D and two box H/ACA) are also conserved in *Schmidtea mediterranea*.

Other RNA motifs

Two examples of relatively well-known schistosome non-coding RNAs are the hammerhead ribozyme motifs within the Sm-α and Sj-α SINE-like elements [27,28]. A `blastn` search of the hammerhead ribozyme motif from the `Rfam` database resulted in ~38,500 candidates for *S. mansoni* in contrast to ~5,000 candidates for *S. japonicum*. While high, this number is not surprising considering the generally high copy number of SINE elements; previously, the copy number for Sm-α elements in the *S. mansoni* genome was estimated to exceed 10,000 [27]. The highly conserved potassium channel RNA editing signal [63,64] is another structured RNA element that was described previously [65]. We found nine copies of this hairpin structure in the *S. mansoni* genome assembly and three in *S. japonicum*.

Uncertain and missing candidates

Both the MRP RNA [2,3,66] and the 7SK RNA [4,5,67] have highly variable, rapidly evolving sequences that make them difficult or impossible to detect in invertebrate genomes. Their ancient evolutionary origin and their extremely conserved molecular house-keeping functions make it more than likely that they are present in the schistosome genomes as well. In both cases, we have not been able to identify unambiguous homologs. There are, however, plausible candidates. We briefly describe them in the following paragraphs since they may warrant further attention and may be a useful starting point for subsequent experimental studies, as exemplified by the history of discovery of the snRNA in *Giardia intestinalis* [68-70].

MRP RNA has multiple functions, among them mitochondrial RNA processing and nucleolar pre-rRNA processing. The *S. mansoni* MRP candidate fits the general secondary structure model of metazoan MRP RNAs [2,3,66] and analysis with `RNAduplex` shows that the candidate contains a pseudoknot which exhibited striking sequence identity with known MRPs. The locus is well-conserved in *S. japonicum*. On the other hand, stems 1 and 12 were divergent compared to known MRPs, and stem 19 also fails to display clear similarities with known MRPs. Although quite likely a true MRP homolog, we therefore consider this sequence only tentative.

7SK RNA is a general transcriptional regulator, repressing transcript elongation through inhibition of transcription elongation factor PTEFb and also suppresses the deaminase activity of APOBEC3C [71]. The *S. mansoni* 7SK candidate has a 5' stem similar to that described in other invertebrates [5], and parts of the middle of the sequence are also recognizable. There is, furthermore, a homologous locus in the genome of *S. japonicum*. However, the 3' stem (which was followed by a poly-T terminator) was not conserved. In addition, a large sequence deletion was evident.

Three major classes of ncRNAs were expected, but not found, in the *S. mansoni* genome. As in all other invertebrates genomes, no candidate sequence was found for a telomerase RNA. *S. mansoni* almost certainly has a canonical telomerase holoenzyme, since it encodes telomerase proteins (Smp_066300 and Smp_066290) and has the same telomeric repeat sequences as many other metazoan animals [72]. Telomerase RNAs are notoriously difficult to find, as they are highly divergent among different species, varying in both size and sequence composition [7,73]. Vault RNAs are known in all major deuterostome lineages [74], and homologs were recently also described in two lophotrochozoan lineages [75]. Since *S. mansoni* has a homolog of the major vault protein (Smp_006740) we would also expect a corresponding RNA component to be present. So far, Y RNAs have been found only in vertebrates [76,77] and in nematodes [78,79], although the Ro RNP, that they are associated with, seems to be present in most or even all eukaryotes.

Conclusion

We have described here a detailed annotation of "housekeeping" ncRNAs in the genomes of the parasitic platyhelminth *Schistosoma mansoni* and *Schistosoma japonicum*. Limited to the best conserved structured RNAs, our work nevertheless uncovered important genomic features such as the existence of a SINE family specific to *Schistosoma mansoni*, which is derived from tRNA-Gln-TTG. Our data furthermore establish the presence of a minor spliceosome in schistosomes and confirms spliced leader *trans*-splicing. With a coverage of at least 90-95% of the genomic DNA, missing data are not a significant problem. The fragmented genome assemblies, however, preclude accurate counts of the multi-copy genes.

Platyhelminths are known to be a fast-evolving phylum [80]. It is not surprising therefore that in particular the small ncRNAs are hard or impossible to detect by simple homology search tools such as `blastn`. Even specialized tools have been successful in identifying only the better conserved genes such as tRNA, microRNAs, RNase P RNA, SRP RNA. The notoriously poorly conserved families, such as snoRNAs, telomerase RNA, or vault RNAs, mostly escaped detection.

