MICROBIAL GENOMICS

RESEARCH ARTICLE

Li *et al.*, *Microbial Genomics* 2022;8:000821

DOI 10.1099/mgen.0.000821

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

Small and intermediate size structural RNAs in the unicellular parasite *Cryptosporidium parvum* as revealed by sRNA-seq and comparative genomics

Yiran Li[1], Rodrigo P. Baptista[1,2]†, Xiaohan Mei[3] and Jessica C. Kissinger[1,2,4,*]

Abstract

Small and intermediate-size noncoding RNAs (sRNAs and is-ncRNAs) have been shown to play important regulatory roles in the development of several eukaryotic organisms. However, they have not been thoroughly explored in *Cryptosporidium parvum*, an obligate zoonotic protist parasite responsible for the diarrhoeal disease cryptosporidiosis. Using Illumina sequencing of a small RNA library, a systematic identification of novel small and is-ncRNAs was performed in *C. parvum* excysted sporozoites. A total of 79 novel is-ncRNA candidates, including antisense, intergenic and intronic is-ncRNAs, were identified, including 7 new small nucleolar RNAs (snoRNAs). Expression of select novel is-ncRNAs was confirmed by RT-PCR. Phylogenetic conservation was analysed using covariance models (CMs) in related *Cryptosporidium* and apicomplexan parasite genome sequences. A potential new type of small ncRNA derived from tRNA fragments was observed. Overall, a deep profiling analysis of novel is-ncRNAs in *C. parvum* and related species revealed structural features and conservation of these novel is-ncRNAs. Covariance models can be used to detect is-ncRNA genes in other closely related parasites. These findings provide important new sequences for additional functional characterization of novel is-ncRNAs in the protist pathogen *C. parvum*.

DATA SUMMARY

The *Cryptosporidium parvum* IOWA (Bunch Grass Farm – BGF) raw small RNA-seq data generated in this study were deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) repository under BioProject accession number PRJNA773742 and SRA number SRR16563270. Supplementary Material can be found with the online version of this article. The constructed covariance models can be accessed from figshare (https://doi.org/10.6084/m9.figshare.17056367) .

INTRODUCTION

Noncoding RNAs (ncRNAs) are RNA molecules that do not encode proteins but are essential components of the transcriptome. The first identified ncRNAs, discovered in the 1950s, were ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs), with small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs) being discovered later [1]. These ncRNAs have conserved secondary structures and are ubiquitously expressed [2]. The ncRNA world is being reshaped by next-generation sequencing technologies, which are revealing additional novel ncRNAs.

Received 30 November 2021; Accepted 01 April 2022; Published 10 May 2022

Author affiliations: [1]Institute of Bioinformatics, University of Georgia, Athens, GA, USA; [2]Center for Tropical and Emerging Global Diseases, University of Georgia, Athens, GA, USA; [3]Department of Physiology and Pharmacology, University of Georgia, Athens, GA, USA; [4]Department of Genetics, University of Georgia, Athens, GA, USA.

***Correspondence:** Jessica C. Kissinger, jkissing@uga.edu

Keywords: ncRNA; is-ncRNA; snoRNA; covariance model; tRNA-derived small RNA; Apicomplexa.

Abbreviations: AGO, argonaute; AS, antisense; CM, covariance model; is-ncRNAs, intermediate-size noncoding RNAs; miRNAs, microRNA; ncRNA, non-coding RNA; piRNA, Piwi-interacting RNA; RNAi, RNA interference; rRNA, ribosomal RNA; RT-PCR, reverse transcription-polymerase chain reaction; SCI, structural conservation index; snoRNAs, small nucleolar RNAs; snRNAs, small nuclear RNA; sRNA-seq, small RNA-seq; SRP, signal recognition particle; SVM, support vector machine; tiRNAs, tRNA halves; tRFs, tRNA derived fragments; tRNA, transfer RNA; tsRNA, tRNA-derived small RNA.

†Present address: Houston Methodist Research Institute, Houston, TX, USA.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. Three supplementary figures and seven supplementary tables are available with the online version of this article.

Impact Statement

ncRNAs are often overlooked in genomic studies, and they are missing from the annotation of most apicomplexan genome sequences. ncRNAs are proving to be of interest because of their growing role in host–parasite interactions and gene regulation. Recent studies on *Cryptosporidium* were able to detect novel lncRNAs. *Cryptosporidium* lacks the genes for the RNAi pathway, raising questions regarding the spectrum of small ncRNAs such as miRNAs and other sRNAs. sRNA-seq is an approach focused on small RNA molecules. Pairing sRNA-seq with newer computational methods creates an opportunity to detect sncRNAs more accurately, enabling the community to further characterize and appreciate their role in *Cryptosporidium* biology.

Based on transcript length, ncRNAs are broadly classified as small (<200 nt) and long (>200 nt) [3]. An important category of small ncRNA, microRNAs (miRNAs) (~22 nt) have been demonstrated to regulate mRNA translation via the RNA interference pathway (RNAi) in a broad range of organisms [4]. tRNA- and snoRNA-derived small ncRNAs (18–40 nt) were discovered as ancient RNA regulatory molecules that exist in eukaryotes [5, 6]. Long ncRNAs are known to participate in numerous biological processes, including transcription, RNA processing, translation and epigenetic modification [7–10]. However, some important classes of ncRNAs have transcript lengths spanning the border between small and long RNA. These ncRNAs include snoRNAs (60–300 nt), whose function has been well established, as well as new variants that have emerged with deep RNA sequencing [8]. In *Caenorhabditis elegans*, many ncRNAs in the size range of 70–500 nt were detected by tiling array. Those ncRNAs showed male-specific expression, stage-specific expression and conservation in other nematodes, suggesting roles in developmental and sex-associated processes [11–13]. ncRNAs 50–500 nt in size have begun to draw attention in recent years. They are aptly referred to as 'intermediate-size ncRNAs' (is-ncRNAs) [11–13]. It has been suggested that transcripts in this size range exert their functions in stable ribonucleoproteins [14]. Intriguingly, many is-ncRNAs are not polyadenylated and will be excluded from polyA-selected RNA libraries, the most commonly used method used to detect lncRNAs. Thus, given their distinct size and potential critical functions, is-ncRNAs need to be studied systemically.

Cryptosporidium is an obligate zoonotic protist parasite that spreads via an oral–faecal route. It causes a diarrhoeal disease called cryptosporidiosis. In recent years, *Cryptosporidium* has been identified as the second most prevalent diarrhoeal pathogen of infants globally after rotavirus [15] and a leading cause of waterborne disease among humans [16]. *Cryptosporidiosis* can be lethal in immunocompromised individuals [17]. *Cryptosporidium parvum* is a medically important zoonotic pathogen species that, together with *Cryptosporidium hominis*, causes the majority of human cryptosporidiosis infections. The first *C. parvum* genome sequence was generated in 2004 [18]. Since this milestone, our molecular understanding of this parasite has progressed significantly in terms of metabolic capabilities [19] and mRNA gene expression during development [20–22]. Genetics and molecular genetics are beginning for *Cryptosporidium* [23], but the community is still far from understanding gene regulation in this dense ~4000 gene 9.2 Mb genome. Interestingly, the RNAi pathway, which is a critical component of post-transcriptional regulation in most eukaryotes, is considered lost in *Cryptosporidium* due to the absence of detectable dicer and argonaute (AGO) genes [24]. This loss creates the possibility that additional, yet to be characterized, ncRNAs may compensate. Thus far, no systematic genome-wide study of small and intermediate size ncRNAs in *Cryptosporidium* has been performed. One *in silico* study of miRNAs based on ESTs has been performed [25].

Both *in vitro* and *in silico* approaches have been implemented successfully to characterize ncRNAs in other organisms [12, 26–28]. Each approach has its advantages and limitations. *In vitro* methods such as small RNA sequencing (sRNA-seq) are able to reflect a comprehensive sRNA transcriptome, thus enabling the discovery of novel RNA genes. However, as RNA expression is time, location and condition dependent, the choice of sequenced biological samples is critical for obtaining breadth. sRNA-seq can also have significant background noise from transcript degradation products that have transcript sizes similar to small RNA. Small RNA library preparation is also susceptible to bias, especially during adaptor ligation, leading to over- or underrepresented RNA [29]. On the other hand, *in silico* methods exploit the features of known functional ncRNAs, including conserved secondary structure, compensatory mutations and thermodynamic stability [30]. Thus, *in silico* approaches can detect some RNA genes without the requirement for expression, and searches can provide structural conservation insights. However, *in silico* approaches usually suffer from a high false-positive rates and the accuracy can be significantly affected by parameter selection [27, 30]. The accuracy and efficiency of RNA gene prediction can be improved by integrating multiple approaches both *in vitro* and *in silico* to balance their strengths and weaknesses.

Advances in deep RNA sequencing and bioinformatics algorithms have paved the way for improved structural is-ncRNA characterization. Recently, several *Cryptosporidium* species have had their genomes sequenced, permitting the addition of comparative genomics for structure-based ncRNA discovery [31]. In this study, our objective was to systemantically profile is-ncRNAs in *C. parvum* and explore their conservation among apicomplexan parasites using both *in vitro* (*C. parvum* only) and *in silico* approaches. A small RNA library from excysted *C. parvum* sporozoites was generated and sequenced using the NextSeq Illumina sequencing platform to detect the expressed is-ncRNA transcriptome in this developmental lifecycle stage. Comparative genomics

was used to predict conserved RNA structures and achieve high-quality is-ncRNA candidates. RNA homology was analysed among other *Cryptosporidium* and apicomplexan parasites with covariance models (CMs). New members of the snoRNA family and novel is-ncRNA genes were recovered. Expression was experimentally confirmed by RT-PCR. A small ncRNA derived from tRNA fragments was observed. This study significantly extends our understanding of small RNAs and their evolutionary conservation in *Cryptosporidium*. The CMs that were built can be used to detect is-ncRNA genes in other closely related parasites once their genome sequences become available.

