BMC Genomics

RESEARCH ARTICLE Open Access

A comparative genome-wide study of ncRNAs in trypanosomatids

Tirza Doniger[1], Rodolfo Katz[1,2], Chaim Wachtel[1,2], Shulamit Michaeli[1,2], Ron Unger[1*]

Abstract

Background: Recent studies have provided extensive evidence for multitudes of non-coding RNA (ncRNA) transcripts in a wide range of eukaryotic genomes. ncRNAs are emerging as key players in multiple layers of cellular regulation. With the availability of many whole genome sequences, comparative analysis has become a powerful tool to identify ncRNA molecules. In this study, we performed a systematic genome-wide in silico screen to search for novel small ncRNAs in the genome of *Trypanosoma brucei* using techniques of comparative genomics.

Results: In this study, we identified by comparative genomics, and validated by experimental analysis several novel ncRNAs that are conserved across multiple trypanosomatid genomes. When tested on known ncRNAs, our procedure was capable of finding almost half of the known repertoire through homology over six genomes, and about two-thirds of the known sequences were found in at least four genomes. After filtering, 72 conserved unannotated sequences in at least four genomes were found, 29 of which, ranging in size from 30 to 392 nts, were conserved in all six genomes. Fifty of the 72 candidates in the final set were chosen for experimental validation. Eighteen of the 50 (36%) were shown to be expressed, and for 11 of them a distinct expression product was detected, suggesting that they are short ncRNAs. Using functional experimental assays, five of the candidates were shown to be novel H/ACA and C/D snoRNAs; these included three sequences that appear as singletons in the genome, unlike previously identified snoRNA molecules that are found in clusters. The other candidates appear to be novel ncRNA molecules, and their function is, as yet, unknown.

Conclusions: Using comparative genomic techniques, we predicted 72 sequences as ncRNA candidates in *T. brucei*. The expression of 50 candidates was tested in laboratory experiments. This resulted in the discovery of 11 novel short ncRNAs in procyclic stage *T. brucei*, which have homologues in the other trypansomatids. A few of these molecules are snoRNAs, but most of them are novel ncRNA molecules. Based on this study, our analysis suggests that the total number of ncRNAs in trypanosomatids is in the range of several hundred.

Background

Non-coding RNA (ncRNA) genes produce functional RNA molecules, but these molecules do not encode for protein products; rather, these RNA molecules directly participate in various cellular processes. For many years, only a few such ncRNA molecules were known, mainly represented by transfer-RNA (tRNA), ribosomal-RNA (rRNA), small nuclear RNA (snRNA) and small nucleolar RNA (snoRNA). The possible existence of additional types of ncRNA molecules was given little consideration, as the fundamental biological principle was that almost all genes are translated into proteins. As a result, most studies have focused their efforts primarily on protein discovery. The appreciation for the role of untranslated RNAs in the cell has changed dramatically over the past decade. Recent work has shown that the incidence and importance of ncRNA molecules has been underestimated [1-3]. ncRNAs are emerging as key players in multiple layers of cellular regulation [4-7]. In addition, it has been speculated that there are many additional types of ncRNA that have yet to be discovered.

However, systematic computational and experimental identification of these molecules has been difficult. The challenge of predicting ncRNAs from primary sequence

* Correspondence: ron@biocom1.ls.biu.ac.il
[1]The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan 52900, Israel
Full list of author information is available at the end of the article

is that they lack the known signals, such as start and stop codons as well as the triplet periodicity, which are distinguishing features of protein coding genes. Furthermore, discriminating between ncRNAs and protein-coding mRNAs is not a trivial task. ncRNAs, especially long ones, may contain open reading frames [8,9].

Over the years, several tools for identifying specific ncRNA family members have been developed. These programs generally exploit the fact that some ncRNA classes have relatively well-defined sequence and/or structural characteristics (i.e. tRNAs [10], snoRNAs (H/ACA [11-14] and C/D [15]) and miRNA [16,17]). General non-family specific tools for identifying ncRNA genes have had more limited success. Many ncRNAs have conserved secondary structures, despite having primary sequences that are often highly variable. This resulted in compensatory changes during evolution that are consistent with the conservation of a consensus secondary structure, and can be detected by a stochastic context-free grammar (SCFG) or hidden Markov models (HMMs) that may be used in conjunction with thermodynamic stability (i.e. qRNA [18], RNAz [19]).

ncRNA molecules can be experimentally detected by selecting for small molecules and preparing a cDNA library as was demonstrated by [20]. Most recently, the next generation sequencing technologies have become powerful tools for ncRNA discovery (see [21]). However, laboratory techniques for identifying RNA molecules are often expensive, time-consuming, and labor-intensive. In addition, these experimental methods have a bias toward highly abundant molecules and can miss RNAs that are only present under specific physiological conditions or during specific developmental stages. Thus, *in silico* methods for identifying RNA molecules have greatly complemented experimental work [22-24].

With the availability of many whole genome sequences, comparative analysis has become a powerful tool to study sequence similarities and differences between various organisms. Comparative genomics is an approach that has been used to aid in the discovery of genes, regulatory elements and gene structure [25-27]. It has also been shown as a powerful tool for identifying ncRNA [28-32].

Comparative genomics can serve as a powerful filter for ncRNA; it sifts genomic DNA and yields a subset of sequences that are enriched for ncRNA sequences. Comparative genome-wide studies for the purpose of detecting ncRNAs have been performed in a range of organisms from bacteria to humans. The number of predicted ncRNAs across the evolutionary scale varies widely. In human and higher vertebrates, computational [33,34] and experimental studies [35,36] indicate a number of putative ncRNAs in the range of tens of thousands. In contrast, in urochordates [37], nematodes [38], and drosophilids [39] the predicted numbers are lower, in the range of several thousand. Lower eukaryotes, such as yeast [29], and *Plasmodium* [40,41] are predicted to have ncRNAs in the range of several hundred. Studies of ncRNAs in prokaryotes, such as *E. coli* and other bacteria [18,24,42,43], suggest that the number of ncRNAs is in the low hundreds.

Trypanosomes are unicellular parasites, and are the cause of several devastating diseases affecting humans (e.g. Chagas disease and African sleeping sickness). Trypanosomatids are known for their non-conventional gene expression mechanisms, including RNA editing [44], and *trans*-splicing, a process that is required for the maturation of all mRNAs in these organisms whereby a small exon, encoded by a small RNA, the SL RNA, is donated to all pre-mRNA [45,46]. Trypanosomes have also been used as model organisms to study ncRNA, and over the years the U snRNAs [46], 7SL RNA [47] and snoRNAs [48-52] were described. However, many ncRNAs that have been found in other eukaryotes have not been identified in trypanosomes, such as many snoRNAs involved in RNA processing, RNase P, and telomerase RNA. These molecules remain elusive despite the fact that computer programs (i.e. Snoscan [15]) exist that are specifically designed to search for some classes of ncRNA (i.e. C/D), and are appropriate for identifying trypanosome homologues in genome-scale searches [51]. Based on experimental data from mapping of ribose methylation sites on ribosomal RNA in *T. brucei*, many C/D molecules that guide those modifications still remain to be discovered [49]. Many of the undiscovered ncRNA may have weak or novel motifs that would be impossible to identify without the use of comparative genomics. There have been several *in silico* genome-wide studies in trypanosomes to search for snoRNAs [14,51,53]. Recently, a genome-wide computational study of functional RNA elements in *T. brucei* [54] was published. The genomes of *T. brucei* and *L. braziliensis* were compared using a binomial-based model to assess conservation followed by a QRNA [18] analysis. After filtering by QRNA score and annotation, a total of 53 ncRNA candidates were reported.

Here, we describe a systematic *in silico* screen to identify conserved non-protein-coding genes across multiple trypanosomatid genomes, and prediction of 72 sequences as novel ncRNA candidates. The expression of 50 candidates was tested in laboratory experiments; 18 molecules were shown to be expressed, and for 11 of them there is strong evidence that they represent novel short ncRNAs in procyclic stage *T. brucei*, or their homologues in the other trypansomatids. The RNAs that do not belong to the previously described most abundant families of small RNAs, such as C/D and H/ACA snoRNAs or RNAs binding the Sm or Lsm proteins, were termed RNAs of Unknown Function (RUFs).