The description of several novel, and in many cases quite derived, schistosome ncRNAs contributes significantly to the understanding of the evolution of the corresponding RNA families. The schistosome ncRNA sequences, furthermore, are an important input to subsequent homology search projects, since they allow the construction of

improved descriptors for sequence/structure-based search algorithms. Last but not least, the ncRNA annotation tracks are an important contribution to the genome-wide annotation datasets of both *S. mansoni* and *S. japonicum*. It not only contributes the protein-based annotation but also helps to identify annotation errors, e.g. cases where putative proteins are annotated that overlap rRNA operons or other ncRNAs.

The house-keeping ncRNAs considered in this study are almost certainly only the proverbial tip of the platyhelminth ncRNAs iceberg. The discovery of a large number of mRNA-like ncRNAs (mlncRNAs) in many eukaryotes (compiled e.g. in the RNAdb[81] and reviewed e.g. in [1]), and in particular in many other invertebrate species (nematodes [82], insects [83,84]) suggests that similar transcripts will also be abundant in schistosomes. The abundant EST data for both schistosome species [85,86] can provide a starting point e.g. for an analysis along the lines of [87]. Computational surveys, furthermore, have provided evidence for large numbers of RNAs with conserved secondary structures in other invertebrates [88-90]. The underlying methods, such as RNAz[91], are inherently comparative, presenting difficulties for application to schistosome genomes due to the large evolutionary distance between schistosome and non-schistosome genomes. This is also the case for a recent approach to identify mRNA-like non-coding RNAs with very low levels of sequence conservation based on their intron structure [92]. A deeper understanding of the non-coding transcriptome of schistosomes will therefore have to rely primarily on experimental approaches, either by means of tiling arrays or by means of high throughput transcriptome sequencing.

Methods

tRNA annotation

We used tRNAscan-SE[93] with default parameters to annotate putative tRNA genes. As additional confirmation, the genome sequence was searched using tRNA consensus sequences from the Rfam database [61]. In order to obtain suitable data for comparison, the genome of the free-living platyhelminth *Schmidtea mediterranea* [17] was searched alongside that of *S. mansoni* and *S. japonicum*.

microRNA annotation

We followed the general protocol outlined in [8] to identify miRNA precursors, using all metazoan miRNAs listed in miRBase [94] [Release 11.0, http://microrna.sanger.ac.uk/sequences/]. The initial search was conducted by blastn with $E < 0.01$ with the mature and mature* miRNAs as query sequences. The resulting candidates were then extended to the length of the precursor sequence of the search query and aligned to the precursors using ClustalW[95]. Secondary structures were predicted using RNAfold[96] for single sequences and RNAalifold[97] for alignments. Candidates that did not fold into miRNA-like hairpin structures were discarded. The remaining sequences were then examined by eye to see if the mature miRNA was well-positioned in the stem portion of each putative precursor sequence. In addition, we used the final candidates to search the *S. japonicum* and *S. mediterranea* genomes to examine whether these sequences are conserved in Schistosoma and/or Platyhelminthes.

snoRNA annotation

We compared all the known human and yeast snoRNAs that are annotated in the snoRNAbase[98] to the *S. mansoni* genome using BLAST[99] and GotohScan[8]. The search for novel snoRNA candidates was performed only on sequences that were not annotated as protein-coding or another ncRNA in the current *S. mansoni* assembly. The SnoReport program [100] was used to identify putative box C/D and box H/ACA snoRNAs on both strands. Only the best predictions, i.e., those that show highly conserved boxes and canonical structural motifs, were kept for further analysis. The remaining candidates are further analysed for possible target interactions with ribosomal RNAs using snoscan[101] for box C/D and RNAsnoop[102] for box H/ACA snoRNA candidates. In addition, the sequences were checked for conservation in *S. japonicum* and *S. mediterranea* using BLAST. To estimate the number of false predictions we compared the candidate snoRNAs with common ncRNA databases, in particular Rfam[61] and NONCODE[62]. All sequences that match a non-snoRNA ncRNA were discarded.

Other RNA families

For other families, we employed the following five steps:

(a) For candidate sequences of ribosomal RNAs, spliceosomal RNAs, the spliced leader (SL) and the SRP RNA, we performed BLAST searches with $E < 0.001$ using the known ncRNA genes from the NCBI and Rfam databases. For the snRNA set, see [44]. For 7SL RNA we used ***X04249***, for 5S and 5.8S rRNAs we used the complete set of Rfam entries, for the SSU and LSU rRNAs, we used ***Z11976*** and ***NR_003287***, respectively. The SL RNAs were searched using SL RNA entries from Rfam and the sequences reported in [26]. For more diverged genes such as minor snRNAs, RNase MRP, 7SK, and RNase P, we used GotohScan[8], an implementation of a full dynamic programming alignment with affine gap costs. In cases where no good candidates were found we also employed descriptor-based search tools such as rnabob http://selab.janelia.org/software.html.