METHODS

sRNA sequencing of *C. parvum* sporozoites

C. parvum IOWA oocysts were obtained from Bunch Grass Farm (BGF) (Deary, ID). A total of 10^8 oocysts were incubated in household bleach (diluted 1:4 in water) on ice for 10 min and then washed twice with cold phosphate-buffered saline (PBS). Excystation of sporozoites was induced by incubating oocysts in 0.8% sodium taurodeoxycholate (Sigma) in PBS at 37 °C for 1 h. Small RNAs (<200 nt) were purified from excysted sporozoites according to the manufacturer's protocols using the Nucleospin miRNA kit (Macherey-Nagel). The NEXTFLEX Small RNA-seq kit v3 was used to prepare a stranded small RNA library and the library was sequenced on the Illumina NextSeq 500 platform with 75 bp single-end reads (SE75). Library preparation and sequencing were conducted at the Georgia Genomics and Bioinformatics Core (GGBC, Athens, GA, USA).

Sequence read preprocessing

The 5′ RNA adapter 5′- TGGAATTCTCGGGTGCCAAGG and the random 4 bp adapter located at the 5′ and 3′ ends of the reads were trimmed using Cutadapt-v2.6 [32]. The lowest acceptable Phred score was set at 30 for sequence quality and the minimum length for a read was set at 15 nt. The filtered and preprocessed reads are referred to as cleaned reads.

Homologous ncRNA predictions in Rfam

The Rfam database is a collection of conserved RNA families [33]. Infernal v1.1.2 [34], which has a CM for each Rfam, was used to predict homologous ncRNA gene families in the *C. parvum* IOWA-ATCC genome sequence assembly [35] downloaded from CryptoDB v46 (https://cryptodb.org/cryptodb/). A CM scores a combination of sequence consensus and RNA secondary structure consensus. The parameters used were: cmscan --cut_ga, --rfam and –nohmmonly. An e-value $<1\times10^{-3}$ was used as the threshold. The results were used as benchmark ncRNA genes in the subsequent analysis.

Transcript assembly from sRNA-seq reads

Following preprocessing, trimmed reads were deduplicated using SAMtools v1.10 [36] and mapped to the *C. parvum* IOWA-ATCC genome sequence assembly downloaded from CryptoDB v46 (https://cryptodb.org/cryptodb/) using Bowtie v1.2.2 [37] with no more than one mismatch allowed, a seed length of 15 and best alignment guaranteed (-v 1, -l 15, --best, --strata). Mapped reads were assembled into transcripts using Blockbuster [38], a tool that clusters reads into blocks of overlapping reads using a Gaussian distribution approach. The minimum distance between two clusters (-distance) was set at 15 bp and the scale standard deviation for the position of each read was replaced by Gaussian profile (-scale) as 0.001. The minimum read number for a block (-minBlockHeight) was set at 2, and for a cluster (-minClusterHeight) was set to 10 (as default). Unmapped reads were analysed by MetaPhlAn2-v2.7.8 [39] to detect potential contamination.

Structural RNA prediction

A comparative genomics-based method was used to detect structurally conserved RNAs in the non-CDS regions of the genome (UTRs, introns, intergenic regions and ncRNAs). First, the genomic sequences of non-CDS regions in the *C. parvum* IOWA-ATCC genome sequence assembly were extracted and searched against seven genome sequences from related species available in CryptoDB v46, including *Cryptosporidium andersoni* 30847, *Cryptosporidium baileyi* TAMU-09Q1, *Cryptosporidium hominis* 30976, *Cryptosporidium meleagridis* UKMEL1, *Cryptosporidium muris* RN66, *Cryptosporidium tyzzeri* UGA55 and *Cryptosporidium ubiquitum* 39726. BLASTN [40] was used with an e-value $<1\times10^{-5}$ and selected the best hit for each query sequence in each target genome. Then, the *C. parvum* genomic regions that had BLASTN hits in more than three species were extracted using BEDTools v2.26.0 [41]. This set of boundary-refined *C. parvum* IOWA genomic sequences were pooled together with the sRNA-seq assembled transcripts that were located in the CDS region on the antisense strand. The pooled sequences were again searched against the seven target genome sequences using the same method to select hits with similarity. CLUSTALW [42] was then used to realign the homologous *Cryptosporidium* spp. sequences and the resulting alignment was fed as input to RNAz v2.1 [30] to detect functional RNA secondary structures. RNAz scanned both strands of the RNA alignment for conserved and stable secondary structures with a window size of 120 bp (as default), a step size of 30 bp and a minimum RNA gene length of 30 bp.

Genome mappability

The tool Mappability [43], which belongs to the GEM-v3.0.3 (GEnome Multitool) program suite, was used to calculate the mappability of each region of the *C. parvum* IOWA-ATCC genome sequence to facilitate the interpretation of the sRNA-seq mapping data. For this analysis, read length (k-mer) was set to 15 bp, which is the shortest read length of the sRNA-seq data in this study, maximum one mismatch allowed, and best mapping guaranteed (as the Bowtie setting used in the study).

Conservation analysis of is-ncRNAs and construction of covariance models

The sequences of is-ncRNA predictions were extracted from the *C. parvum* IOWA-ATCC genome sequence with an additional 10 bp included on each end to provide a better chance of recovering complete RNA structures. To retrieve the best BLASTN hit to serve as a seed sequence alignment, each sequence query was initially used to search target *Cryptosporidium* genome sequences via BLASTN from NCBI BLAST v 2.10.0 with strict criteria: e-value of 1×10^{-5}, similarity >70%, coverage >70%. Then, the seed sequence alignment was realigned by mlocarna from LocARNA v2.0.0RC8 [44] based on both sequence and secondary structure to build a consensus structure. The multiple sequence alignments with a consensus structure were curated by trimming unpaired bases at the boundaries to emphasize the core consensus structure. Then, a CM model was built on the refined multiple sequence alignment for each novel ncRNA candidate using cmbuild from Infernal-v1.1.2 [34]. An e-value $<1\times10^{-3}$ was used as the threshold. The constructed CM compressed bundle of all models (https://doi.org/10.6084/m9.figshare.17056367) was used to search for homologous ncRNAs in more distant apicomplexan species (*Cryptosporidium bovis* isolate 42 482 (GCA_009768925.1), *Cryptosporidium ryanae* isolate 45019 (GCA_009792415.1) and *Cryptosporidium viatorum* isolate UKVIA1 (GCA_004337795.1); *Plasmodium falciparum* 3D7 and *Plasmodium vivax* P01 from PlasmoDB v46 (https://plasmodb.org/plasmo/); and *Toxoplasma gondii* ME49 from ToxoDB v46 (https://toxodb.org/toxo/) that did not have a BLAST-based hit using cmsearch from Infernal-v1.1.2. If a new homologous ncRNA was detected, it was added to the alignment and the above steps were iterated to update the CM model until all genome sequence searches were exhausted and no additional homologs were found.

ncRNA structure prediction and clustering ncRNA candidates were clustered by RNAclust-v1.3 to identify groups of RNA sequences that share similar secondary structure motifs [45]. The output of clustering is a hierarchical tree that was visualized using the Interactive Tree of Life v4 (iTOL) [46]. ncRNA consensus structures were visualized using the RNAalifold web server (http://rna.tbi.univie.ac.at/cgi-bin/RNAWebSuite/RNAalifold.cgi).

snoRNA and other ncRNA prediction

Snoscan-v0.91 [47] and SnoReport-v2.0 [48] were used to predict snoRNA genes using default parameters. Snoscan utilizes a probabilistic model with states related to features such as restriction sites on RNA secondary structure, the score of each characteristic box and the relative position of each box to identify C/D box snoRNAs. rRNA sequences extracted from the annotated *C. parvum* IOWA-ATCC genome sequence were provided to the program as potential rRNA targets for each snoRNA. SnoReport uses a support vector machine (SVM) to classify C/D box and H/ACA box snoRNAs. SnoReport does not use information regarding putative target sites; thus, it can discover orphan snoRNAs. snoRNA structures and the consensus folding pattern were predicted and visualized using RNAfold from the ViennaRNA-2.4.6 package [49] and VARNA-v3.93 [50]. Results from the two methods were checked manually to assess the structures and motifs.

Another sRNA detector, miRDeep2 v.0.0.8 [51], was used to detect both novel and catalogued miRNAs from sRNA-seq data with default parameters. To detect tRNA-derived small RNAs, preprocessed reads (see above) were annotated using MINTmap [52], which has been applied successfully in many tRNA fragment studies [53, 54].

RT-PCR validation of is-ncRNAs

Total RNA from excysted sporozoites was isolated using Trizol Reagent (Invitrogen) according to the manufacturer's protocol. The cDNA was synthesized using the RevertAid First-Strand cDNA Synthesis kit (Thermo Scientific, K1621) according to the manufacturer's recommendations with 1 µg RNA input and a specific primers pool (10 µM). Specific primers for reverse transcription (Table S1), available in the online version of this article) were designed using the PrimerQuest tool from IDT. PCR amplification was performed using an initial denaturation at 95 °C for 1 min, followed by 35 cycles at 95 °C for 5 s and 64 °C for 10 s. The PCR products were electrophoresed on a 2% agarose gel. A 100 bp DNA ladder (Thermo Scientific, SM0241) was electrophoresed along with all amplicons to indicate the DNA size.