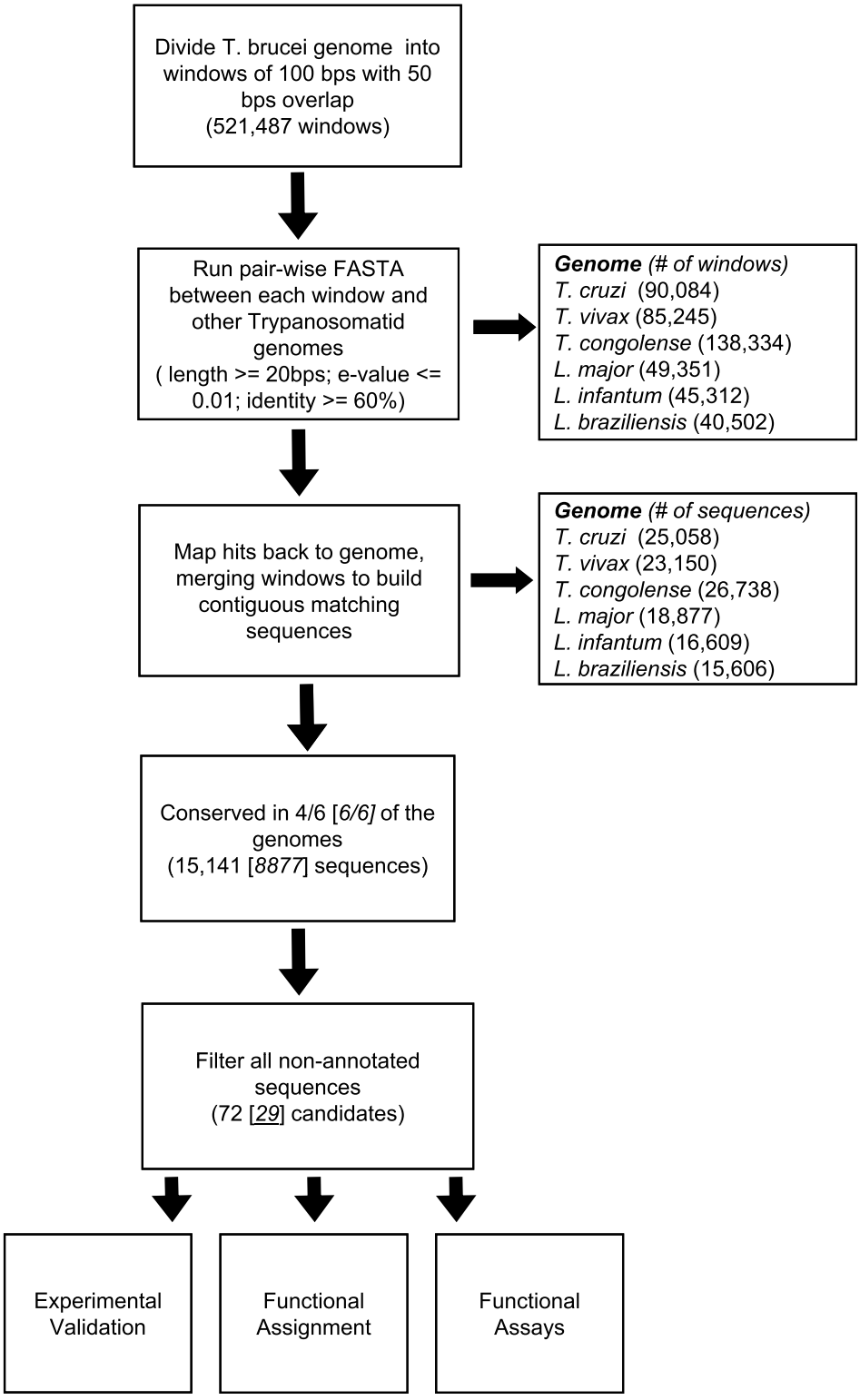

Figure 1 Schematic representation of the genome wide search pipeline. Using *T. brucei* as the reference genome, the chart describes each stage, and the number of candidates found at each stage. In the first stage, the number in the parentheses represents the number of homologous windows found. In the following stage, the number in parentheses represents the number of sequences found. The number of sequences conserved in six out of six genomes is given in the square brackets, following the number of sequences conserved in four out of six genomes.

Results

We report here the identification of novel ncRNAs based on the conservation among seven trypanosomatid species. Figure 1 shows the flow of the genome wide ncRNA search pipeline using *T. brucei* as the reference genome. As detailed in the methods, the pipeline is made up of five stages. We began our search with the *T. brucei* genome divided into windows of 100 nts with a 50 nt overlap in between windows, and performed a FASTA search against each one of the six other trypanosomatid genomes. Figure 1 shows the parameters used and the number of results obtained for each stage.

Assessment of performance

To assess the performance of our prediction scheme, we tested the protocol on the set of known ncRNA molecules of *T. brucei* (GeneDB version 4). When we required conservation in all of the six genomes, we were able to recover almost half of the known ncRNAs. When we loosened our constraints and required conservation in at least four of the six genomes, we were able to return almost 2/3 of the known ncRNAs (Table 1). The threshold of four genomes was chosen, as three of the genomes were from the *Leishmania* genus and three were from the *Trypanosoma* genus; thus conservation over at least four of the six genomes would force the conservation to bridge the divergence between *Leishmania* and *Trypanosoma*. A list of the 559 annotated ncRNA in GeneDB v4 is given in Additional File 1.

During the analysis, all known and hypothetical protein coding genes were filtered by comparing the coordinates of the candidate sequences to those of the annotation. This filter left a pool of 125 potential ncRNA candidates that are conserved in a minimum of four of the six genomes. However, the initial filtering of annotated sequences was based on comparing the coordinates of the sequences as appears in GeneDB. This comparison is fast, but it may miss proteins because of coordinate annotation problems, which are quite common. Thus, we checked the 125 candidates further by direct sequence comparison to see if they match any annotated gene. As annotation in the *T. brucei* genome is incomplete, we compared our candidates against the annotated genes of both *T. brucei* and *L. major*. Using BLAST comparison versus the *T. brucei* and *L. major* annotated sequences, a significant number of candidates (47 of the 125) were found to be highly similar to known coding sequences. Most of these sequences were simply a result of incomplete genome annotation. For example, several of the ribosomal RNA proteins (LmjF28.2460 ribosomal protein S29, putative and LmjF36.3750 40S ribosomal protein S27), which are highly conserved, were not annotated in *T. brucei*. We also found six previously described RNAs that are not reported in GeneDB. For example, the screen identified selenocysteine-tRNA [55], whose sequence had been unannotated in the genome, while instead sRNA-76 [56] was labeled as selenocysteine-tRNA, and was also identified in the final set (candidate 7). A list of the additional RNA genes that have been reported previously in the literature, but have not yet been incorporated or are misannotated in the GeneDB genome annotation is provided as Supplementary Material (Additional File 2). These include MRP RNA [49], snR30 [48], U5 [57], tRNA-sec [55], sRNA-76 [56], and several previously identified snoRNA clusters [14,49,51].

At this point we were left with a total of 72 candidates that are conserved in 4/6 genomes, out of which 29 are conserved in 6/6 genomes. Table 2 summarizes the number of sequences found in the 6/6 and 4/6 genome conservation analysis categorized according to their annotation. The complete list of all the sequences of the 72 ncRNA candidates is provided as Additional File 3. Searches of the RFAM database using BLAST on these 72 sequences did not provide any additional annotation information, suggesting that these may be trypanosome specific ncRNAs, or alternatively the sequence similarity to other organisms is too low to be detected. Note that

Table 1 Assessment of performance on known ncRNA found in GeneDB

Type of ncRNA	Annotated in Genedb v4	All 6 genomes	4 of 6 genomes
rRNA	106	22 (21%)	92 (87%)
snRNA	6	3 (50%)	5 (83%)
snoRNA	353	110 (31%)	188 (53%)
tRNA	65	64 (98%)	65 (100%)
misc RNA	29	28 (97%)	29 (100%)
Total	559	227 (41%)	379 (68%)

Using the set of known ncRNA as the standard to evaluate the performance of our algorithm in detecting ncRNAs, the numbers of ncRNA molecules detected are listed, according to families, including those conserved in all six genomes or in 4/6 genomes (note that 6/6 conservation is a subset of the 4/6 conservation). The percent of the ncRNA family that was detected in the screen is listed in the parentheses.