(b) In a second step, known and predicted sequences were aligned using ClustalW[95] and visualized with ClustalX[103]. To identify functional secondary structure, RNAfold, RNAalifold, and RNAcofold[104] were

used. Combined primary and secondary structures were visualized using `stockholm-format` alignment files in the `emacs` editor utilizing `ralee` mode [105]. Alignments are provided at the Supplemental Data online.

(c) Putatively functional sequences were distinguished from likely pseudogenes by analysis of flanking genomic sequence. To this end, the flanking sequences of snRNA and SL RNA copies were extracted and analyzed for conserved sequence elements using `MEME`[106]. Only snRNAs with plausible promoter regions were reported.

(d) Additional consistency checks were employed for individual RNA families, including phylogenetic analysis by neighbor-joining [107] to check that candidate sequences fall at phylogenetically reasonable positions relative to previously known homologs. For RNase MRP RNA candidates, `RNAduplex` http://www.tbi.univie.ac.at/RNA/RNAduplex.html was used to find the pseudoknot structure. In order to confirm that the SL RNA candidate is indeed *trans*-spliced to mRNA transcripts, we searched the *FAPESP Genoma Schistosoma mansoni* website for ESTs including fragments of the predicted SL RNA. We found 52 ESTs with `blastn`$E < 0.001$ that span the predicted region of the SL RNA (nt 8-38), indicating that this RNA does indeed function as a spliced leader.

(e) Accepted candidate sequences were used as `BLAST` queries against the *S. mansoni* genome to determine their copy number in the genome assembly.

Additional Data Online

The website http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/08-014 provides extensive machine readable information, including sequence files, alignments, and genomic coordinates.

Authors' contributions

CSC, PB, and PFS designed the study. CSC, MM, DR, JH, CBS, SK, CSA, and PFS performed the computational analyses. CSC wrote the first draft of the manuscript. All authors contributed to the final assessment of the data as well as the writing of the final version of the manuscript. CSC, MM, DR, JH should be considered as joint first authors.

Additional material

Additional file 1

***Supplemental figures and captions**. contains supplemental Figures S1 - S4 mentioned in the main text.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-464-S1.PDF]

Acknowledgements

This work was supported in part by the European Union through grants in the 6th and 7th Framework Programe of the European Union (projects EMBIO, SYNLET, and EDEN), the *Deutsche Forschungsgemeinschaft* and the auspices of SPP SPP-1174 "Deep Metazoan Phylogeny", the *Freistaat Sachsen*, and the DAAD-AleCol program.