RESULTS

sRNA-seq read mapping to the *C. parvum* IOWA-ATCC genome sequence

A small RNA library was sequenced and quality controlled reads were kept. Read lengths ranged from 15 to 68 nt after adapter trimming (Fig. S1). The minimum read length was set as 15 nt because the shortest type of ncRNA, such as miRNA, is 18–25 nt, and 4^{15} is larger than the genome size of *C. parvum* (~9 Mb) and thus should be unique. Reads were then mapped to the most

Table 1. Distribution of Illumina sRNA-seq reads mapped to the *C. parvum* IOWA-ATCC genome sequence

Region	No. in class*	% of genome sequence*	No. of sRNA-seq reads	% of mapped reads
Total raw reads			103 954 523	
Total cleaned reads			71 801 360	
Mapped to genome			27 049 025	
CDSs	5265	76.82	4 158 980	15.38
intronic	1370	1.29	1 446 130	5.35
UTR (5′+3′)	497+370	1.02	197 613	0.73
intergenic	3956	20.68	4 338 561	16.04
rRNA	12	0.12	14 182 503	52.43
Benchmark ncRNAs†				
RNaseP	1	<0.01	1674	0.01
snoRNA	3	<0.01	187 970	0.69
snRNA	5	0.01	152 698	0.56
SRP	1	<0.01	22 884	0.08
tRNA	45	0.04	2 360 012	8.72

*Refers to the *C. parvum* IOWA-ATCC genome sequence.
†Conserved *C. parvum* ncRNA families in Rfam.
SRP, signal recognition particle; RNaseP, ribonuclease P; snoRNA, small nucleolar RNA; snRNA, small nuclear RNA.

complete *C. parvum* sequence, the IOWA-ATCC assembly, and the distribution of reads was analysed (Table 1). Reads that did not map represented *C. parvum* at lower stringency or bacterial contaminants (Table S2). *C. parvum* ncRNAs annotated by Infernal from the Rfam database of known RNA families are considered reliable and conserved, thus they were used as benchmark genes in the rest of the study. Parameters and filters were used as suggested in Ref. [33] that will identify all subsequences that match any Rfam family with a score above the gathering cutoff (GA) selected by the Rfam curators. Fifty-five benchmark genes were identified in *C. parvum*. They consist of 45 tRNAs (Table S3), 5 snRNAs (1 copy for each of U1, U2, U4, U5, U6), 3 C/D box type snoRNAs (U3, SNORD36, SNORD96), 1 copy of signal recognition particle (SRP) and 1 copy of RNaseP. mRNA CDS regions account for 77.84% of the annotated genome sequence, yet only 15.38% of the sRNA-seq reads mapped to CDSs, indicating the high level of RNA integrity and successful small RNA enrichment in the library construction. ncRNAs are revealed by sRNA-seq and comparative genomics analysis

Mapped sRNA-seq reads were assembled into transcripts using Blockbuster. miRNAs were not expected at this step and were investigated separately (see below). A total of 1383 transcripts were assembled from the mapped sRNA-seq reads (Table S4). Of these, 66.2% derived from mRNA exonic regions (857 on the same strand as the mRNA and 58 on the opposite strand), while 3.5% (n=48) were located in intronic regions and 30.4% (n=420) were located in intergenic regions. All of the 55 benchmark gene transcripts were recovered. Most of the transcripts that mapped to mRNA exons are distributed to the margins of expression levels and sequence length, which suggests transcriptional noise such as mRNA degradation (Fig. S2). Benchmark ncRNA genes displayed higher levels of expression (Fig. 1a) and length ranges of ~70–500 nt except for one snRNA U3. The U3 transcript assembly that was generated was fused with the downstream mRNA from gene CPATCC_0004520 due to transcript overlap. It was fixed manually by splitting the fused transcript at the location of lowest read coverage (Fig. S3).

We set the maximum length of transcripts as 600 nt to minimize the transcriptional noise from mRNA degradation products and maximize the capture of ncRNAs. This threshold was chosen for two reasons. First, the primary ncRNA can be longer than the mature transcript. Second, conserved structural RNAs such SRP and RNaseP are in the range of 300–500 nt and were observed to be enriched in sRNA-seq libraries [12]. The minimum length of transcripts was set to 30 nt. ncRNAs shorter than 30 nt are not likely to fold into functional structures. Small ncRNA genes such as miRNA (~22 nt), piRNA (24–32 nt) and siRNA (20–25 nt) function by base-pairing with targets rather than conserved structure. Their detection is based on the particular primary miRNA (pri-miRNA) structure and the pattern of unique mapped reads. Transcripts that overlapped with a CDS from the same strand were filtered out and not considered further.

In total, 341 assemblies remained post-filtering (Fig. 1b). Of these, 67 are located on mRNA exon antisense strands; 40 are located in intronic regions; 61 overlap benchmark genes (55 on the same strand, while 6 are on the antisense strand); and the remaining

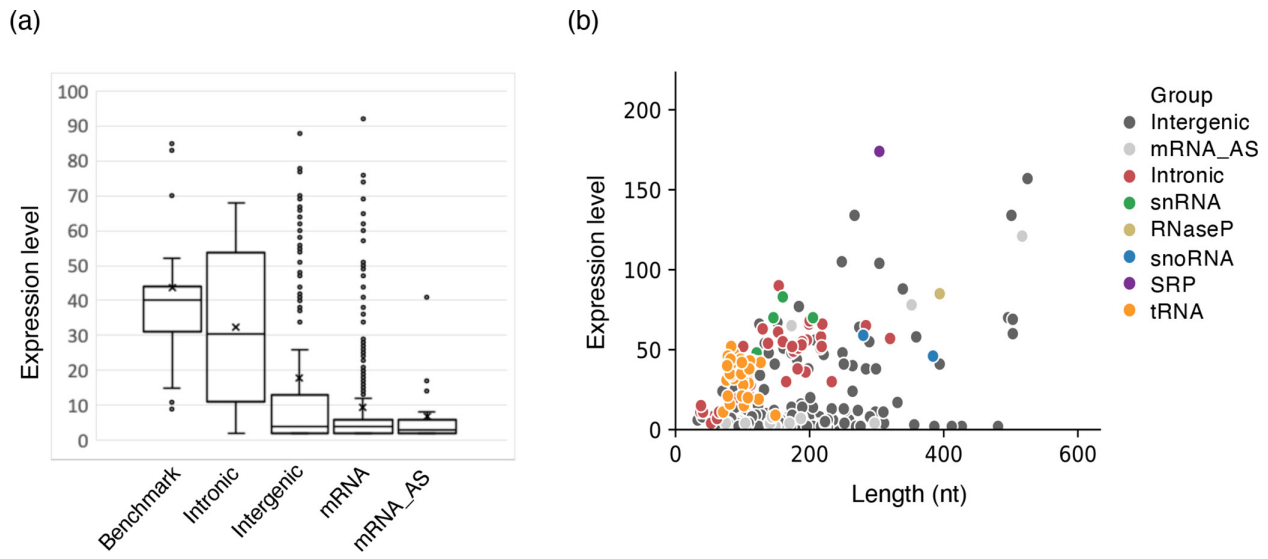

Fig. 1. Analysis of sRNA-seq-generated transcripts. (a) The expression level of 1383 assembled sRNA-seq transcripts sorted by the genomic region to which they map. The mRNA regions were grouped into sense (CDS +UTR) and antisense (mRNA_AS) regions. ×, indicates the average expression level. (b) Scatter plot of sequence length (nt) and expression level of the 341 select (filtered) transcripts. The genomic location of transcripts is indicated by different colours. Expression levels were calculated using Blockbuster based on the number of supporting reads.

173 are in intergenic regions. The transcripts located in intronic, intergenic and mRNA antisense regions without an RNA family assigned are putative novel ncRNA genes.

To ascertain RNA structure information for the 341 predicted ncRNA genes, a second approach was applied using comparative genomics and the program RNAz to screen for conserved structural RNAs among seven *Cryptosporidium* species for which genome sequences are available (Fig. 2). The approach takes advantage of RNA thermodynamic stability and structural conservation, as elucidated from multiple sequence alignments. Thermodynamic stability is quantified by means of a *z*-score that indicates the degree of stability that is greater than would be expected by chance. Structural conservation is quantified by a structural conservation index (SCI), a measure of good consensus structure based on covariant nucleotides capable of maintaining a stem structure. The SCI will be high if individual structures are consistent with the consensus structure of the sequences being analysed. RNAz uses these two features and a support vector machine (SVM) learning algorithm to classify an alignment as capable of 'structural RNA' or 'other'.

Since this study was interested in ncRNAs, multiple sequence alignments for intergenic, intronic and antisense locations (5222 regions) were built according to the *C. parvum* IOWA-ATCC annotation, regardless of whether or not a transcript mapped to this region (Fig. 2). A relaxed criterion of $P>0.5$ was selected in order to consider an element as analysed by RNAz as a potential structural RNA, since the putative ncRNA transcripts would be further characterized with multiple criteria in downstream analyses. RNAz uses a machine learning technique to calculate the *P*-value using training data from Rfam. According to RNAz documentation, the false-positive rate at this cutoff was found to be ≈4% [30].

RNAz analysis generated 435 structural RNA candidates, 346 (79.7%) located in intergenic regions, 15 (3.5%) in currently annotated UTR regions, 14 (3.2%) in intronic regions, 6 (1.4%) in CDS antisense regions and 53 (12.2%) in known RNA genes [including 47 (88.7%) corresponding to benchmark genes and 6 (11.3%) rRNAs] (Table S5). These sequences summed up to 146881 bp or 6.9% of the total length of the input sequence derived from the 5222 non-CDS regions.

When the results of the 2 approaches were combined, 107 ncRNAs were predicted by both methods. Additionally, 37 ncRNAs from the sRNA-seq analysis that had higher expression levels than the third quartile of the benchmark genes were also retained for further analysis. Since RNA structure prediction methods are usually insenstiitve to RNA strand, candidates that overlapped with themselves on the opposite strand were further examined and the less expressed transcript was removed ($n=6$). Four potential transcripts were also removed on the basis of read-through, here defined as being within 15 bp of an annotated mRNA on the same strand. In the end, a total of 134 ncRNA sequences were kept as high-quality RNA gene candidates and used for additional evolutionary studies (Table S6).