Table 2 Number of candidates from the different subtypes of RNA

	4 of 6 genomes	All 6 genomes
Total	15141	8877
Annotated proteins	7871	5482
Hypothetical proteins	6819	3139
Known ncRNA	379	227
Not annotated	72	29

Total loci found when requiring conservation in all six genomes or when requiring at least four genomes. The total was then further categorized based on each loci's annotation.

our method cannot detect the strand that contains the candidate molecule as conservation is the same for both strands. However, since trypanosomes have polycistronic transcription, we can obtain information about the direction of transcription from that of flanking genes. In cases where flanking genes were not sufficient to determine the direction of transcription, the sequences from both strands were subjected to the experimental validation step described below.

We checked for redundancy between the 72 candidates and found that Candidates # 85 and #90 shared 98% identity to each other and 70% identity with candidate #78. Candidates # 89 and #99 shared almost 100% identity to each other and 88% identity with candidate # 124. Candidates 68 and 70 shared 63% identity. Interestingly none of these candidates were among the molecules that we were able to validate experimentally.

Experimental Verification

Fifty of the 72 candidates in the final set were chosen for experimental validation. Fifteen were chosen from the sequences that were conserved in six of six genomes, and the rest were chosen randomly from the remaining candidates. The list of candidates sent for experimental verification appears in the comments to Additional File 3, and Additional File 4 provides the list of primers. Eighteen of the 50 candidates were shown to be expressed in cells. Expression was detected by primer extension assay that exactly determines the 5' end of the molecule. The strength of the signal reflected the abundance of the RNA, as the same amount of radio labelled primer and RNA were used in each experiment. Note that we did not use an internal control of very abundant RNA because it often affects the ability of non-abundant RNA to prime. Rather, we performed the primer extension using U3 snRNA as an internal control. This RNA was chosen because it is stable and tends not to degrade. However, the presence of the U3 oligo in the reaction reduced the efficiency of extension from our tested RNA (see Additional File 5).

Out of the 50 molecules, 32 did not show expression in the primer extension experiment described above. 18 molecules did yield extension products and 11 of the 18 had a distinct extension product suggesting that they are distinct small RNAs. The others yielded multiple bands, which may reflect the extension of a long polycistronic RNA, but probably not of a single small RNA (see Figure 2). While we mapped the 5' end of the candidates by primer extension, the full size of the products is unknown, as there is no information about their 3' end. However, for most of the distinct bands (and for some of the multiple bands) the size predicted by the bioinformatic analysis was quite reliable. This is a surprising and encouraging result considering the thresholds and cut-offs that are inherently somewhat arbitrary in bioinformatic analysis.

Note that even some of the candidates that were not expressed may still be ncRNAs that are expressed in another part of the parasite's life cycle. We analyzed expression only in procyclic form, and it is possible that the other RNAs are stage specific and are expressed only at 37°C when the parasite lives in the mammalian host. Indeed, we previously identified snoRNAs that are expressed better in the bloodstream form [49]. However, for the purposes of evaluating the performance of our procedure, we considered candidates that did not show a distinct band in our assay as false predictions.

Although Northern blot would be a better approach to show that the identified candidates are indeed small RNAs, the majority of the novel RNAs identified by this study were not abundant. There are only two that were abundant as determined by primer extension: tRNA-sec and candidate #28. tRNA-sec does appear abundant in the Northern blot, but candidate #28, while appearing strong by primer extension, gives a relatively weak band by Northern analysis (see Additional File 6); hence the remaining molecules, which were not abundant on the primer extension assay, are not likely to be clearly detected by Northern analysis. Note that in these two cases where we compared primer extension with Northern analysis, the sizes of the molecules were consistent.

In order to evaluate the performance of our prediction scheme we needed to estimate the True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) rates of the ncRNA prediction. TP represents the predictions that turn out to be correct, and our analysis yielded 379 (the number of known ncRNA molecules that we "identified") plus the 11 molecules we confirmed experimentally. FN can be estimated by the number of known ncRNA that our methods missed and there are 180 such molecules (559 known ncRNA molecules minus the 379 detected). FP corresponds to the number of predictions that were shown to be wrong which is 39 (50-11). Calculating the TN values is meaningless since most of the genome is not comprised of ncRNA. Thus, the calculated TN value would be in the millions, and while this would make the performance measures that are dependent on TN (like Specificity which is defined as TN/(TN+FP)) seem to be extremely good, this doesn't reflect true performance characteristics.

However, even when ignoring TN, we can estimate the Sensitivity (defined as TP/(TP+FN) to be 0.68 and the Positive Predictive Value (PPV also known as Precision, defined by TP/(TP+FP)) to be 0.9. Note that if we consider the additional 18 molecules that show expression (although with multiple bands) as positive predictions, as well, the score would be even somewhat higher.

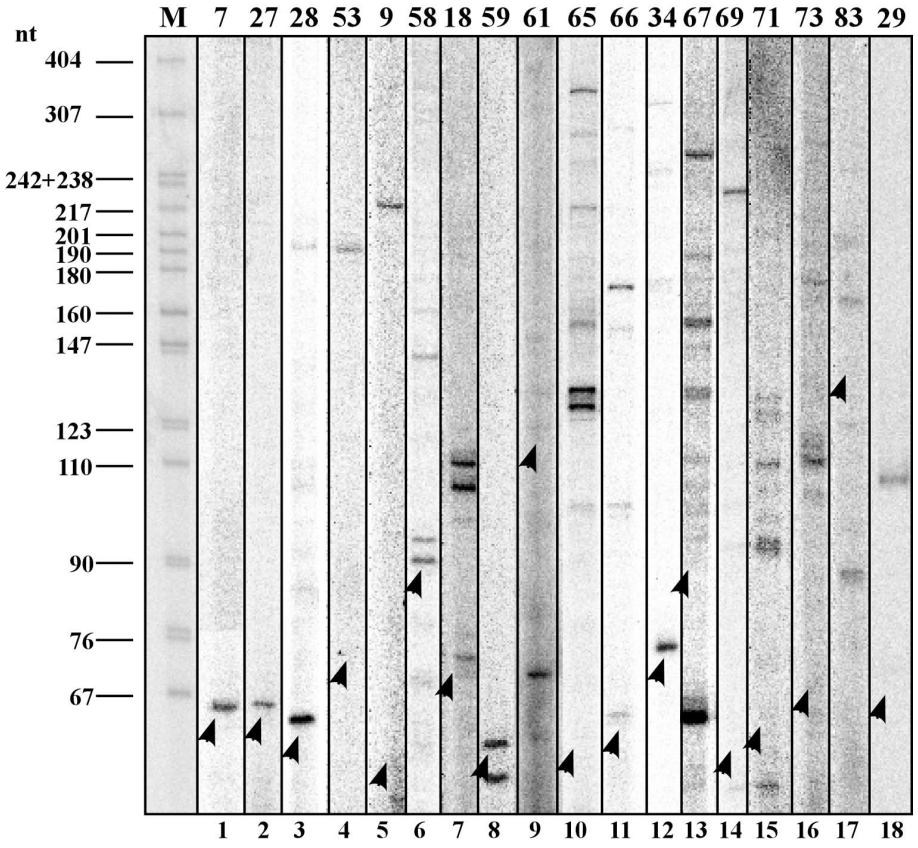

Figure 2 Results of expressed candidates in primer extension assay. RNA was subjected to primer extension using the oligonucleotide specified in Additional File 4. The products were separated on a 6% polyacrylamide. M-DNA marker, labelled *pBR322* DNA *MspI* digest. The arrowheads indicate the 5' end of full length transcripts. The numbers above the lanes indicate the candidates listed by candidate number as found in column 1 in Additional File 3, while the numbers below the lanes are sequential. Note that the gel is a composite. However, each experiment was performed in the same way. The same batch of total RNA was used, and the same amount of gel purified primer was used (50,000 cpm). Of the lanes analyzed, 11 (1,2,3,4,5,7,8,11,12,14,18) show extension products that are distinct or highly dominant and agree with the size predicted by the bioinformatic analysis.