References

1. Amaral PP, Dinger ME, Mercer TR, Mattick JS: **The eukaryotic genome as an RNA machine.** *Science* 2008, **319:**1787-1789.
2. Piccinelli P, Rosenblad MA, Samuelsson T: **Identification and analysis of ribonuclease P and MRP RNA in a broad range of eukaryotes.** *Nucleic Acids Res* 2005, **33:**4485-4495.
3. Woodhams MD, Stadler PF, Penny D, Collins LJ: **RNAse MRP and the RNA Processing Cascade in the Eu-karyotic Ancestor.** *BMC Evol Biol* 2007, **7:**S13.
4. Gruber AR, Koper-Emde D, Marz M, Tafer H, Bernhart S, Obernosterer G, Mosig A, Hofacker IL, Stadler PF, Benecke BJ: **Invertebrate 7SK snRNAs.** *J Mol Evol* 2008, **107-115:**66.
5. Gruber A, Kilgus C, Mosig A, Hofacker IL, Hennig W, Stadler PF: **Arthropod 7SK RNA.** *Mol Biol Evol* 2008, **1923-1930:**25.
6. Chen JL, Blasco MA, Greider CW: **Secondary Structure of Vertebrate telomerase RNA.** *Cell* 2000, **100:**503-514.
7. Xie M, Mosig A, Qi X, Li Y, Stadler PF, Chen JJL: **Size Variation and Structural Conservation of Vertebrate Telomerase RNA.** *J Biol Chem* 2008, **283:**2049-2059.
8. Hertel J, de Jong D, Marz M, Rose D, Tafer H, Tanzer A, Schierwater B, Stadler PF: **Non-Coding RNA Annotation of the Genome of *Trichoplax adhaerens*.** *Nucleic Acids Res* 2009, **37:**1602-1615.
9. Blair D, Davis GM, Wu B: **Evolutionary relationships between trematodes and snails emphasizing schistosomes and paragonimids.** *Parasitology* 2001, **123(Suppl):**S229-S243.
10. Brant SV, Loker ES: **Can specialized pathogens colonize distantly related hosts? Schistosome evolution as a case study.** *PLoS Pathog* 2005, **1:**167-169.
11. Webster BL, Southgate VR, Littlewood DTJ: **A revision of the interrelationships of Schistosoma including the recently described Schistosoma guineensis.** *Int J Parasitol* 2006, **36:**947-955.
12. Jiménez-Guri E, Philippe H, Okamura B, Holland PWH: ***Buddenbrockia* is a cnidarian worm.** *Science* 2007, **317:**116-118.
13. Wilson RA, Ashton PD, Braschi S, Dillon GP, Berriman M, Ivens A: **'Oming in on schistosomes: prospects and limitations for post-genomics.** *Trends Parasitol* 2007, **23:**14-20.
14. Berriman M, Haas BJ, LoVerde PT, Wilson RA, Dillon GP, Cerqueira GC, Mashiyama ST, Al-Lazikani B, Andrade LF, Ashton PD, Aslett MA, Bartholomeu DC, Blandin G, Caffrey CR, Coghlan A, Coulson R, Day TA, Delcher A, DeMarco R, Djikeng A, Eyre T, Gamble JA, Ghedin E, Gu Y, Hertz-Fowler C, Hirai H, Hirai Y, Houston R, Ivens A, Johnston DA, Lacerda D, Macedo CD, McVeigh P, Ning Z, Oliveira G, Overington JP, Parkhill J, Pertea M, Pierce RJ, Protasio AV, Quail M, Rajandream MA, Rogers J, Sajid M, Salzberg SL, Stanke M, Tivey AR, White O, Williams DL, Wortman J, Wu W, Zamanian M, Zerlotini A, Fraser-Liggett CM, Barrell BG, El-Sayed NM: **The genome of the blood fluke *Schistosoma mansoni*.** *Nature* 2009, **460:**352-358.
15. Schistosoma japonicum Genome Sequencing and Functional Analysis Consortium: **The *Schistosoma japonicum* genome reveals features of host-parasite interplay.** *Nature* 2009, **460:**345-351.
16. Hirai H, Taguchi T, Saitoh Y, Kawanaka M, Sugiyama H, Habe S, Okamoto M, Hirata M, Shimada M, Tiu WU, Lai K, Upatham ES, Agatsuma T: **Chromosomal differentiation of the *Schistosoma japonicum* complex.** *Int J Parasitol* 2000, **30:**441-452.
17. Robb SMC, Ross E, Alvarado AS: **`SmedGD`: the *Schmidtea mediterranea* genome database.** *Nucleic Acids Res* 2008, **36:**D599-D606.
18. Haas BJ, Berriman M, Hirai H, Cerqueira GG, Loverde PT, El-Sayed NM: ***Schistosoma mansoni* genome: closing in on a final gene set.** *Exp Parasitol* 2007, **117:**225-228.
19. Hu W, Yan Q, Shen DK, Liu F, Zhu ZD, Song HD, Xu XR, Wang ZJ, Rong YP, Zeng LC, Wu J, Zhang X, Wang JJ, Xu XN, Wang SY, Fu G, Zhang XL, Wang ZQ, Brindley PJ, McManus DP, Xue CL, Feng Z, Chen Z, Han ZG: **Evolutionary and biomedical implications of a *Schistosoma japonicum* complementary DNA resource.** *Nat Genet* 2003, **35:**139-147.