The 134 high-quality RNA genes were further verified by the uniqueness of the sequence in the genome. False-positive predictions could be derived from expression data that derived from non-uniquely mapped reads, which are common in sRNA-seq

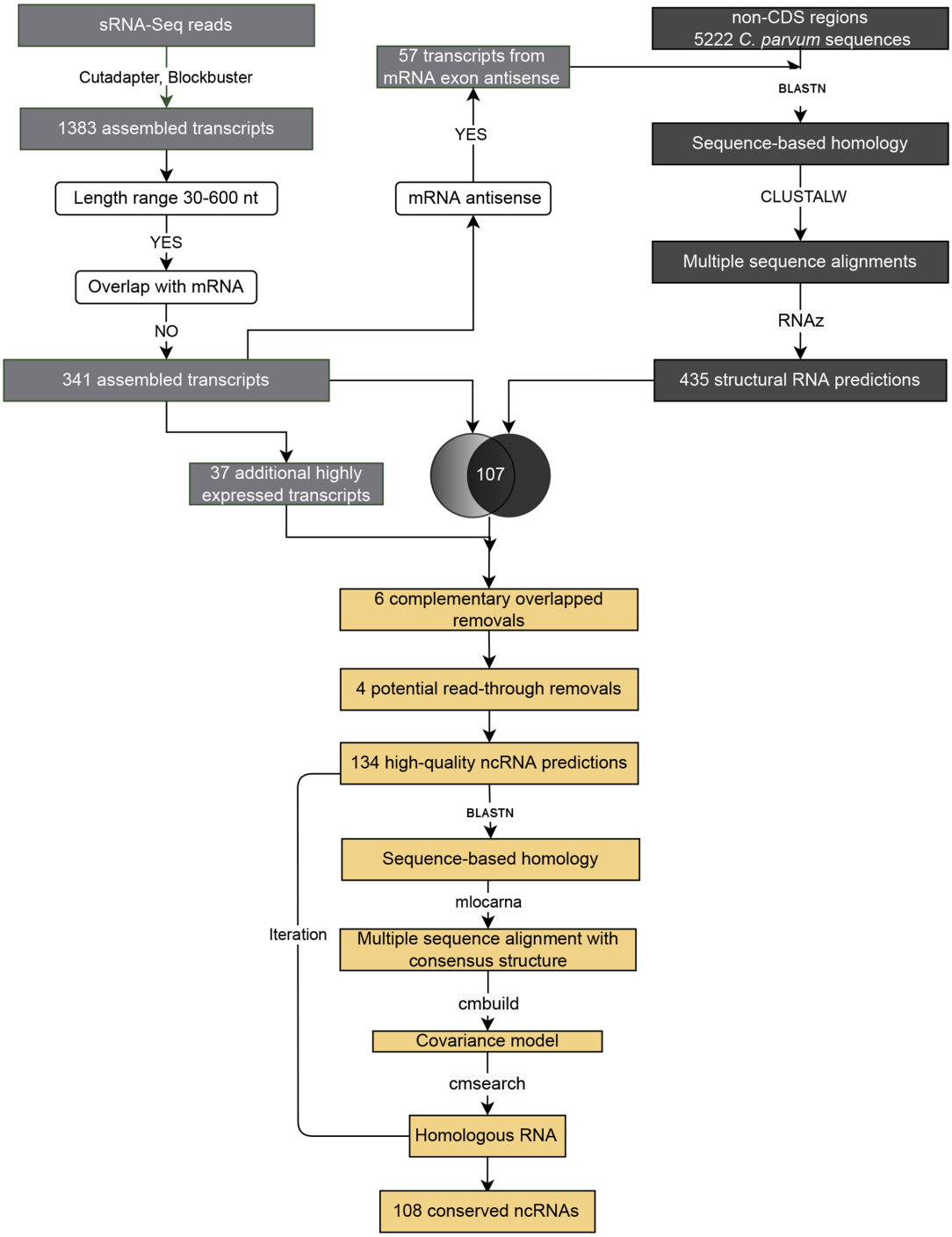

Fig. 2. Pipeline used to ascertain conserved small RNAs with high expression or conserved ability to form secondary structures.

data due to relatively short read length, short gene length and multiple gene family members. Thus, GEM Mappability was used. It is a mapping-based tool that shreds the genome sequence into k-mers and examines the level of mapping uniqueness of the sequencing data. The average mappability score was ~1, which indicated that the 134 RNA genes are good for downstream analysis.

ncRNA homology search and covariance model construction for ncRNA homology detection is challenging due to poor primary sequence conservation. Methods developed by the Rfam database were used to identify 'families' of homologous ncRNAs, which include powerful statistical models of sequence and structure profiles known as CMs. CMs are a widely used and robust algorithm for RNA homology searches. The approach is closely related to the profile hidden Markov model. CMs capture sequence and

structure information from a structurally annotated RNA multiple sequence alignment. This approach provides reliable prediction accuracy and sensitivity [55].

The strategy used was: (1) use RNA sequence information to detect homologous sequences in closely related species to construct good 'seed' alignments that capture both sequence conservation and variation; (2) embed secondary structure information to improve alignment accuracy and build the CM; and (3) iteratively use the CM model to detect RNA homologues and update the alignments.

First, the 134 high-quality ncRNA candidates were clustered based on their structural similarity (Fig. 3a). tRNAs formed the largest group, while other benchmark genes distributed separately. CM model searches of each ncRNA against the genome sequence of *C. parvum* resulted in 3 to 35 hits for each tRNA, while 87/89 remaining ncRNAs (98%) resulted in only 1 hit in the genome. Two novel intergenic snRNAs (Cp_snc_8 and Cp_snc_116) each had two hits. This is consistent with the reduced gene family in *C. parvum*, where benchmark ncRNAs only have one copy for each family except for tRNAs.

Second, for each of the ncRNA candidates in *C. parvum*, homology searches against 10 *Cryptosporidium* species and 3 apicomplexans, including *T. gondii* ME49, *P. falciparum* 3D7 and *P. vivax* P01 were performed. The *Cryptosporidium* species have different host preferences and cover a wide phylogenetic range [56]. They include five species that are closer to *C. parvum* (*C. tyzzeri*, *C. hominis*, *C. meleagridis*, *C. viatorum* and *C. ubiquitum*) and five that are evolutionarily distant from *C. parvum* (*C. baileyi*, *C. ryanae*, *C. bovis*, *C. muris* and *C. andersoni*), based on the genome similarity and synteny (Fig. 3a). The distribution of orthologues has three peaks: 53 is-ncRNAs conserved in all 13 species that are all known RNA families in Rfam; 28 conserved in 10 species that are ncRNAs conserved in all 10 *Cryptosporidium* species; and 15 only conserved in the 5 closely related *Cryptosporidium* species (Fig. 3b). Overall, benchmark ncRNAs are likely to be conserved in the Apicomplexa, while novel ncRNAs are conserved in the genus *Cryptosporidium*. The length of the ncRNAs ranges from 45 to 531 nt, primarily from 70 to 200 nt (Fig. 3c). Four genes longer than 500 bp are novel intergenic lncRNAs without known function.

Overall, 23 ncRNAs encoded in intronic regions were discovered, including 4 snoRNAs and 19 novel ncRNAs (Fig. 3a). Given that *Cryptosporidium* has a much lower number of introns than other eukaryotes and most other apicomplexans [18], these small ncRNAs might be exerting an evolutionary selective pressure to keep some of the introns in mRNAs.

The new members of known ncRNA families are emerging. snoRNAs are a class of essential ncRNAs that primarily guide chemical modifications of other RNAs, including rRNAs, tRNAs and snRNAs. Some snoRNAs lack this sequence complementarity to other RNAs, so-called orphan snoRNAs. There are two main classes of snoRNAs, the C/D box snoRNAs and the H/ACA box snoRNAs. C/D box snoRNAs are ~60–100 nt long and associated with RNA methylation. They are classified by the characteristic elements called boxes – C box (RUGAUGA, where R is a purine) and D box (CUGA) – near their 5′ and 3′ ends. H/ACA snoRNAs are ~120–160 nt long and associated with RNA pseudouridylation. They typically form a double hairpin loop structure containing an H box (ANANNA) located between the two hairpins loops and an ACA box (ACA) following at the 3′ end.

Three C/D box snoRNAs of *C. parvum* are reported in CryptoDB and Rfam, but no H/ACA snoRNAs are listed. The annotation of snoRNAs in *Cryptosporidium* is incomplete due to the extensive sequence and structural variation present in this family. Thus, a sub-screen of the 134 ncRNA candidates using two snoRNA specified tools, snoReport and snoScan, was performed to identify putative H/ACA box and C/D box snoRNA genes. With manual examination of their motifs and structures, three additional C/D box and four H/ACA box snoRNAs could be annotated (Fig. 4). Of these, one C/D box and two H/ACA box snoRNAs are conserved among all *Cryptosporidium* species searched in this study. The others are also conserved in at least seven species. Of the 10 snoRNAs genes, 4 are intronic, 3 are intergenic, 1 is antisense (H/ACA box snoRNAs) and 2 are located in UTRs.

In summary, homology searches using the 134 high-quality ncRNA predictions in 13 apicomplexan parasites revealed 55 known RNA clades in Rfam and 79 novel ncRNA genes (7 new snoRNAs and 72 ncRNAs without known function) (Table 2). Of these high-quality ncRNA predictions, 104 have homology in at least seven species and are considered structurally conserved ncRNAs (Table S6).

Absence of miRNAs and discovery of tRNA-derived small RNAs

We used miRDeep2 to detect potential miRNAs in *C. parvum* using the sRNA-seq data. As expected, no known miRBase miRNAs were found. A few novel miRNAs were predicted by miRDeep2, but they either had low read support or low prediction scores. Thus, it is likely that no canonical miRNA structures exist in *C. parvum*, consistent with the lack of RNAi machinery. However, upon investigation, some of these 'novel miRNA' predictions are derived from tRNA genes, raising the possibility that the RNA fragments resulted from tRNA degradation or are enzymatically cleaved biological small RNAs.