To examine if the novel RNAs belong to known families of RNA, we examined their level in *T. brucei* cells depleted of core RNA proteins by RNAi-silencing. NOP58 silenced cells (previously described [49]) were used to classify RNAs as C/D snoRNAs, and CBF5 silenced cells [48] were used to identify H/ACA RNAs. Five of the identified RNA species were assigned to their respective families (4 C/Ds, and 1 H/ACA), and the others remain RNAs of unknown function (RUFs). The level of the RUFs was examined in cells silenced for the C/D and H/ACA core proteins as described above, and in cells depleted for Lsm8 and SmD1, and their levels were unchanged, suggesting that these are novel small RNAs, not belonging to known classes of small RNAs, and have binding proteins that are yet to be discovered.

Novel snoRNAs

Most eukaryotic C/D box and H/ACA snoRNAs guide 2'-O methylation (Nm) and pseudouridylation on specific nucleotides on the rRNA or snRNAs, and are also involved in rRNA processing [5]. To date, 64 C/D snoRNAs and 48 H/ACA snoRNAs [14,49-51] have been described in *T. brucei*, and 62 C/D and 37 H/ACA snoRNAs [53] were described in *L. major*. Among the candidates, four C/D box (candidates 28, 29, 34, and 69) and one H/ACA snoRNA (candidate 9) (See Figure 3a for cluster structure and experimental gels) were found. Candidates 28 and 29 were found as a cluster, and upon further inspection of the flanking region, two additional C/D snoRNAs were identified in this cluster. Candidates 9, 34 and 69 were found in the genome as single-copy genes. Proposed interaction domains for several of these snoRNAs are presented in Figure 3b, while no putative target was identified for the others. Interestingly, a continuous 13 bp complementarity was identified between TB2Cs1C1 and another C/D snoRNA TB9Cs3C2 [51]. The box structure of the four C/D snoRNA presented in Figure 3a is depicted in Additional File 7. Positive

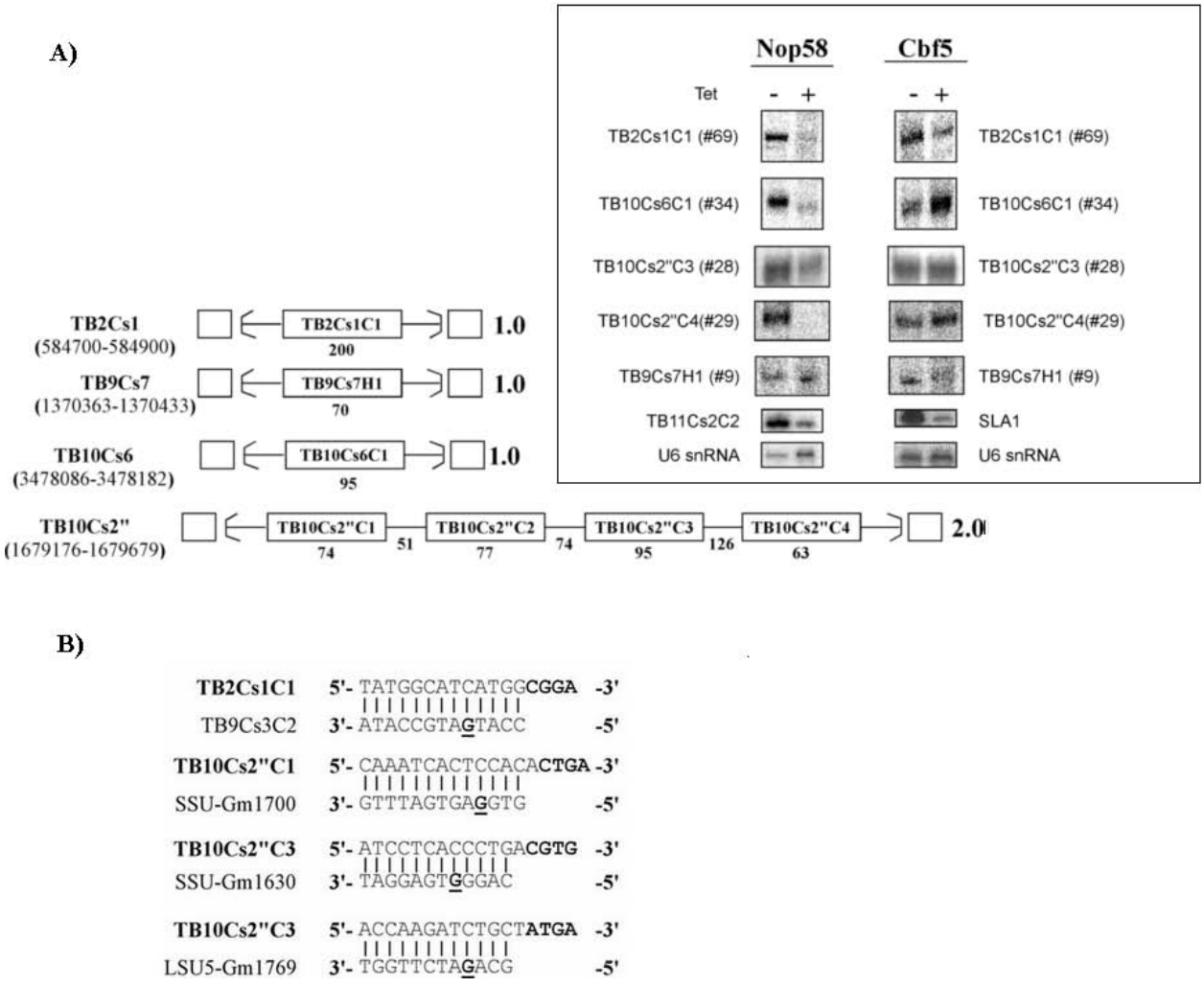

Figure 3 Effects of NOP58 and CBF5 knockdown. (a) Effects of NOP58 and CBF5 silencing on novel snoRNAs and schematic representation of the clusters encoding the snoRNAs. RNA was prepared from cells carrying either NOP58 or CBF5 silencing constructs before induction with Tetracycline (-), and 3 days after addition of tetracycline (+). The RNA was analyzed by primer extension and separated on a 6% denaturing polyacrylamide gel. The level of U6 snRNA was used to examine the amount of RNA the samples. The same RNA was used for the different primer extension assays. The positions of the snoRNAs in the genome are given on the left. The sizes of the snoRNAs and the intergenic regions (in nucleotides) are given below each diagram. The number of times the gene cluster (delineated by parentheses) is repeated is given on the right. The inset shows the controls: TB11Cs2C2, a known C/D molecule, and SLA1, a known H/ACA molecule, for which a significant reduction in expression following silencing can be seen, and U6 snRNA, a non-snoRNA molecule which is not affected. Novel C/D and H/ACA molecules were named as such based on the nomenclature of Liang et al. 2005 [49]. (b) Putative interaction domains for the C/D molecules TB2Cs1C1, TB10Cs2"C1, and TB10Cs2"C3. For TB10Cs2"C3, two putative targets are listed.

and negative controls for these experiments are included as Additional File 8.