20. Verjovski-Almeida S, R D, Martins EA, Guimarães PE, Ojopi EP, Paquola AC, Piazza JP, Nishiyama MY Jr, Kitajima JP, Adamson RE, Ashton PD, Bonaldo MF, Coulson PS, Dillon GP, Farias LP, Gregorio SP, Ho PL, Leite RA, Malaquias LC, Marques RC, Miyasato PA, Nascimento AL, Ohlweiler FP, Reis EM, Ribeiro MA, Sá RG, Stukart GC, Soares MB, Gargioni C, Kawano T, Rodrigues V, Madeira AM, Wilson RA, Menck CF, Setubal JC, Leite LC, Dias-Neto E: **Transcriptome analysis of the acoelomate human parasite *Schistosoma mansoni*.** *Nat Genet* 2003, **35:**148-157.
21. Verjovski-Almeida S, Venancio TM, Oliveira KC, Almeida GT, DeMarco R: **Use of a 44k oligoarray to explore the transcriptome of *Schistosoma mansoni* adult worms.** *Exp Parasitol* 2007, **117:**236-245.
22. Schulmeister A, Heyers O, Morales ME, Brindley PJ, Lucius R, Meusel G, Kalinna BH: **Organization and functional analysis of the *Schistosoma mansoni* cathepsin D-like aspartic protease gene promoter.** *Biochim Biophys Acta* 2005, **1727:**27-34.
23. Copeland CS, Mann VH, Brindley PJ: **Both sense and antisense strands of the LTR of the *Schistosoma mansoni* Pao-like retrotransposon Sinbad drive luciferase expression.** *Mol Genet Genomics* 2007, **277:**161-170.
24. Brejová B, Vinaz T, Chen Y, Wang S, Zhoa G, Brown DG, Li M, Zhou Y: **Finding genes in *Schistosoma japonicum*: annotating novel genomes with help of extrinsic evidence.** *Nucleic Acids Res* 2009, **37:**e52.
25. Mourier T, Carret C, Kyes S, Christodoulou Z, Gardner PP, Jeffares DC, Pinches R, Barrell B, Berriman M, Griffiths-Jones S, Ivens A, Newbold C, Pain A: **Genome-wide discovery and verification of novel structured RNAs in *Plasmodium falciparum*.** *Genome Res* 2008, **18:**281-292.
26. Rajkovic A, Davis RE, Simonsen JN, Rottman FM: **A spliced leader is present on a subset of mRNAs from the human parasite *Schistosoma mansoni*.** *Proc Natl Acad Sci USA* 1990, **87:**8879-8883.
27. Ferbeyre G, Smith JM, Cedergren R: **Schistosome satellite DNA encodes active hammerhead ribozymes.** *Mol Cell Biol* 1998, **18:**3880-3888.
28. Laha T, McManus DP, Loukas A, Brindley PJ: **Sjα elements, short interspersed element-like retroposons bearing a hammerhead ribozyme motif from the genome of the oriental blood fluke *Schistosoma japonicum*.** *Biochim Biophys Acta* 2000, **1492:**477-482.
29. Copeland CS, Heyers O, Kalinna BH, Bachmair A, Stadler PF, Hofacker IL, Brindley PJ: **Structural and evolutionary analysis of the transcribed sequence of *Boudicca*, a *Schistosoma mansoni* retrotransposon.** *Gene* 2004, **329:**103-114.
30. Rollinson D, Kaukas A, Johnston DA, Simpson AJ, Tanaka M: **Some molecular insights into schistosome evolution.** *Int J Parasitol* 1997, **27:**11-28.
31. Littlewood DT, Lockyer AE, Webster BL, Johnston DA, Le TH: **The complete mitochondrial genomes of *Schistosoma haematobium* and *Schistosoma spindale* and the evolutionary history of mitochondrial genome changes among parasitic flatworms.** *Mol Phylogenet Evol* 2006, **39:**452-467.
32. DeMarco R, Verjovski-Almeida S: **Expressed Sequence Tags (ESTs) and Gene Discovery: *Schistosoma mansoni*.** *Bioinformatics in Tropical Disease Research: A Practical and Case-Study Approach* 2008:B06 [http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=bioinfo]. Bethesda, MD: National Library of Medicine
33. Sheppard K, Akochy PM, Söll D: **Assays for transfer RNA-dependent amino acid biosynthesis.** *Methods* 2008, **44:**139-145.
34. Ambrogelly A, Palioura S, Söll D: **Natural expansion of the genetic code.** *Nat Chem Biol* 2007, **3:**29-35.
35. Hubert N, Walczak R, Sturchler C, Myslinski E, Schuster C, Westhof E, Carbon P, Krol A: **RNAs mediating cotrans-lational insertion of selenocysteine in eukaryotic selenoproteins.** *Biochimie* 1996, **78:**590-596.
36. Coleman JR, Papamichail D, Skiena S, Futcher B, Wimmer E, Mueller S: **Virus attenuation by genome-scale changes in codon pair bias.** *Science* 2008, **320:**1784-1787.
37. Smit AFA, Hubley R, Green P: **`RepeatMasker`. [Version, open-3.2.5 [RMLib: 20080611].** [http://www.repeatmasker.org/].
38. Spotila LD, Hirai H, Rekosh DM, Lo Verde PT: **A retroposon-like short repetitive DNA element in the genome of the human blood fluke, *Schistosoma mansoni*.** *Chromosoma* 1989, **97:**421-428.