To test this hypothesis, sRNA-seq reads were mapped to *C. parvum* tRNAs and the results were analysed using the tool MINTmap [52], which is designed for profiling tRNA fragments from sRNA-seq (Table S7). A total of 7717 tRNA fragments (with different mapping positions on tRNAs) were identified, most of which had low read counts and were filtered out. The length of the remaining 93 tRNA fragments from 25 tRNA genes were plotted with reads per million mapped reads (RPM) >20 (Fig. 5a) (Table

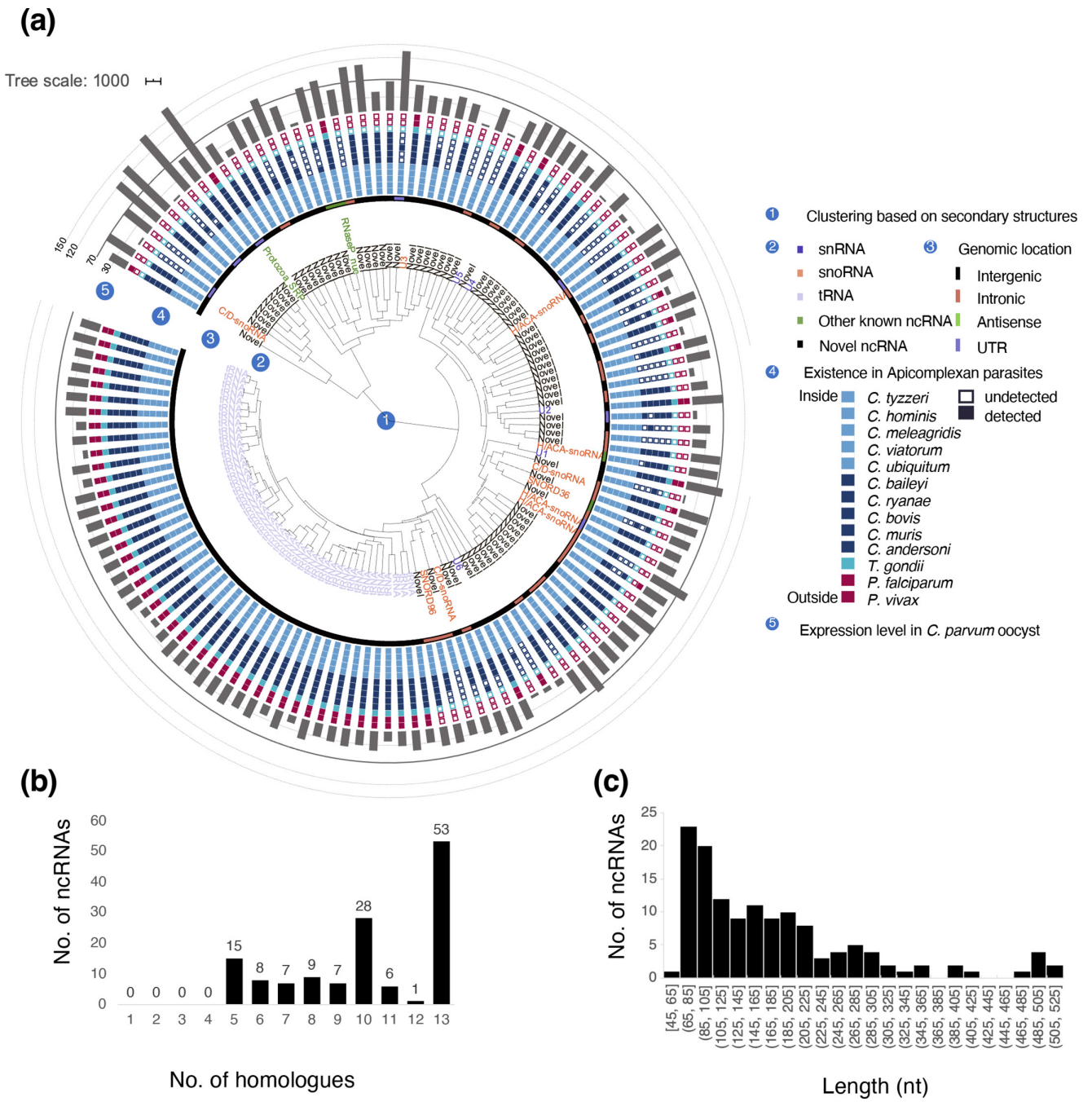

Fig. 3. Conservation of predicted ncRNA genes. (a) Conservation and expression of the 134 ncRNA predictions. RNA families are coloured differently. C/D-snoRNA and H/ACA-snoRNA are new snoRNA members predicted in this study. Circles one to five (inner to outer) indicate: 1, structural similarity-based clustering; 2, the type of small RNA; 3, the genomic location of the gene; 4, the extent of homology in other apicomplexan parasites; and 5, the expression level (based on the number of supporting reads detected. In circle four, the five *Cryptosporidium* species that are evolutionarily close to *C. parvum* are indicated by a light blue colour, while distant species [56] are in dark blue. (b) Number of homologues detected for each ncRNA candidate. (c) Transcript length distribution of predicted ncRNA genes in 10 nt sliding windows.

S3). Two fragment lengths are enriched, 16–23 nt and 30–38 nt. The tRNA fragment boundaries are located at the 5′ start, the D loop, the anticodon loop and the 3′ end of the tRNA (Fig. 5b). The results demonstrate that the tRNA fragments are not random RNA degradation but processed, raising the possibility that they may be functional small RNAs.

There are 45 *C. parvum* tRNA genes currently annotated and 8 are unique, i.e. there is only 1 tRNA for the amino acid (Asn, Asp, Cys, His, Met, Phe, Trp, Tyr) (Table S3). The single-copy $tRNA^{Asp}$ is the most highly expressed tRNA in oocysts (Fig. 5c), followed

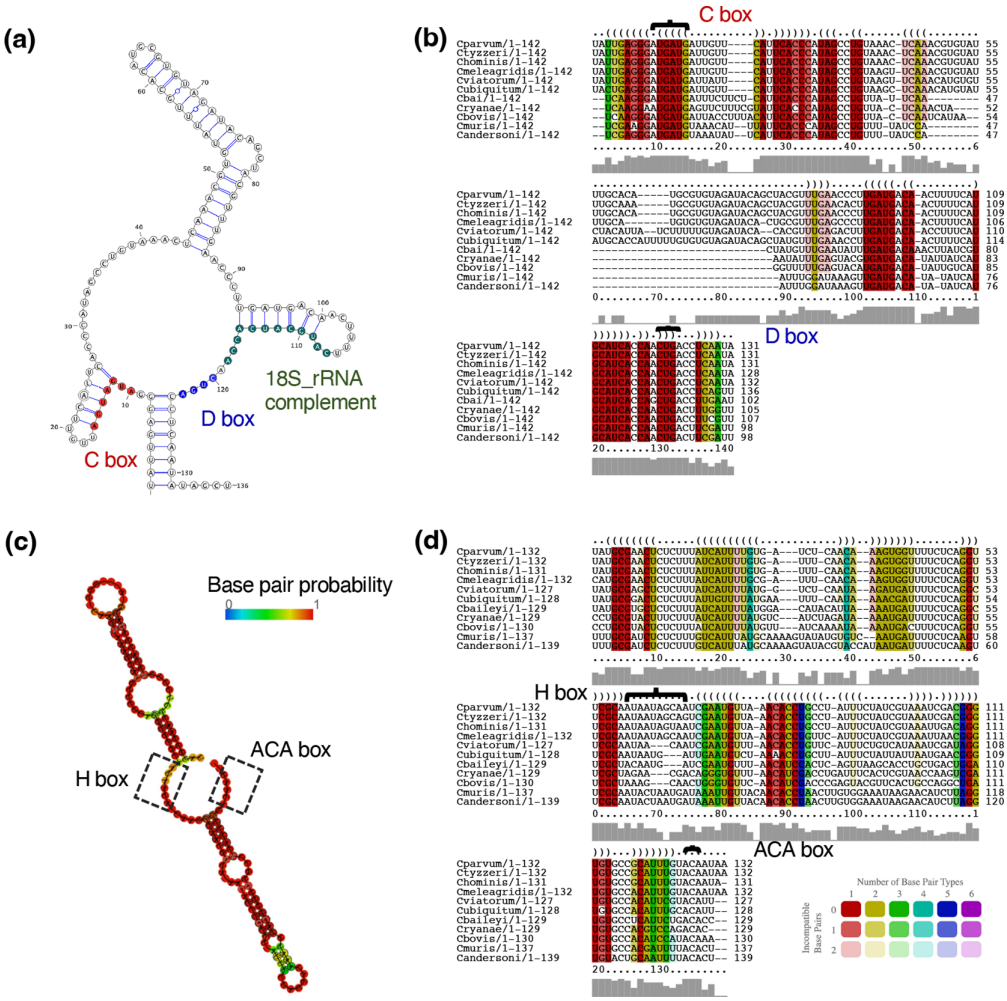

Fig. 4. Representatives of predicted novel snoRNAs. (a) C/D box snoRNA structure and (b) sequence alignment with consensus structure of Cp_snc_103. (c) H/ACA box snoRNA structure and (d) sequence alignment with consensus structure of Cp_snc_108.

by tRNA^{Ile} and tRNA^{Gly}. tRNA^{Asp} also has the most abundant tRNA fragments, ranging in size from 30 to 40 nt. Over 86% of the tRNA^{Asp}-derived fragments are the 3′-tRNA halves (Fig. 5d). On the other hand, the fragments of tRNA^{Ile} were enriched in the region from the T loop to the 5′ ends, ranging in size from 19 to 20 nt. The CCA sequences at the 3′ end of mature tRNA were not detected in the characterized tRNA fragments.

Experimental confirmation of the novel is-ncRNAs

For expression validation by RT-PCR, 24 is-ncRNA predictions were randomly selected for testing. All tested is-ncRNA candidates had a fragment of the expected size amplified (Fig. 6).

DISCUSSION

We systematically characterized small to intermediate size ncRNAs in *Cryptosporidium parvum* by both experimental and *in silico* methods. A small RNA library from *C. parvum* IOWA BGF sporozoites was sequenced using Illumina. Structural RNA elements were analysed using a comparative genomics approach. A total of 134 new high-quality is-ncRNA genes were annotated. CMs were also developed and applied to search for ncRNA homologues in 10 additional *Cryptosporidium* species and 3 other apicomplexan parasites. The CM built in this study can be used to detect ncRNA genes in other closely related parasites. Of the 134 new is-ncRNAs, 104 are conserved among *Cryptosporidium* species. They include 55 known ncRNAs described in Rfam, 7 new members of snoRNA (3 C/D box and 4 H/ACA box snoRNAs) and 42 novel ncRNAs of unknown function. A search for miRNAs in our sporozoite RNA-seq expression data found none, which is consistent with the previous assumption that no canonical miRNAs exist in *C. parvum*, since key components of the RNAi machinery are missing, including argonaute and dicer

Table 2. Classification of is-ncRNAs predicted in this study

	Group	Type	Count
Known is-ncRNAs	snRNA	U1	1
		U2	1
		U4	1
		U5	1
		U6	1
	snoRNA	U3	1
		SNORD36	1
		SNORD96	1
	SRP		1
	RNaseP		1
	tRNA		45
Novel is-ncRNA	snoRNA	C/D box	3
		H/ACA box	4
	Antisense		5
	Intergenic		43
	Intronic		23
	UTR		1

[18]. However, a recent bioinformatics study identified a single putative miRNA encoded in *C. parvum* sporozoite ESTs [25]. To address this finding, the genetic locus was examined further using the sporozoite sRNA-seq data and available RNA-Seq data. This locus encodes a SANT/Myb domain/WD40YVTN repeat-containing protein (CPATCC_0005100). The region around the putative miRNA is not capable of forming a canonically base-paired stem–loop structure. No miRNA was detected in our expression or computational analyses using miRDeep2.