Novel RNA Candidates of unknown Function (RUFs)

The remaining candidates were not readily identifiable as belonging to any of the known ncRNA families. These sequences were highly conserved across multiple trypansomatids, and were not found in open reading frames. Several examples of multiple sequence alignments depicting the high conservation of these RUFs among different trypanosomatid species are shown in Figure 4. In addition, two candidates show potential base-pair complementarity to areas on ribosomal RNA. TB11-RUF5 has potential perfect complementarity to 13 continuous base pairs on LSU-β (296-308), and TB11-RUF2 has potential perfect complementarity to 12 continuous base pairs on LSU-β (337-348). Other candidates have potential complementarity to additional areas in the genome. TB8-RUF1 has potential complementarity of 19 out of 20 residues to a known coding sequence Tb927.8.1590/Tb08.29O9.320 (upl3 ubiquitin-protein ligase), and perfect 15 base-pair complementarity to Tb927.7.2080/Tb07.43M14.530

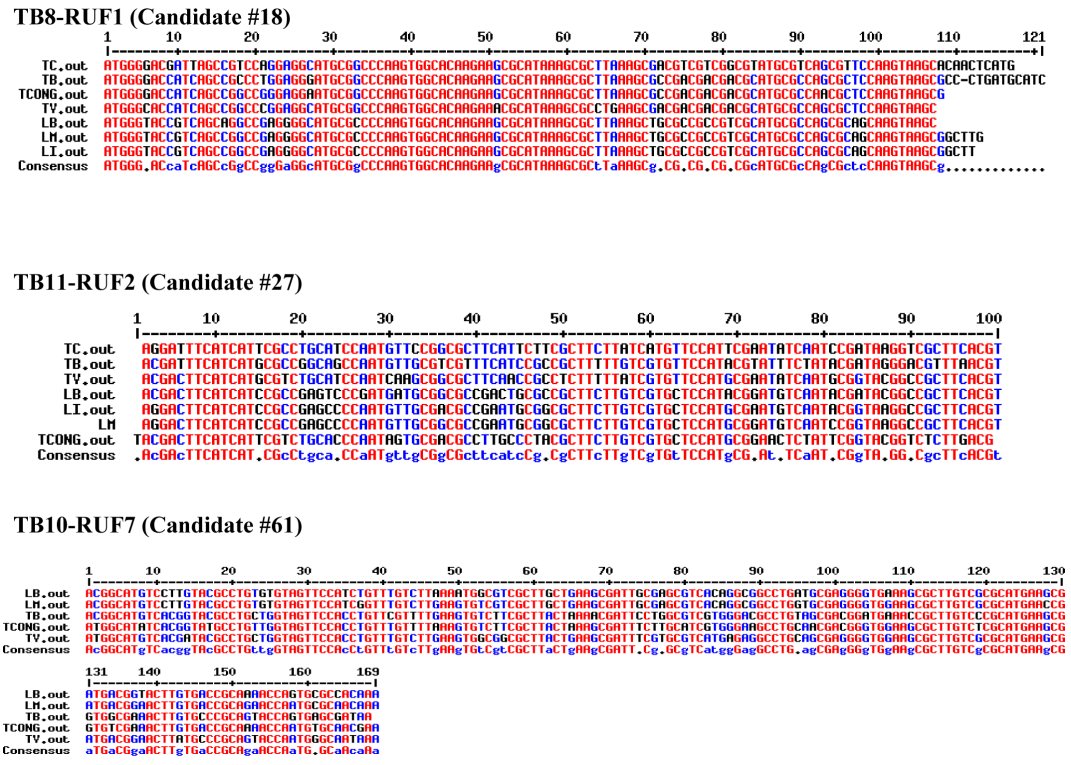

Figure 4 Examples of multiple sequence alignments of RUFs among different trypanosomatid species. Several examples illustrating the high level of conservation of the RUFs: These multiple sequence alignments were produced using the MulitAlign server [79]. High consensus (> 90% of the column is conserved) is indicated in red. Low consensus is indicated in blue (> = 50% conservation). Neutral columns are indicated in black.

(methyltransferase, putative). TB7-RUF8 has potential perfect complementarity to 16 continuous base pairs of Tb10.70.5440 (chaperone protein DNAJ, putative). The biological significance of this finding is currently unknown, since the statistical significance of complementarity with a run of even 15-20 nucleotides is not high when the entire genome is scanned. However, the target genes mentioned above are key regulators of proteolysis, chromatin state and protein folding, and these putative RUFs may function in regulating their level. This will require further experimental validation.

Discussion

We divide the Discussion into two sections; the first section deals with the technical aspects of the comparative genomics procedure, while the second will describe the implications of our findings on the repertoire of ncRNA molecules in Trypanosomes.

FASTA versus BLAST for RNA comparative genomics

For the purpose of RNA comparative genomics, one has to choose the most appropriate tool to efficiently compare the genomes with optimal sensitivity for detecting homologous ncRNA. BLAST [58] and FASTA [59] are the two popular heuristic programs for searching query sequences against a sequence database. Several papers have been published benchmarking the performance of BLAST and FASTA in protein-coding similarity searches [60,61]. One study [62] evaluated the sensitivity and specificity for the detection of ncRNA based on a variety of homology methods including BLAST and FASTA. Overall, FASTA was found to be more sensitive in detecting ncRNA than BLAST. In addition, FASTA's performance in detecting ncRNA was found to be comparable to WU-BLAST [63], though FASTA's run- time was faster. Nonetheless in the ncRNA community at large, the most popular tool of choice has been and continues to be BLASTN (i.e. [12,30,39,42,64]). As a test case for the preferential homology search methodology, the detection of a known snoRNA cluster (LM25Cs1) in *L. major* was examined. BLASTn and FASTA searches were performed using as the query a 100 kb area in *L. major* which included the snoRNA cluster, versus the whole *T. brucei* genome as the database. Based on our results from this small sample, FASTA, using the default settings, is more sensitive at identifying ncRNA even when we used more sensitive parameters for BLASTN (-r 1, -q -1 instead of the default +1/-3, personal communication William

Pearson). We also tested the sensitivity of performing the sequence comparison programs on the whole 100 kb, and on windows of 100 bps with 20% and 50% overlap. The result of these experiments, which is consistent with [62], is that FASTA should be preferred over BLAST for ncRNA searches.

Implications for the repertoire of ncRNA in Trypanosomes

In this study, a systematic *in silico* screen for conserved ncRNA among seven trypanosomatids is presented. In total, we found close to 100 candidates. One reason for the relatively low number of the additional ncRNAs that we found stems at least in part from the fact that studies from our labs, and those of others (i.e. [47,49,56,57,65,66]), already characterized the repertoire of Trypanosome snoRNAs, snRNAs, and other ncRNA species. Many of the recent studies which utilized comparative genomics to identify ncRNAs examined organisms that had very little previous ncRNA annotation. For example, in a study of *Plasmodium*, Chakrabarti et al. [40] identified several snRNAs (U1-U5), telomerase RNA, and about 30 snoRNAs. We believe that the fact that we were able to detect a third to half of the known ncRNA in trypanosomes by the bioinformatic method used indicates that our computational procedure is thorough.

In a recent study, Mao et al. [54] evaluated the conservation between *T. brucei* and *L. braziliensis* using a binomial-based model. QRNA was then used to identify likely ncRNA candidates. A total of 378 sequences were found with a significant QRNA score. Among the 378 sequences, 117 sequences were found to be highly significant when compared to randomized versions of the same sequence. Of the 117, 53 were unannotated. We evaluated the overlap between our final set and Mao's set of 378. We found three common sequences. They were: VSG pseudogene (candidate #121), a retro-transposon hotspot (candidate #98), and a novel C/D snoRNA (candidate #29, named TB10Cs2"C4). Note that although Mao et al. reported a low false positive rate - their algorithm only detected about 50% of the tRNAs, 20% of the rRNAs and 0% of the known snoRNAs. Comparing the performance of our procedure with this work, we conclude that the procedure used in our study is efficient and can serve as a useful tool for other systems, as well.

We propose that our findings can be used to estimate the total number of small ncRNA molecules in Trypanosomes. Of the 50 candidates tested, 18 novel ncRNAs were validated in procyclic stage trypanosomes. The experimental validation of a sample of 50 candidates suggest that about 1/3 of the candidates exist as novel small RNAs. On the other hand, when we tested our procedure on the known ncRNA we found that about 2/3 of the molecules have sufficient sequence conservation to be discovered by comparative sequence methods. Assuming that the rest of the ncRNA repertoire has similar characteristics and combining the two observations above, we can suggest that the total number of ncRNA molecules yet to be discovered in trypanosomatids is unlikely to be more than a few hundred.

There are several caveats to this claim. First, in our search, we did not consider the large amount (about 60% of the genes) of conserved hypothetical proteins. Many hypothetical proteins have been annotated as such because their sequence is found in open reading frames. However, some of these sequences may actually harbor ncRNA molecules. Several snoRNAs have been found within open reading frames. For example, Tb03.30p12.690, labelled as a hypothetical protein, overlaps with a C/D snoRNA TB3Cs2C1.