39. Simpson AJ, Dame JB, Lewis FA, McCutchan TF: **The arrangement of ribosomal RNA genes in *Schistosoma mansoni*. Identification of polymorphic structural variants.** *Eur J Biochem* 1984, **139:**41-45.
40. van Keulen H, Loverde PT, Bobek LA, Rekosh DM: **Organization of the ribosomal RNA genes in *Schistosoma mansoni*.** *Mol Biochem Parasitol* 1985, **15:**215-230.
41. Nei M, Rooney AP: **Concerted and birth-and-death evolution of multigene families.** *Annu Rev Genet* 2005, **39:**121-152.
42. Scheibye-Alsing K, Hoffmann S, Frankel AM, Jensen P, Stadler PF, Mang Y, Tommerup N, Gilchrist MJ, Hillig ABN, Cirera S, Jørgensen CB, Fredholm M, Gorodkin J: **Sequence Assembly.** *Comp Biol Chem* 2009, **33:**121-136.
43. Staley JP, Woolford JL Jr: **Assembly of ribosomes and spliceosomes: complex ribonucleoprotein machines.** *Curr Opin Cell Biol* 2009, **21:**109-118.
44. Marz M, Kirsten T, Stadler PF: **Evolution of Spliceosomal snRNA Genes in Metazoan Animals.** *J Mol Evol* 2008, **67:**594-607.
45. Kreivi JP, Lamond AI: **RNA splicing: unexpected spliceosome diversity.** *Curr Biol* 1996, **6:**802-805.
46. Pouchkina-Stantcheva NN, Tunnacliffe A: **Spliced leader RNA-mediated trans-splicing in phylum Rotifera.** *Mol Biol Evol* 2005, **22:**1482-1489.
47. Marz M, Vanzo N, Stadler PF: **Carnival of SL RNAs: Structural variants and the possibility of a common origin.** *J Bioinf Comp Biol* 2009 [http://www.bioinf.uni-leipzig.de/Publications/PREPRINTS/09-009.pdf].
48. McNair A, Zemzoumi K, Lütcke H, Guillerm C, Boitelle A, Capron A, Dissous C: **Cloning of a signal-recognition-particle subunit of *Schistosoma mansoni*.** *Parasitol Res* 1995, **81:**175-177.
49. Kirsebom LA: **RNase P RNA mediated cleavage: substrate recognition and catalysis.** *Biochimie* 2007, **89:**1183-1194.
50. Kikovska E, Svärd SG, Kirsebom LA: **Eukaryotic RNase P RNA mediates cleavage in the absence of protein.** *Proc Natl Acad Sci USA* 2007, **104:**2062-2067.
51. Williams AE: **Functional aspects of animal microRNAs.** *Cell Mol Life Sci* 2008, **65:**545-562.
52. Krautz-Peterson G, Skelly PJ: ***Schistosoma mansoni*: the dicer gene and its expression.** *Exp Parasitol* 2008, **118:**122-128.
53. Gomes MS, Cabral FJ, Jannotti-Passos LK, Carvalho O, Rodrigues V, Baba EH, Sá RG: **Preliminary analysis of miRNA pathway in *Schistosoma mansoni*.** *Parasitol Int* 2009, **58:**61-68.
54. Liu F, Lu J, Hu W, Wang SY, Cui SJ, Chi M, Yan Q, Wang XR, Song HD, Xu XN, Wang JJ, Zhang XL, Zhang X, Wang ZQ, Xue CL, Brindley PJ, McManus DP, Yang PY, Feng Z, Chen Z, Han ZG: **New perspectives on host-parasite interplay by comparative transcriptomic and proteomic analyses of *Schistosoma japonicum*.** *PLoS Pathog* 2006, **2:**e29.
55. Xue X, Sun J, Zhang Q, Wang Z, Huang Y, Pan W: **Identification and characterization of novel microRNAs from *Schistosoma japonicum*.** *PLoS ONE* 2008, **3:**e4034.
56. Palakodeti D, Smielewska M, Graveley BR: **MicroRNAs from the Planarian *Schmidtea mediterranea*: a model system for stem cell biology.** *RNA* 2006, **12:**1640-1649.
57. Palakodeti D, Smielewska M, Lu YC, Yeo GW, Graveley BR: **The PIWI proteins SMEDWI-2 and SMEDWI-3 are required for stem cell function and piRNA expression in planarians.** *RNA* 2008, **14:**1174-1186.
58. Matera AG, Terns R, Terns : **Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs.** *Nat Rev Mol Cell Biol* 2007, **8:**209-220.
59. Dieci G, Preti M, Montanini B: **Eukaryotic snoRNAs: A paradigm for gene expression flexibility.** *Genomics* 2009, **94:**83-88.
60. Lukowiak AA, Granneman S, Mattox SA, Speckmann WA, Jones K, Pluk WJ, Venrooij Hand, Terns RM, Terns MP: **Interaction of the U3-55k protein with U3 snoRNA is mediated by the box B/C motif of U3 and the WD repeats of U3-55k.** *Nucleic Acids Res* 2000, **28:**3462-3471.
61. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A: **`Rfam`: annotating non-coding RNAs in complete genomes.** *Nucleic Acids Res* 2005, **33:**D121-D124.
62. Liu C, Bai B, Skogerbø G, Cai L, Deng W, Zhang Y, Bu DB, Zhao Y, Chen R: **`NONCODE`: an integrated knowledge database of non-coding RNAs.** *Nucleic Acids Res* 2005, **33:**D112-D115.