Seventy *C. parvum* (strain IOWA II) unspecified RNAs with length <500 nt are annotated in CryptoDB v46. After filtering for hits that matched our new annotated mRNA genes and rRNAs (strain IOWA-ATCC), 60 were left as potential ncRNAs genes. Ten of these ncRNAs are not in our annotation. Among the 10, 7 were expressed in the sRNA-seq and assembled into transcripts but not considered structural ncRNAs (5 without RNAz structure predictions, 2 removed as potential mRNA read-through noise). The remaining three ncRNAs from CryptoDB v46 were not detected in sRNA-seq or structural RNA predictions. The remaining 50 ncRNA genes in CryptoDB were annotated as conserved ncRNA genes in this study. In conclusion, sRNA-seq from this study recovered most of the known ncRNAs available in CryptoDB (57 out of 60) and Rfam (55 out of 55). While the RNA gene prediction in this study was limited to one sporozoite small RNA library, the observation of most of the known ncRNAs available in CryptoDB indicates that our sRNA-seq coverage was deep enough to detect most of the small to intermediate size ncRNA gene expression and structural RNAs in *C. parvum* oocysts/sporozoites. Even though there are not any sRNA-seq data from other developmental stages (they are difficult to obtain due to severe host contamination), candidates that are related to housekeeping functions should be detectable in sporozoites. The deep coverage obtained also compensates, in part, for the lack of sRNA-seq replicates in this study, but some ncRNAs may have been missed. Most of the transcriptional noise such as mRNA degradation can be distinguished from biological ncRNA genes based on their lower expression levels and lack of conserved structures.

The benchmark genes are the most conserved RNAs within *Cryptosporidium* and also have detectable homologues in *Toxoplasma* and *Plasmodium*. The phylogenetic distance within *Cryptosporidium* is considerable [56]. Of the 104 candidates that are conserved across >7 of the 10 *Cryptosporidium* species examined, 42 are novel with unknown function. It is worth noting that most *Cryptosporidium* genome sequences have many sequencing gaps. Homologues in other species may be missing due to fragmented genome sequences.

Candidates that show a narrow phylogenetic distribution (homology only detected in closely related *Cryptosporidium* species or *C. parvum*-specific) are not considered conserved is-ncRNAs in this study. It is possible that they are derived from transcriptional noise or are false positives by RNAz with no biological function. However, some candidates may have species-specific functions, especially those with good expression levels and significant secondary structures as predicted by RNAz. To distinguish them from

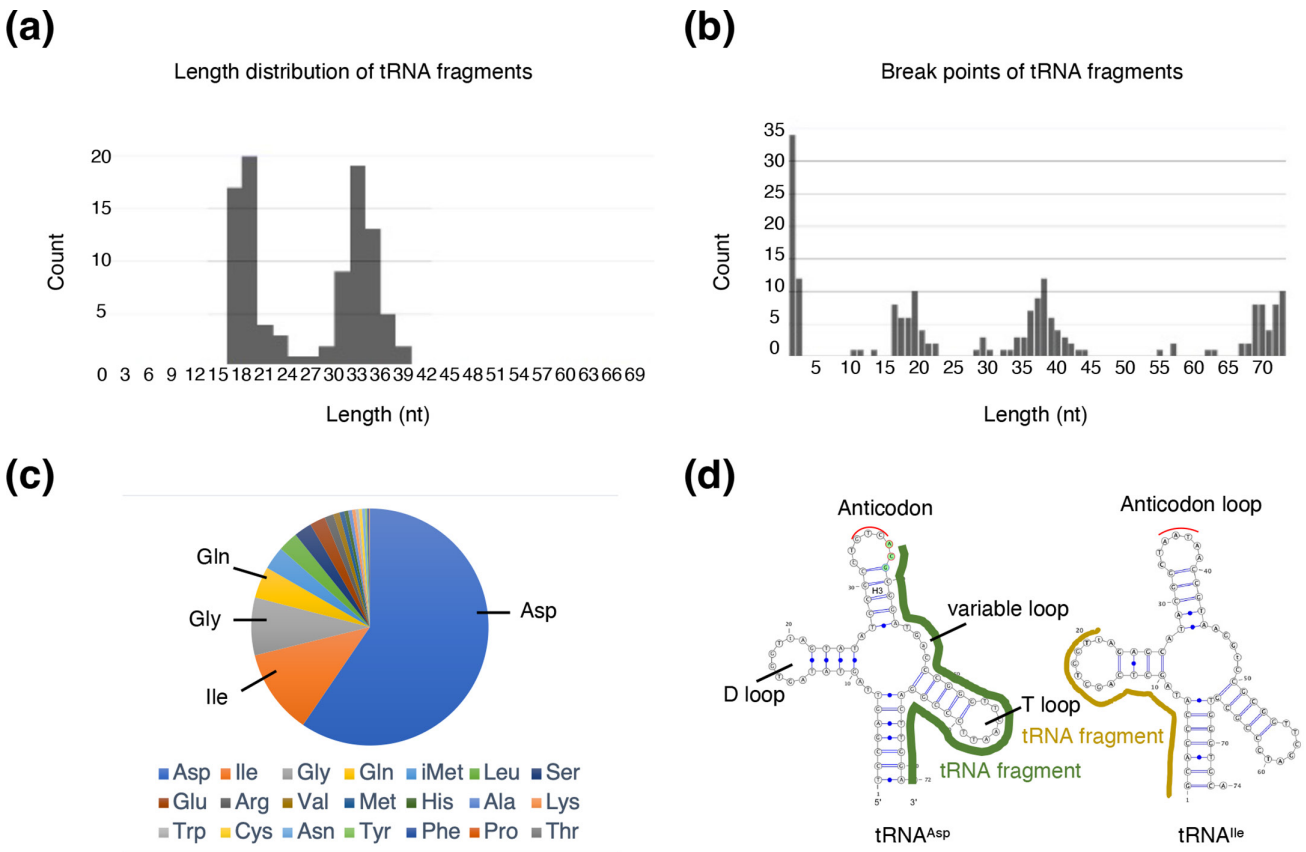

Fig. 5. Analysis of tRNA fragments in sRNA-seq reads. (a) Length distribution of tRNA fragments. (b) Position of breakpoints of tRNA fragments. (c) tRNA read distribution by amino acid. (d) Cleavage sites of tRNA fragments in mature tRNA, using the secondary structure of $tRNA^{Asp}$ and $tRNA^{Ile}$ as examples.

transcriptional noise, additional expression evidence, such as sRNA-seq with replicates or sRNA-seq from other *Cryptosporidium* species, will be needed.

The conserved ncRNA candidates are restricted to RNAs whose spatial structure is of functional importance. Out of the top 100 highly expressed transcripts in sRNA-seq (excluding candidates from CDS regions on the same strand), 31 (31%) are not predicted as structural RNAs by RNAz, including 2 of the annotated ncRNAs (cgd2_1365 and cgd5_2965), both located in intronic regions. It is possible that these ncRNAs do not have a stable and conserved spatial structure or serve as a source for other small RNAs. In humans, intronic RNAs can be further processed into smaller RNAs such as miRNAs or circRNA [56, 57]. Another possibility

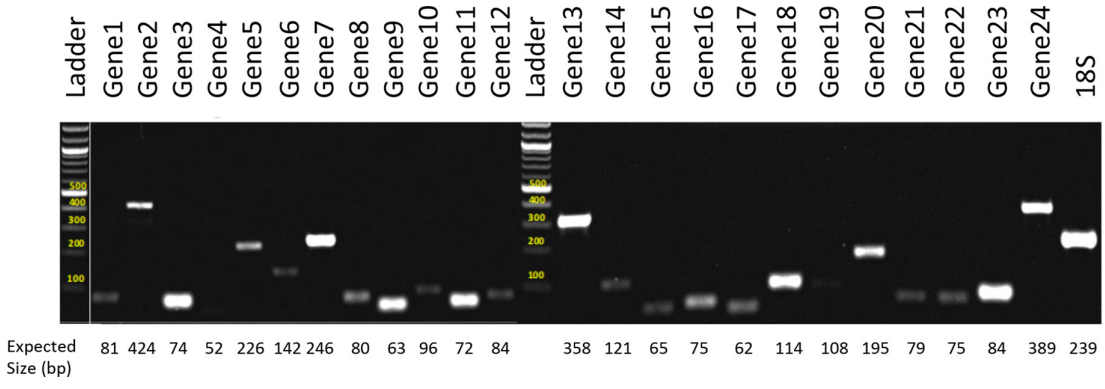

Fig. 6. Experimental validation of predicted novel is-ncRNAs via RT-PCR. The expected size of each amplified fragment is shown below the lane. sncRNA IDs corresponding to genes 1–24 are in Table S1. A 100 bp ladder was used.

is that the genomic locations are more critical to their functions than structures. In *P. falciparum*, two classes of ncRNA with lengths of 75 and 175 nt are transcribed from centromere regions with a suggested role in centromeric chromatin maintenance. It is also worth considering the limitations of *in silico* structure predictions that employ free energy minimization. The approach is not fully parameterized by experimental data. Currently, it is not possible to take in all variables that affect RNA structures, including, but not limited to, Mg2+level, heavy metals, pH and temperature [58]. Furthermore, extensive evidence argues that some RNAs fold into functional structures that are not at the minimum free energy and exist in dynamic states rather than a single explicit structure [59]. Thus, the lack of a conserved structure for some candidates does not necessarily mean that they lack biological function. Interpretation of conserved RNA structures requires caution.