In addition, it is possible that there are many ncRNAs that are organism specific and cannot be detected by comparative methods. We notice that our study failed to identify several RNAs that are expected to exist in trypanosomatids such as telomerase RNA and RNAse P. Interestingly, Piccinelli et al. [67] studied RNase P and MRP in a variety of eukaryotes, but were unable to identify them in trypanosomatids. This is likely due to the fact that these RNAs are highly divergent even among closely related trypanosomatids. An interesting finding in this context is the detection of snoRNAs (TB2Cs1C1, TB10Cs6C1 and TB9Cs7H1) that are present in the genome as singletons, and are not part of the usual cluster organization of snoRNA in trypanosomatids. While obviously these two molecules were conserved enough among the different trypanosomatid species to be detected, other singleton molecules may be more diverse and hence harder to detect, suggesting that more such snoRNAs may exist.

Third, our extrapolation was based on our observation that only about 1/3 of candidate molecules were shown to be expressed. We cannot rule out the possibility that these candidate molecules are expressed at different stages in the life cycle of the parasite or under ambient environmental conditions. In *C. elegans* [68], it was shown that many ncRNAs are developmentally regulated and exhibit stage-specific function.

Conclusion

Taken together these issues limit our ability to quantitatively estimate, the number of ncRNA molecules in trypanosomatids. However, even if each one of these factors are off by a factor of two, our overall estimate should be in error by less than a single order of magnitude. Thus, we believe that our results supply an "order of magnitude" qualitative argument suggesting that there are relatively few remaining small ncRNA to be

identified. Since we found several dozen candidates, we estimate that not more than several hundred ncRNA molecules exist in each of the trypanosomatid genomes. Many of these molecules may be additional members of known ncRNA families, so that the expected number of novel families is limited.

It has been suggested that the genome of higher eukaryotes contain many thousands of as yet undiscovered ncRNA molecules. Washietl et al., [19], suggested that this repertoire includes short and long ncRNA molecules. Indeed, there is mounting evidence [69] that there are thousands of long ncRNA molecules (although their functional relevance is still under debate). However, we must note that there is no experimental evidence to support the claim of a large number of short ncRNA, except for the large variety of very short ncRNA (miRNA, piRNA) which are associated with the Dicer/Argonaut silencing system. Our findings support the view that at least for unicellular eukaryotes, the repertoire of small ncRNA is not likely to grow much beyond what is already known, and will remain in the hundreds and not thousands.

Methods

Genomic Data sources

Trypanosoma brucei (TB) genomic DNA and sequence annotation (version 4) was downloaded from GeneDB (http://www.genedb.org). GeneDB contains all available sequences from the 11 megabase chromsomes of *T. brucei* strain TREU927/4 GUTat10.1 generated by the *T. brucei* genome projects at The Institute for Genomic Research (TIGR's *T. brucei* project) and The Wellcome Trust Sanger Institute (Sanger's *T. brucei* project). *Trypanosoma cruzi* (version 4) (TC), *Leishmania major* (version 5.2) (LM), *Leishmania infantum* (version 2) (LI), and *Leishmania braziliensis* (version 1) (LB) genomic sequence data was also downloaded from GeneDB (ftp://ftp.sanger.ac.uk/pub/pathogens/). The nuclear genome of *Trypanosoma cruzi* CL Brener is being sequenced by the TIGR-Seattle Biomedical Research Institute-Karolinska Institute *T. cruzi* Sequencing Consortium (TSK-TSC) (http://www.jcvi.org/). The genome of *L. major Friedlin*, the reference strain (MHOM/IL/80/Friedlin, zymodeme MON-103), was sequenced as part of a multi-centre collaboration (Sanger Institute/EULEISH, Seattle Biomedical Research Institute, FMRP). The shotgun sequences of *T. vivax* (TV) and *T. congolense* (TCONG) were downloaded from GeneDB (ftp://ftp.sanger.ac.uk/pub/databases/). The Sanger Institute has also carried out a 5× coverage of the nuclear genome of *T. vivax*, as well as *T. congolense*. The sizes of the *T. brucei*, *T. cruzi*, *L. major* genomes are 25 Mb, 60 Mb, and 32 Mb respectively. These genomes have been published [70-72]. The *L. major* and *T. brucei* genomes are fully assembled. The *T. cruzi* genome has been fully sequenced, but its assembly is still in its preliminary stages. The *T. cruzi* genome is available as many large contigs.

Sequence similarity searches

The basis for our search strategy was to designate one trypanosomatid genome as a "reference" and to find all sequences in the other organisms that are similar to it. We chose to use *T. brucei* as the reference genome since it is fully sequenced, reasonably annotated, and because we have the experimental setup for candidate validation. While genome synteny maybe the preferable method to align genomes, we note that some of our genomes are not assembled and are still only available as shotgun sequences; thus we had to chose an alignment method that is based on relatively small windows. The reference genome, *T. brucei*, was divided into a window size of 100 bps with a sliding window of 50 bps. These sequences were searched for similarity against the other trypanosomatid genomes using FASTA [59]. We found FASTA to be more useful for this project than BLAST (see the Discussion). A Bio-PERL/PERL [73] script was written to post-process the FASTA results. Sequence matches were further analyzed if they fit the following criteria: 25 bps or longer, an e-value less than or equal to 0.01, and percent identity equal or greater than 60%. FASTA matches that passed the filter were then mapped back to the *T. brucei* genome. Areas that were less than 10 bps apart were concatenated. Conservation was defined by the number of genomes that had matches to the same corresponding segment of the genome. We considered areas that were conserved in at least four of the six genomes and those that were conserved in all six genomes. Sequences annotated as protein coding or hypothetical protein coding, were then filtered out.

General ncRNA Detection tools

BLAST [58] was run using the *T. brucei* ncRNA candidates versus the RFAM database (v6.1) [74] to search for sequence similarity to any known ncRNA.

Experimental Methods

Primer extension

RNA was prepared from *T. brucei* cells using the TRI-Reagent (Sigma). Primer extension analysis was performed as described [75,76] using 5'-end-labeled oligonucleotides specific to each target RNA. The extension products were analyzed on a 6% polyacrylamide/7 M urea gel and visualized by autoradiography. For examining the level of ncRNAs under silencing of the core RNA binding proteins, RNA was prepared from untreated cells and 3 days after the induction of silencing, as previously described [48,49,77,78].

Sequence Availability

Sequence data from this study were deposited in GeneDB. Accession numbers can be found in Additional File 9.

Additional material

Additional file 1: List of the annotated GeneDB v4 RNA genes in *T. brucei*. List of the annotated ncRNA found in GeneDB ver4, which were used as a standard to assess the success of our screen.

Additional file 2: List of missing/mis-annoted ncRNA. List of the additional RNA genes that have been reported previously in the literature, but have not yet been incorporated or are misannotated in the GeneDB genome annotation.

Additional file 3: Complete list of candidate ncRNA. The complete list of all the sequences of the 72 ncRNA candidates conserved in four of the six genomes. The first 29 were conserved in all of the six genomes.

Additional file 4: Oligonucleotides used in primer extension for candidates presented in Figure 2.

Additional file 5: Results of expressed candidates in primer extension assay with an internal control. Primer extension was performed as in Figure 2 with the addition of an internal control to each sample. The primer extension reactions contained a primer specific for the candidate as well as a primer specific to U3 snoRNA.

Additional file 6: Northern blot analysis of two of the candidates. RNA was prepared from PS cells, separated on a 10% denaturing polyacrylamide gel, and subjected to Northern analysis with the indicated oligonucleotide anti-sense probes.

Additional file 7: The box structure of the novel C/D molecules. For C/D snoRNA TB10Cs2"C3, TB10Cs2"C4, TB10Cs6C1, TB2Cs1C1 the canonical C and D box structure is shown.

Additional file 8: Effects of NOP58 and CBF5 knockdown on all candidates. RNA was prepared from cells carrying either NOP58 or CBF5 silencing constructs before induction with Tetracycline (-) and 3 days after addition of tetracycline (+). The RNA was analyzed by primer extension and separated on a 6% denaturing polyacrylamide gel. The level of U6 snRNA was used to examine the amount of RNA the samples. The same RNA was used for the different primer extension assays.

Additional file 9: GeneDB accession numbers of the new snoRNA molecules that were reported in this study. List of the newly annotated sequences with their GeneDB id.

Acknowledgements

This research was supported by a grant from the Israel-US Binational Science Foundation (BSF), and by an International Research Scholar's Grant from the Howard Hughes Foundation to S.M. S.M. holds the David and Inez Myers Chair in RNA silencing of diseases.