63. Bhalla T, Rosenthal JJC, Holmgren M, Reenan R: **Control of human potassium channel inactivation by editing of a small mRNA hairpin.** *Nature Struct Mol Biol* 2004, **11:**950-956.
64. Yang Y, Lv J, Gui B, Yin H, Wu X, Zhang Y, Jin Y: **A-to-I RNA editing alters less-conserved residues of highly conserved coding regions: implications for dual functions in evolution.** *RNA* 2008, **14:**1516-1525.
65. Kim E, Day TA, Bennett JL, Pax RA: **Cloning and functional expression of a Shaker-related voltage-gated potassium channel gene from *Schistosoma mansoni* (Trematoda: Digenea).** *Parasitology* 1995, **110:**171-180.
66. López MD, Rosenblad MA, Samuelsson T: **Conserved and variable domains of RNase MRP RNA.** *RNA Biology* 2009, **6:**208-221.
67. Marz M, Donath A, Verstraete N, Nguyen VT, Stadler PF, Bensaude O: **Evolution of 7SK RNA and its Protein Partners in Metazoa.** *Mol Biol Evol* 2009 in press.
68. Collins LJ, Poole AM, Penny D: **Using ancestral sequences to uncover potential gene homologues.** *Appl Bioin-formatics* 2003, **2(Suppl 3):**85-95.
69. Chen XS, Rozhdestvensky TS, Collins LJ, Schmitz J, Penny D: **Combined experimental and computational approach to identify non-protein-coding RNAs in the deep-branching eukaryote *Giardia intestinalis*.** *Nucleic Acids Res* 2007, **35:**4619-4628.
70. Chen XS, White WT, Collins LJ, Penny D: **Computational identification of four spliceosomal snRNAs from the deep-branching eukaryote *Giardia intestinalis*.** *PLoS One* 2008, **3(8):**e3106.
71. Barrandon C, Spiluttini B, Bensaude O: **Non-coding RNAs regulating the transcriptional machinery.** *Biol Cell* 2008, **100:**83-95.
72. Hirai H, LoVerde PT: **Identification of the telomeres on *Schistosoma mansoni* chromosomes by FISH.** *J Parasitol* 1996, **82:**511-512.
73. Theimer CA, Feigon J: **Structure and function of telomerase RNA.** *Curr Opin Struct Biol* 2006, **16:**307-318.
74. Stadler PF, Chen JJL, Hackermüller J, Hoffmann S, Horn F, Khaitovich P, Kretzschmar AK, Mosig A, Prohaska SJ, Qi X, Schutt K, Ullmann K: **Evolution of Vault RNAs.** *Mol Biol Evol* 2009, **26:**1975-1991.
75. Mosig A, Zhu L, Stadler PF: **Strategies for Homology-Based ncRNA Gene Annotation.** *Brief Funct Genomics Proteomics* 2009 in press.
76. Mosig A, Guofeng M, Stadler BMR, Stadler PF: **Evolution of the Vertebrate Y RNA Cluster.** *Th Biosci* 2007, **126:**9-14.
77. Perreault J, Perreault JP, Boire G: **Ro-associated Y RNAs in metazoans: evolution and diversification.** *Mol Biol Evol* 2007, **24:**1678-1689.
78. Van Horn DJ, Eisenberg D, O'Brien CA, Wolin SL: **Caenorhabditis elegans embryos contain only one major species of Ro RNP.** *RNA* 1995, **1:**293-303.
79. Boria I, Gruber AR, Tanzer A, Bernhart S, Lorenz R, Mueller MM, Hofacker IL, Stadler PF: **Nematode sbRNAs: homologs of vertebrate Y RNAs.** *Tech. Rep. BIOINF-09-020* 2009 [http://www.bioinf.uni-leipzig.de/Publications/PREPRINTS/09-020.pdf]. Bioinformatics, University of Leipzig
80. Lartillot N, Brinkmann H, Philippe H: **Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model.** *BMC Evolutionary Biology* 2007, **7:**S4.
81. Pang KC, Stephen S, Dinger ME, Engström PG, Lenhard B, Mattick JS: `RNAdb 2.0` **-- an expanded database of mammalian non-coding RNAs.** *Nucleic Acids Res* 2007, **35:**D178-D182.
82. Shin H, Hirst M, Bainbridge MN, Magrini V, Mardis E, Moerman DG, Marra MA, Baillie DL, Jones SJ: **Transcriptome analysis for *Caenorhabditis elegans* based on novel expressed sequence tags.** *BMC Biol* 2008, **6:**30.
83. Inagaki S, Numata K, Kondo T, Tomita M, Yasuda K, Kanai A, Kageyama Y: **Identification and expression analysis of putative mRNA-like non-coding RNA in Drosophila.** *Genes Cells* 2005, **10:**1163-1173.
84. Tupy JL, Bailey AM, Dailey G, Evans-Holm M, Siebel CW, Misra S, Celniker SE, Rubin GM: **Identification of putative noncoding polyadenylated transcripts in *Drosophila melanogaster*.** *Proc Natl Acad Sci USA* 2005, **102:**5495-5500.