Of the 435 structural RNAs predicted by RNAz, only 179 (41.1%) have sRNA-seq expression support. False positives are not the only considerations. is-ncRNA expression may only occur in certain developmental stages or conditions. Also, RNAz cannot distinguish functional ncRNA genes from cis-regulatory elements, which are part of mRNAs. To be clear, only structural is-ncRNA genes are targeted in this study. When screening for structural elements, coding regions were excluded that could encode structural information. In some organisms, CDS regions harbour more structure than UTRs [60–62]. An RNA structurome should include both coding and noncoding RNAs, which can be revealed by genome-wide RNA structure probing methods such as structure-seq and SHAPE-Map.

Excepting tRNAs, most of the *C. parvum* is-ncRNA genes annotated in this study are single copy. In invertebrates, sequences encoding H/ACA and C/D box snoRNAs are typically located in introns of their host gene in the same orientation [63], while in plants they are transcribed as polycistronic transcripts [64]. snoRNAs located in UTR regions are observed in *P. falciparum* and *C. parvum*, a finding that has only been reported in a few other organisms. Like other apicomplexan parasites, *C. parvum* has a compact genome of ~9.2 M bp in which 76.88% of the genome is protein-coding region. Many metabolic pathways appear broken, genes are streamlined and only about a quarter of genes contain intron structures. These reductive genomic features may be relevant to shaping ncRNA function and evolution in *Cryptosporidium*. Two uncharacterized protein-encoding genes (CPATCC_0009230 and CPATCC_0032370) in which>80% of the gene body is covered by is-ncRNAs on the antisense strand are observed. An H/ACA box snoRNA and a novel ncRNA are nested in CPATCC_0009230, which encodes a hypothetical protein. The benchmark is-ncRNA gene signal recognition particle RNA (SRP) is nested in CPATCC_0032370, also encoding a hypothetical protein. This raises the possibility of the co-evolution of ncRNAs with their target mRNA genes. One possible *in silico* way to test this hypothesis is to analyse selection pressures on the sequence and secondary structure features independently. Bialignments as models of evolution of RNA sequence and structure are in the exploratory stage of development [65].

tRNA-derived small RNA (tsRNA) have been identified in several protozoans, including *Giardia lamblia* [66], *Tetrahymena thermophila* [67], *Trypanosoma brucei* [68] and *Plasmodium falciparum* [69]. In humans, they have been shown to be involved in cellular proliferation and the RNAi pathway, and to regulate gene translation, similar to miRNAs [70]. tRNA-derived fragments can be classified into two main groups: tiRNAs (or tRNA halves) and tRFs (tRNA-derived fragments). tRNA halves (tiRNAs) are produced by specific cleavage in the anticodon loop of a mature tRNA by ribonucleases to produce 30–40 nt 5′-tRNA and 3′-tRNA halves. Studies indicate that tiRNA cleavage rarely occurs under normal conditions but happens under various stress conditions [71]. tRFs are 18–22 nt, and are derived from mature tRNAs after cleavage at the D-loop, T-loop or anticodon stem. In this study, the $tRNA^{Asp}$ fragments were primarily 3′-tiRNAs, while $tRNA^{Ile}$ was observed to produce 5′-tRFs. 5′-tRFs are the most abundant type in *P. falciparum*, with the majority (90%) of tsRNA derived from tRNAs that encode eight amino acids: Pro, Phe, Asn, Gly, Cys, Gln, His and Ala [69]. In *G. lamblia*, the most abundant type of tsRNA varied among life cycle stages, and is thought to be involved in parasite differentiation [66]. In *T. brucei*, $tRNA^{Thr}$ 3′-tiRNAs were reported to be produced during nutrient deprivation, where they become one of the most abundant tRNA-derived RNA fragments [68]. These results suggest that tsRNA is likely to have different origins and functions in different parasites and environmental conditions. Although tsRNAs have been related to DICER and AGO in many studies, the biogenesis and functional mechanisms of tsRNA are not understood clearly. Dicer-independent post-transcriptional gene expression regulation by tsRNA has also been reported [72]. Thus, tsRNAs harbour similarity to miRNAs but may use an alternative pathway to achieve RNAi. Additionally, tsRNAs have been reported to be a common species in extracelluar vasicles (EVs). In *Leishmania*, tsRNAs, mostly derived from $tsRNA^{Asp}$ and $tsRNA^{Gln}$ 5′ end or mid-5′ end, were seen to be selectively and specifically enriched in secreted exosomes, raising the possibility of tsRNA involvement in host–parasite interactions [73]. The role of tsRNAs in *Cryptosporidium* needs further exploration.

The apicomplexan RNAi pathway is significantly different from that of other eukaryotes. Like *Cryptosporidium*, *Plasmodium* also lacks key components in the RNAi pathway detectable endogenous miRNAs. However, in one study, a host miRNA–Argonaute 2 complex was detected in EVs and was shown to regulate gene expression in *P. falciparum* [74]. These findings raised the intriguing possibility that the parasite may be able to utilize host Argonaute 2 [69]. Some *Cryptosporidium* lncRNAs have been reported to be translocated into the host cell nucleus [75]. Whether is-ncRNA also plays an important role in gene regulation or host–parasite interactions merits further exploration.

There are likely many more conserved ncRNAs in *Cryptosporidium*. *Cryptosporidium* studies are challenging and are limited by some factors: (i) the lack of ncRNA from different time point data due to the lack of a continuous *in vitro* culture system spanning the entire

life cycle; (ii) a lack of complete genome sequences for other species. The fragmented nature of the available genome sequences from other species could affect the ncRNA detection by homology and copy number.

Here, a vesicles annotation of small to intermediate size ncRNAs in *Cryptosporidium* is provided, extending our understanding of the noncoding transcriptome and its conservation in *Cryptosporidium* evolution. sRNA-seq, together with a comparative genomics approach, revealed many novel structural ncRNAs, providing an opening to their study. Given the absence of a conventional RNAi pathway, ncRNAs may provide insights into unique mechanisms of gene regulation in *Cryptosporidium* parasites.

Funding information
The current project was supported by NIH 1R21AI144779.

Author contributions
Y.L., R.P.B., J.C.K.: conceptualization, methodology, original draft preparation. Y.L., R.P.B.: data curation. Y.L., X.M.: formal analysis. Y.L.: visualization. J.C.K.: supervision. J.C.K., R.P.B., J.C.K.: funding, writing – review and editing.

Conflicts of interest
The authors declare that there are no conflicts of interest.

References

1. Cech TR, Steitz JA. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell* 2014;157:77–94.
2. Losko M, Kotlinowski J, Jura J. Long noncoding RNAs in metabolic syndrome related disorders. *Mediators Inflamm* 2016;2016:5365209.
3. Matrajt M. Non-coding RNA in apicomplexan parasites. *Mol Biochem Parasitol* 2010;174:1–7.
4. Fabian MR, Sonenberg N, Filipowicz W. Regulation of mRNA translation and stability by microRNAs. *Annu Rev Biochem* 2010;79:351–379.
5. Anderson P, Ivanov P. tRNA fragments in human health and disease. *FEBS Lett* 2014;588:4297–4304.
6. Falaleeva M, Stamm S. Processing of snoRNAs as a new source of regulatory non-coding RNAs: snoRNA fragments form a new class of functional RNAs. *Bioessays* 2013;35:46–54.
7. Zhang X, Wang W, Zhu W, Dong J, Cheng Y, *et al*. Mechanisms and functions of long non-coding RNAs at multiple regulatory levels. *Int J Mol Sci* 2019;20:22.
8. Li Y, Baptista RP, Kissinger JC. Noncoding RNAs in apicomplexan parasites: an update. *Trends Parasitol* 2020;36:835–849.
9. Statello L, Guo CJ, Chen LL, Huarte M. Gene regulation by long non-coding RNAs and its biological functions. *Nat Rev Mol Cell Biol* 2021;22:96–118.
10. Li Y, Baptista RP, Sateriale A, Striepen B, Kissinger JC. Analysis of long non-coding RNA in *Cryptosporidium parvum* reveals significant stage-specific antisense transcription. *Front Cell Infect Microbiol* 2020;10:608298.
11. Wang Y, Chen J, Wei G, He H, Zhu X, *et al*. The *Caenorhabditis elegans* intermediate-size transcriptome shows high degree of stage-specific expression. *Nucleic Acids Res* 2011;39:5203–5214.
12. Wei C, Xiao T, Zhang P, Wang Z, Chen X, *et al*. Deep profiling of the novel intermediate-size noncoding RNAs in intraerythrocytic *Plasmodium falciparum*. *PLoS One* 2014;9:e92946.
13. Yan D, He D, He S, Chen X, Fan Z, *et al*. Identification and analysis of intermediate size noncoding RNAs in the human fetal brain. *PLoS One* 2011;6:e21652.
14. St Laurent G 3rd, Wahlestedt C. Noncoding RNAs: couplers of analog and digital information in nervous system function? *Trends Neurosci* 2007;30:612–621.
15. Bouzid M, Hunter PR, Chalmers RM, Tyler KM. *Cryptosporidium* pathogenicity and virulence. *Clin Microbiol Rev* 2013;26:115–134.
16. Checkley W, White AC, Jaganath D, Arrowood MJ, Chalmers RM, *et al*. A review of the global burden, novel diagnostics, therapeutics, and vaccine targets for cryptosporidium. *Lancet Infect Dis* 2015;15:85–94.
17. Pumipuntu N, Piratae S. Cryptosporidiosis: A zoonotic disease concern. *Vet World* 2018;11:681–686.
18. Abrahamsen MS, Templeton TJ, Enomoto S, Abrahante JE, Zhu G, *et al*. Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science* 2004;304:441–445.
19. Pawlowic MC, Somepalli M, Sateriale A, Herbert GT, Gibson AR, *et al*. Genetic ablation of purine salvage in *Cryptosporidium parvum* reveals nucleotide uptake from the host cell. *Proc Natl Acad Sci USA* 2019;116:21160–21165.
20. Mauzy MJ, Enomoto S, Lancto CA, Abrahamsen MS, Rutherford MS. The *Cryptosporidium parvum* transcriptome during in vitro development. *PLoS One* 2012;7:e31715.
21. Mirhashemi ME, Noubary F, Chapman-Bonofiglio S, Tzipori S, Huggins GS, *et al*. Transcriptome analysis of pig intestinal cell monolayers infected with *Cryptosporidium parvum* asexual stages. *Parasit Vectors* 2018;11:176.
22. Tandel J, English ED, Sateriale A, Gullicksrud JA, Beiting DP, *et al*. Life cycle progression and sexual development of the apicomplexan parasite *Cryptosporidium parvum*. *Nat Microbiol* 2019;4:2226–2236.
23. Vinayak S, Pawlowic MC, Sateriale A, Brooks CF, Studstill CJ, *et al*. Genetic modification of the diarrhoeal pathogen *Cryptosporidium parvum*. *Nature* 2015;523:477–480.
24. Keeling PJ. Reduction and compaction in the genome of the apicomplexan parasite *Cryptosporidium parvum*. *Dev Cell* 2004;6:614–616.
25. Ahsan MI, Chowdhury MSR, Das M, Akter S, Roy S, *et al*. *In silico* identification and functional characterization of conserved miRNAs in the genome of *Cryptosporidium parvum*. *Bioinform Biol Insights* 2021;15:11779322211027665.
26. Nawaz MZ, Jian H, He Y, Xiong L, Xiao X, *et al*. Genome-wide detection of small regulatory RNAs in deep-sea bacterium *Shewanella piezotolerans* WP3. *Front Microbiol* 2017;8:1093.
27. Washietl S, Hofacker IL, Lukasser M, Hüttenhofer A, Stadler PF. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat Biotechnol* 2005;23:1383–1390.
28. Ferrero G, Cordero F, Tarallo S, Arigoni M, Riccardo F, *et al*. Small non-coding RNA profiling in human biofluids and surrogate tissues from healthy individuals: description of the diverse and most represented species. *Oncotarget* 2018;9:3097–3111.
29. Dard-Dascot C, Naquin D, d'Aubenton-Carafa Y, Alix K, Thermes C, *et al*. Systematic comparison of small RNA library preparation protocols for next-generation sequencing. *BMC Genomics* 2018;19:118.
30. Gruber AR, Findeiß S, Washietl S, Hofacker IL, Stadler PF. RNAz 2.0: improved noncoding RNA detection. *Pac Symp Biocomput* 2010:69–79.