Author details

[1]The Mina and Everard Goodman Faculty of Life Sciences, Bar-Ilan University, Ramat-Gan 52900, Israel. [2]Advanced Materials and Nanotechnology Institute, Bar-Ilan University, Ramat-Gan 52900, Israel.

Authors' contributions

TD carried out all the computational work in this study. The experimental work was performed by RK and CW. SM and RU coordinated the project. TD, SM and RU wrote the manuscript. All authors have read and approved the final manuscript.

Received: 25 January 2010 Accepted: 4 November 2010
Published: 4 November 2010

References

1. Amaral PP, Dinger ME, Mercer TR, Mattick JS: **The eukaryotic genome as an RNA machine.** *Science* 2008, **319**:1787-1789.
2. Costa FF: **Non-coding RNAs: lost in translation?** *Gene* 2007, **386**:1-10.
3. Dinger ME, Amaral PP, Mercer TR, Mattick JS: **Pervasive transcription of the eukaryotic genome: functional indices and conceptual implications.** *Brief Funct Genomic Proteomic* 2009, **8**:407-423.
4. Erdmann VA, Barciszewska MZ, Hochberg A, de GN, Barciszewski J: **Regulatory RNAs.** *Cell Mol Life Sci* 2001, **58**:960-977.
5. Kiss T: **Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions.** *Cell* 2002, **109**:145-148.
6. Amaral PP, Mattick JS: **Noncoding RNA in development.** *Mamm Genome* 2008, **19**:454-492.
7. Kugel JF, Goodrich JA: **In new company: U1 snRNA associates with TAF15.** *EMBO Rep* 2009, **10**:454-456.
8. Dinger ME, Pang KC, Mercer TR, Mattick JS: **Differentiating protein-coding and noncoding RNA: challenges and ambiguities.** *PLoS Comput Biol* 2008, **4**:e1000176.
9. Frith MC, Bailey TL, Kasukawa T, Mignone F, Kummerfeld SK, Madera M, Sunkara S, Furuno M, Bult CJ, Quackenbush J, *et al*: **Discrimination of non-protein-coding transcripts from protein-coding mRNA.** *RNA Biol* 2006, **3**:40-48.
10. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence.** *Nucleic Acids Res* 1997, **25**:955-964.
11. Edvardsson S, Gardner PP, Poole AM, Hendy MD, Penny D, Moulton V: **A search for H/ACA snoRNAs in yeast using MFE secondary structure prediction.** *Bioinformatics* 2003, **19**:865-873.
12. Schattner P, Decatur WA, Davis CA, Ares M Jr, Fournier MJ, Lowe TM: **Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the Saccharomyces cerevisiae genome.** *Nucleic Acids Res* 2004, **32**:4281-4296.
13. Muller S, Charpentier B, Branlant C, Leclerc F: **A dedicated computational approach for the identification of archaeal H/ACA sRNAs.** *Methods Enzymol* 2007, **425**:355-387.
14. Myslyuk I, Doniger T, Horesh Y, Hury A, Hoffer R, Ziporen Y, Michaeli S, Unger R: **Psiscan: a computational approach to identify H/ACA-like and AGA-like non-coding RNA in trypanosomatid genomes.** *BMC Bioinformatics* 2008, **9**:471.
15. Lowe TM, Eddy SR: **A computational screen for methylation guide snoRNAs in yeast.** *Science* 1999, **283**:1168-1171.
16. Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP: **Vertebrate microRNA genes.** *Science* 2003, **299**:1540.
17. Hertel J, Stadler PF: **Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data.** *Bioinformatics* 2006, **22**: e197-e202.
18. Rivas E, Klein RJ, Jones TA, Eddy SR: **Computational identification of noncoding RNAs in E. coli by comparative genomics.** *Curr Biol* 2001, **11**:1369-1373.
19. Washietl S, Hofacker IL, Stadler PF: **Fast and reliable prediction of noncoding RNAs.** *Proc Natl Acad Sci USA* 2005, **102**:2454-2459.
20. Huttenhofer A, Kiefmann M, Meier-Ewert S, O'Brien J, Lehrach H, Bachellerie JP, Brosius J: **RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse.** *EMBO J* 2001, **20**:2943-2953.
21. Mardis ER: **The impact of next-generation sequencing technology on genetics.** *Trends Genet* 2008, **24**:133-141.
22. Brown JW, Clark GP, Leader DJ, Simpson CG, Lowe T: **Multiple snoRNA gene clusters from Arabidopsis.** *RNA* 2001, **7**:1817-1832.
23. Huang ZP, Chen CJ, Zhou H, Li BB, Qu LH: **A combined computational and experimental analysis of two families of snoRNA genes from Caenorhabditis elegans, revealing the expression and evolution pattern of snoRNAs in nematodes.** *Genomics* 2007, **89**:490-501.
24. Voss B, Georg J, Schon V, Ude S, Hess WR: **Biocomputational prediction of non-coding RNAs in model cyanobacteria.** *BMC Genomics* 2009, **10**:123.
25. Liu Y, Liu XS, Wei L, Altman RB, Batzoglou S: **Eukaryotic regulatory element conservation analysis and identification using comparative genomics.** *Genome Res* 2004, **14**:451-458.
26. Wang X, Haberer G, Mayer KF: **Discovery of cis-elements between sorghum and rice using co-expression and evolutionary conservation.** *BMC Genomics* 2009, **10**:284.