85. Zerlotini A, Heiges M, Wang H, Moraes RL, Dominitini AJ, Ruiz JC, Kissinger JC, Oliveira G: `SchistoDB`**: a *Schistosoma mansoni* genome resource.** *Nucleic Acids Res* 2009, **37:**D579-D582.
86. Liu F, Chen P, Cui SJ, Wang ZQ, Han ZG: `SjTPdb`**: integrated transcriptome and proteome database and analysis platform for *Schistosoma japonicum*.** *BMC Genomics* 2008, **9:**304.
87. Seemann SE, Gilchrist MJ, Hofacker IL, Stadler PF, Gorodkin J: **Detection of RNA structures in porcine EST data and related mammals.** *BMC Genomics* 2007, **8:**316.
88. Missal K, Rose D, Stadler PF: **Non-coding RNAs in *Ciona intestinalis*.** *Bioinformatics* 2005, **21(S2):**i77-i78.
89. Missal K, Zhu X, Rose D, Deng W, Skogerbø G, Chen R, Stadler PF: **Prediction of Structured Non-Coding RNAs in the Genome of the Nematode *Caenorhabitis elegans*.** *J Exp Zool: Mol Dev Evol* 2006, **306B:**379-392.
90. Rose DR, Hackermüller J, Washietl S, Findeiß S, Reiche K, Hertel J, Stadler PF, Prohaska SJ: **Computational RNomics of Drosophilids.** *BMC Genomics* 2007, **8:**406.
91. Washietl S, Hofacker IL, Stadler PF: **Fast and reliable prediction of noncoding RNAs.** *Proc Natl Acad Sci USA* 2005, **102:**2454-2459.
92. Hiller M, Findeiß S, Lein S, Marz M, Nickel C, Rose D, Schulz C, Backofen R, Prohaska SJ, Reuter G, Stadler PF: **Conserved Introns Reveal Novel Transcripts in Drosophila melanogaster.** *Genome Res* 2009, **19:**1289-1300.
93. Lowe T, Eddy S: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucl Acids Res* 1997, **25:**955-964.
94. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ: `miRBase`**: tools for microRNA genomics.** *Nucleic Acids Res* 2008, **36:**D154-D158.
95. Thompson JD, Higgs DG, Gibson TJ: **CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice.** *Nucl Acids Res* 1994, **22:**4673-4680.
96. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P: **Fast Folding and Comparison of RNA Secondary Structures.** *Monatsh Chem* 1994, **125:**167-188.
97. Hofacker IL, Fekete M, Stadler PF: **Secondary Structure Prediction for Aligned RNA Sequences.** *J Mol Biol* 2002, **319:**1059-1066.
98. Lestrade L, Weber MJ: **snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs.** *Nucleic Acids Res* 2006, **34:**D158-D162.
99. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215:**403-410.
100. Hertel J, Hofacker IL, Stadler PF: `snoReport`**: Computational identification of snoRNAs with unknown targets.** *Bioinformatics* 2008, **24:**158-164.
101. Lowe TM, Eddy SR: **A Computational Screen for Methylation Guide snoRNAs in Yeast.** *Science* 1999, **283:**1168-1171.
102. Tafer H, Kehr S, Hertel J, Stadler PF: `RNAsnoop`**: Efficient target prediction for box H/ACA snoRNAs.** *Tech. Rep. BIOINF-09-025* 2009 [http://www.bioinf.uni-leipzig.de/Publications/PREPRINTS/09-025.pdf]. Bioinformatics, University of Leipzig
103. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools.** *Nucleic Acids Res* 1997, **25:**4876-4882.
104. Bernhart SH, Tafer H, Mückstein U, Flamm C, Stadler PF, Hofacker IL: **Partition Function and Base Pairing Probabilities of RNA Heterodimers.** *Algorithms Mol Biol* 2006, **1:**3.
105. Griffiths-Jones S: `RALEE` **--RNA ALignment editor in** `Emacs`**.** *Bioinformatics* 2005, **21:**257-259.
106. Bailey TL, Williams N, Misleh C, Li WW: **MEME: discovering and analyzing DNA and protein sequence motifs.** *Nucleic Acids Res* 2006, **34:**W369-W373.
107. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4:**406-425.
108. Hofacker IL: **Vienna RNA secondary structure server.** *Nucleic Acids Res* 2003, **31:**3429-3431.
109. Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A, Flamm C, Hofacker IL, Stadler PF: **The Students of Bioinformatics Computer Labs 2004 and 2005: The Expansion of the Metazoan MicroRNA Repertoire.** *BMC Genomics* 2006, **7:**15.