31. Heiges M, Wang H, Robinson E, Aurrecoechea C, Gao X, *et al*. CryptoDB: a *Cryptosporidium* bioinformatics resource update. *Nucleic Acids Res* 2006;34:D419-22.
32. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet j* 2011;17:10.
33. Kalvari I, Nawrocki EP, Argasinska J, Quinones-Olvera N, Finn RD, *et al*. Non-coding RNA analysis using the Rfam database. *Curr Protoc Bioinformatics* 2018;62:e51.
34. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 2013;29:2933–2935.
35. Baptista RP, Li Y, Sateriale A, Sanders MJ, Brooks KL, *et al*. Long-read assembly and comparative evidence-based reanalysis of *Cryptosporidium* genome sequences reveal expanded transporter repertoire and duplication of entire chromosome ends including subtelomeric regions. *Genome Res* 2022;32:203–213.
36. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al*. The sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
37. Langmead B. Aligning short sequencing reads with Bowtie. *Curr Protoc Bioinformatics* 2010;Chapter 11:Unit .
38. Langenberger D, Bermudez-Santana C, Hertel J, Hoffmann S, Khaitovich P, *et al*. Evidence for human microRNA-offset RNAs in small RNA sequencing data. *Bioinformatics* 2009;25:2298–2301.
39. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, *et al*. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat Methods* 2012;9:811–814.
40. Altschul SF, Boguski MS, Gish W, Wootton JC. Issues in searching molecular sequence databases. *Nat Genet* 1994;6:119–129.
41. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841–842.
42. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;22:4673–4680.
43. Derrien T, Estellé J, Marco Sola S, Knowles DG, Raineri E, *et al*. Fast computation and applications of genome mappability. *PLoS One* 2012;7:e30377.
44. Will S, Joshi T, Hofacker IL, Stadler PF, Backofen R. LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *RNA* 2012;18:900–914.
45. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol* 2007;3:e65.
46. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res* 2019;47:W256–W259.
47. Lowe TM, Eddy SR. A computational screen for methylation guide snoRNAs in yeast. *Science* 1999;283:1168–1171.
48. de Araujo Oliveira JV, Costa F, Backofen R, Stadler PF, Machado Telles Walter ME, *et al*. SnoReport 2.0: new features and a refined Support Vector Machine to improve snoRNA identification. *BMC Bioinformatics* 2016;17:464.
49. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, Tafer H, Flamm C, *et al*. ViennaRNA Package 2.0. *Algorithms Mol Biol* 2011;6:26.
50. Darty K, Denise A, Ponty Y. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* 2009;25:1974–1975.
51. Friedländer MR, Mackowiak SD, Li N, Chen W, Rajewsky N. miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Res* 2012;40:37–52.
52. Loher P, Telonis AG, Rigoutsos I. MINTmap: fast and exhaustive profiling of nuclear and mitochondrial tRNA fragments from short RNA-seq data. *Sci Rep* 2017;7:41184.
53. Xie Y, Yao L, Yu X, Ruan Y, Li Z, *et al*. Action mechanisms and research methods of tRNA-derived small RNAs. *Signal Transduct Target Ther* 2020;5:109.
54. Molla-Herman A, Angelova MT, Ginestet M, Carré C, Antoniewski C, *et al*. tRNA fragments populations analysis in mutants affecting tRNAs processing and tRNA methylation. *Front Genet* 2020;11:518949.
55. Eddy SR, Durbin R. RNA sequence analysis using covariance models. *Nucleic Acids Res* 1994;22:2079–2088.
56. Šlapeta J. *Cryptosporidiosis* and *Cryptosporidium* species in animals and humans: a thirty colour rainbow? *Int J Parasitol* 2013;43:957–970.
57. Panda AC, De S, Grammatikakis I, Munk R, Yang X, *et al*. High-purity circular RNA isolation method (RPAD) reveals vast collection of intronic circRNAs. *Nucleic Acids Res* 2017;45:12.
58. Bevilacqua PC, Ritchey LE, Su Z, Assmann SM. Genome-wide analysis of RNA secondary structure. *Annu Rev Genet* 2016;50:235–266.
59. Schlick T, Pyle AM. Opportunities and challenges in RNA structural modeling and design. *Biophys J* 2017;113:225–234.
60. Wickiser JK, Winkler WC, Breaker RR, Crothers DM. The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch. *Mol Cell* 2005;18:49–60.
61. Li F, Zheng Q, Vandivier LE, Willmann MR, Chen Y, *et al*. Regulatory impact of RNA secondary structure across the *Arabidopsis* transcriptome. *Plant Cell* 2012;24:4346–4359.
62. Wan Y, Qu K, Ouyang Z, Kertesz M, Li J, *et al*. Genome-wide measurement of RNA folding energies. *Mol Cell* 2012;48:169–181.
63. Makarova JA, Kramerov DA. Analysis of C/D box snoRNA genes in vertebrates: The number of copies decreases in placental mammals. *Genomics* 2009;94:11–19.
64. Leader DJ, Clark GP, Watters J, Beven AF, Shaw PJ, *et al*. Clusters of multiple different small nucleolar RNA genes in plants are expressed as and processed from polycistronic pre-snoRNAs. *EMBO J* 1997;16:5742–5751.
65. Waldl M, Will S, Wolfinger MT, Hofacker IL, Stadler PF. Bi-alignments as models of incongruent evolution of RNA sequence and structure. *Bioinformatics* 2019:631606.
66. Liao J-Y, Guo Y-H, Zheng L-L, Li Y, Xu W-L, *et al*. Both endo-siRNAs and tRNA-derived small RNAs are involved in the differentiation of primitive eukaryote Giardia lamblia. *Proc Natl Acad Sci U S A* 2014;111:14159–14164.
67. Couvillion MT, Sachidanandam R, Collins K. A growth-essential *Tetrahymena* Piwi protein carries tRNA fragment cargo. *Genes Dev* 2010;24:2742–2747.
68. Fricker R, Brogli R, Luidalepp H, Wyss L, Fasnacht M, *et al*. A tRNA half modulates translation as stress response in *Trypanosoma brucei*. *Nat Commun* 2019;10:118.
69. Wang Z, Wei C, Hao X, Deng W, Zhang L, *et al*. Genome-wide identification and characterization of transfer RNA-derived small RNAs in *Plasmodium falciparum*. *Parasit Vectors* 2019;12:36.
70. Green D, Fraser WD, Dalmay T. Transfer RNA-derived small RNAs in the cancer transcriptome. *Pflugers Arch* 2016;468:1041–1047.
71. Li S, Xu Z, Sheng J. tRNA-derived small RNA: a novel regulatory small non-coding RNA. *Genes (Basel)* 2018;9:E246.
72. Kuscu C, Kumar P, Kiran M, Su Z, Malik A, *et al*. tRNA fragments (tRFs) guide Ago to regulate gene expression post-transcriptionally in a Dicer-independent manner. *RNA* 2018;24:1093–1105.
73. Lambertz U, Oviedo Ovando ME, Vasconcelos EJR, Unrau PJ, Myler PJ, *et al*. Small RNAs derived from tRNAs and rRNAs are highly enriched in exosomes from both old and new world *Leishmania* providing evidence for conserved exosomal RNA Packaging. *BMC Genomics* 2015;16:151.
74. Wang Z, Xi J, Hao X, Deng W, Liu J, *et al*. Red blood cells release microparticles containing human argonaute 2 and miRNAs to target genes of *Plasmodium falciparum*. *Emerg Microbes Infect* 2017;6:e75.
75. Wang Y, Gong A-Y, Ma S, Chen X, Li Y, *et al*. Delivery of parasite RNA transcripts into infected epithelial cells during *Cryptosporidium* infection and its potential impact on host gene transcription. *J Infect Dis* 2017;215:636–643.