27. Sieglaff DH, Dunn WA, Xie XS, Megy K, Marinotti O, James AA: **Comparative genomics allows the discovery of cis-regulatory elements in mosquitoes.** *Proc Natl Acad Sci USA* 2009, **106**:3053-3058.
28. McCutcheon JP, Eddy SR: **Computational identification of non-coding RNAs in Saccharomyces cerevisiae by comparative genomics.** *Nucleic Acids Res* 2003, **31**:4119-4128.
29. Steigele S, Huber W, Stocsits C, Stadler PF, Nieselt K: **Comparative analysis of structured RNAs in S. cerevisiae indicates a multitude of different functions.** *BMC Biol* 2007, **5**:25.
30. Song D, Yang Y, Yu B, Zheng B, Deng Z, Lu BL, Chen X, Jiang T: **Computational prediction of novel non-coding RNAs in Arabidopsis thaliana.** *BMC Bioinformatics* 2009, **10(Suppl 1)**:S36.
31. Chen CL, Zhou H, Liao JY, Qu LH, Amar L: **Genome-wide evolutionary analysis of the noncoding RNA genes and noncoding DNA of Paramecium tetraurelia.** *RNA* 2009, **15**:503-514.
32. Kavanaugh LA, Dietrich FS: **Non-coding RNA prediction and verification in Saccharomyces cerevisiae.** *PLoS Genet* 2009, **5**:e1000321.
33. Washietl S, Hofacker IL, Lukasser M, Huttenhofer 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.
34. Torarinsson E, Sawera M, Havgaard JH, Fredholm M, Gorodkin J: **Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure.** *Genome Res* 2006, **16**:885-889.
35. Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, Stadler PF, Hertel J, Hackermuller J, Hofacker IL, *et al*: **RNA maps reveal new RNA classes and a possible function for pervasive transcription.** *Science* 2007, **316**:1484-1488.
36. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, *et al*: **The transcriptional landscape of the mammalian genome.** *Science* 2005, **309**:1559-1563.
37. Missal K, Rose D, Stadler PF: **Non-coding RNAs in Ciona intestinalis.** *Bioinformatics* 2005, **21(Suppl 2)**:ii77-ii78.
38. Missal K, Zhu X, Rose D, Deng W, Skogerbo G, Chen R, Stadler PF: **Prediction of structured non-coding RNAs in the genomes of the nematodes Caenorhabditis elegans and Caenorhabditis briggsae.** *J Exp Zool B Mol Dev Evol* 2006, **306**:379-392.
39. Rose D, Hackermuller J, Washietl S, Reiche K, Hertel J, Findeiss S, Stadler PF, Prohaska SJ: **Computational RNomics of drosophilids.** *BMC Genomics* 2007, **8**:406.
40. Chakrabarti K, Pearson M, Grate L, Sterne-Weiler T, Deans J, Donohue JP, Ares M Jr: **Structural RNAs of known and unknown function identified in malaria parasites by comparative genomics and RNA analysis.** *RNA* 2007, **13**:1923-1939.
41. Mourier T, Carret C, Kyes S, Christodoulou Z, Gardner PP, Jeffares DC, Pinches R, Barrell B, Berriman M, Griffiths-Jones S, *et al*: **Genome-wide discovery and verification of novel structured RNAs in Plasmodium falciparum.** *Genome Res* 2008, **18**:281-292.
42. Axmann IM, Kensche P, Vogel J, Kohl S, Herzel H, Hess WR: **Identification of cyanobacterial non-coding RNAs by comparative genome analysis.** *Genome Biol* 2005, **6**:R73.
43. Coenye T, Drevinek P, Mahenthiralingam E, Shah SA, Gill RT, Vandamme P, Ussery DW: **Identification of putative noncoding RNA genes in the Burkholderia cenocepacia J2315 genome.** *FEMS Microbiol Lett* 2007, **276**:83-92.
44. Stuart KD, Schnaufer A, Ernst NL, Panigrahi AK: **Complex management: RNA editing in trypanosomes.** *Trends Biochem Sci* 2005, **30**:97-105.
45. Agabian N: **Trans splicing of nuclear pre-mRNAs.** *Cell* 1990, **61**:1157-1160.
46. Liang XH, Haritan A, Uliel S, Michaeli S: **trans and cis splicing in trypanosomatids: mechanism, factors, and regulation.** *Eukaryot Cell* 2003, **2**:830-840.
47. Michaeli S, Podell D, Agabian N, Ullu E: **The 7SL RNA homologue of Trypanosoma brucei is closely related to mammalian 7SL RNA.** *Mol Biochem Parasitol* 1992, **51**:55-64.
48. Barth S, Hury A, Liang XH, Michaeli S: **Elucidating the role of H/ACA-like RNAs in trans-splicing and rRNA processing via RNA interference silencing of the Trypanosoma brucei CBF5 pseudouridine synthase.** *J Biol Chem* 2005, **280**:34558-34568.
49. Barth S, Shalem B, Hury A, Tkacz ID, Liang XH, Uliel S, Myslyuk I, Doniger T, Salmon-Divon M, Unger R, *et al*: **Elucidating the role of C/D snoRNA in rRNA processing and modification in Trypanosoma brucei.** *Eukaryot Cell* 2008, **7**:86-101.
50. Doniger T, Michaeli S, Unger R: **Families of H/ACA ncRNA molecules in trypanosomatids.** *RNA Biol* 2009, **6**:370-374.
51. Liang XH, Uliel S, Hury A, Barth S, Doniger T, Unger R, Michaeli S: **A genome-wide analysis of C/D and H/ACA-like small nucleolar RNAs in Trypanosoma brucei reveals a trypanosome-specific pattern of rRNA modification.** *RNA* 2005, **11**:619-645.
52. Uliel S, Liang XH, Unger R, Michaeli S: **Small nucleolar RNAs that guide modification in trypanosomatids: repertoire, targets, genome organisation, and unique functions.** *Int J Parasitol* 2004, **34**:445-454.
53. Liang XH, Hury A, Hoze E, Uliel S, Myslyuk I, Apatoff A, Unger R, Michaeli S: **Genome-wide analysis of C/D and H/ACA-like small nucleolar RNAs in Leishmania major indicates conservation among trypanosomatids in the repertoire and in their rRNA targets.** *Eukaryot Cell* 2007, **6**:361-377.
54. Mao Y, Najafabadi HS, Salavati R: **Genome-wide computational identification of functional RNA elements in Trypanosoma brucei.** *BMC Genomics* 2009, **10**:355.
55. Cassago A, Rodrigues EM, Prieto EL, Gaston KW, Alfonzo JD, Iribar MP, Berry MJ, Cruz AK, Thiemann OH: **Identification of Leishmania selenoproteins and SECIS element.** *Mol Biochem Parasitol* 2006, **149**:128-134.
56. Beja O, Ullu E, Michaeli S: **Identification of a tRNA-like molecule that copurifies with the 7SL RNA of Trypanosoma brucei.** *Mol Biochem Parasitol* 1993, **57**:223-229.
57. Dungan JM, Watkins KP, Agabian N: **Evidence for the presence of a small U5-like RNA in active trans-spliceosomes of Trypanosoma brucei.** *EMBO J* 1996, **15**:4016-4029.
58. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**:403-410.
59. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison.** *Proc Natl Acad Sci USA* 1988, **85**:2444-2448.
60. Shpaer EG, Robinson M, Yee D, Candlin JD, Mines R, Hunkapiller T: **Sensitivity and selectivity in protein similarity searches: a comparison of Smith-Waterman in hardware to BLAST and FASTA.** *Genomics* 1996, **38**:179-191.
61. Brenner SE, Chothia C, Hubbard TJ: **Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships.** *Proc Natl Acad Sci USA* 1998, **95**:6073-6078.
62. Freyhult EK, Bollback JP, Gardner PP: **Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA.** *Genome Res* 2007, **17**:117-125.
63. Gish W: **WU-blast.** 1996.
64. Stein LD, Bao Z, Blasiar D, Blumenthal T, Brent MR, Chen N, Chinwalla A, Clarke L, Clee C, Coghlan A, *et al*: **The genome sequence of Caenorhabditis briggsae: a platform for comparative genomics.** *PLoS Biol* 2003, **1**:E45.
65. Wieland B, Bindereif A: **Unexpected diversity in U6 snRNA sequences from trypanosomatids.** *Gene* 1995, **161**:129-133.
66. Xu Y, Ben-Shlomo H, Michaeli S: **The U5 RNA of trypanosomes deviates from the canonical U5 RNA: the Leptomonas collosoma U5 RNA and its coding gene.** *Proc Natl Acad Sci USA* 1997, **94**:8473-8478.
67. 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.
68. He H, Cai L, Skogerbo G, Deng W, Liu T, Zhu X, Wang Y, Jia D, Zhang Z, Tao Y, *et al*: **Profiling Caenorhabditis elegans non-coding RNA expression with a combined microarray.** *Nucleic Acids Res* 2006, **34**:2976-2983.
69. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, *et al*: **Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals.** *Nature* 2009, **458**:223-227.
70. El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, Ghedin E, Worthey EA, Delcher AL, Blandin G, *et al*: **The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease.** *Science* 2005, **309**:409-415.
71. Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, Sisk E, Rajandream MA, Adlem E, Aert R, *et al*: **The genome of the kinetoplastid parasite, Leishmania major.** *Science* 2005, **309**:436-442.
72. Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, Lennard NJ, Caler E, Hamlin NE, Haas B, *et al*: **The**

genome of the African trypanosome Trypanosoma brucei. *Science* 2005, **309**:416-422.
73. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, *et al*: **The Bioperl toolkit: Perl modules for the life sciences.** *Genome Res* 2002, **12**:1611-1618.
74. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR: **Rfam: an RNA family database.** *Nucleic Acids Res* 2003, **31**:439-441.
75. Liang XH, Liu L, Michaeli S: **Identification of the first trypanosome H/ACA RNA that guides pseudouridine formation on rRNA.** *J Biol Chem* 2001, **276**:40313-40318.
76. Xu Y, Liu L, Lopez-Estrano C, Michaeli S: **Expression studies on clustered trypanosomatid box C/D small nucleolar RNAs.** *J Biol Chem* 2001, **276**:14289-14298.
77. Mandelboim M, Barth S, Biton M, Liang XH, Michaeli S: **Silencing of Sm proteins in Trypanosoma brucei by RNA interference captured a novel cytoplasmic intermediate in spliced leader RNA biogenesis.** *J Biol Chem* 2003, **278**:51469-51478.
78. Liu Q, Liang XH, Uliel S, Belahcen M, Unger R, Michaeli S: **Identification and functional characterization of lsm proteins in Trypanosoma brucei.** *J Biol Chem* 2004, **279**:18210-18219.
79. Corpet F: **Multiple sequence alignment with hierarchical clustering.** *Nucleic Acids Res* 1988, **16**:10881-10890.

doi:10.1186/1471-2164-11-615
Cite this article as: Doniger *et al.*: A comparative genome-wide study of ncRNAs in trypanosomatids. *BMC Genomics* 2010 **11**:615.