PhD Thesis

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

UNIVERSIDADE FEDERAL DE MINAS GERAIS

Genomic and transcriptomic surveys for the study of ncRNAs with a focus on tropical parasites

Mainá Bitar

Belo Horizonte

February 2015

Universidade Federal de Minas Gerais

PhD Thesis

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOINFORMÁTICA

Genomic and transcriptomic surveys for the study of ncRNAs with a focus on tropical parasites

PhD candidate: Mainá Bitar

Advisor: Glória Regina Franco

Co-advisor: Martin Alexander Smith

Mainá Bitar Lourenço

Genomic and transcriptomic surveys for the study of ncRNAs with a focus on tropical parasites

Versão final
Tese apresentada ao Programa Interunidades de Pós-Graduação em Bioinformática do Instituto de Ciências Biológicas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do título de Doutor em Bioinformática.
Orientador: Prof^{a}. Dr^{a}. Glória Regina Franco

BELO HORIZONTE
2015

043 Bitar, Mainá.

Genomic and transcriptomic surveys for the study of ncRNAs with a focus on tropical parasites [manuscrito] / Mainá Bitar. – 2015.

134 f. : il. ; 29,5 cm.

Orientador: Glória Regina Franco. Coorientador: Martin Alexander Smith.

Tese (doutorado) – Universidade Federal de Minas Gerais, Instituto de Ciências Biológicas. Programa de Pós-Graduação em Bioinformática.

1. Bioinformática - Teses. 2. Trypanosoma cruzi. 3. Schistosoma mansoni. 4. Genômica. 5. Transcriptoma. 6. Trans-Splicing. I. Franco, Glória Regina. II. Smith, Martin Alexander. III. Universidade Federal de Minas Gerais. Instituto de Ciências Biológicas. IV. Título.

CDU: 573:004

Ficha catalográfica elaborada por Fabiane C. M. Reis – CRB 6/2680

Universidade Federal de Minas Gerais
Instituto de Ciências Biológicas
Programa de Pós-Graduação em Bioinformática

ATA DA DEFESA DE TESE

50/2015
entrada
2º/2011
CPF:
055.276.257-14

Mainá Bitar Lourenço

Às treze horas do dia **20 de fevereiro de 2015**, reuniu-se, no Instituto de Ciências Biológicas da UFMG, a Comissão Examinadora de Tese, indicada pelo Colegiado de Programa, para julgar, em exame final, o trabalho intitulado: **"Genomic and Transcriptomic Surveys for the Study of ncRNAs with a focus on tropical parasites"**, requisito para obtenção do grau de Doutora em **Bioinformática.** Abrindo a sessão, a Presidente da Comissão, **Dra. Glória Regina Franco**, após dar a conhecer aos presentes o teor das Normas Regulamentares do Trabalho Final, passou a palavra à candidata, para apresentação de seu trabalho. Seguiu-se a arguição pelos Examinadores, com a respectiva defesa da candidata. Logo após, a Comissão se reuniu, sem a presença da candidata e do público, para julgamento e expedição de resultado final. Foram atribuídas as seguintes indicações:

Prof./Pesq.	Instituição	CPF	Indicação
Dra. Glória Regina Franco	UFMG	623 387 496 34	aprovada
Dra. Juliana Lopes Rangel Fietto	UFV	88979 881649	aprovada
Dr. Fabio Passetti	FIOCRUZ/RJ	271690878-80	Aprovada
Dra. Elida Mara Leite Rabelo	UFMG	448589906-30	Aprovada
Dr. Gerald Weber	UFMG	02491500850	APROVADA

Pelas indicações, a candidata foi considerada: APROVADA

O resultado final foi comunicado publicamente à candidata pela Presidente da Comissão. Nada mais havendo a tratar, a Presidente encerrou a reunião e lavrou a presente ATA, que será assinada por todos os membros participantes da Comissão Examinadora.

Belo Horizonte, 20 de fevereiro de 2015.

Dra. Glória Regina Franco - Orientadora Gloria Regina Franco

Dra. Juliana Lopes Rangel Fietto

Dr. Fabio Passetti Fabio Passetti

Dra. Elida Mara Leite Rabelo Elida Mara Leite Rabelo

Dr. Gerald Weber Gerald Weber

Dr. Vasco Ariston de C. Azevedo
Prof. Titular e Coordenador do Programa de
Pós Graduação em Bioinformática
ICB/UFMG

Esta tese é dedicada à minha mãe,

que me deu a liberdade para sonhar

e a força para viver a realidade.

AGRADECIMENTOS

À Chefa, por ser meu exemplo de cientista, professora e de tantas outras coisas. Por me ensinar a ser melhor a cada dia em tudo que eu faço e pela nossa simbiose que vai além da pesquisa. Espero um dia poder retribuir tudo o que você fez por mim desde que essa *Gloriosa* história começou.

À minha mãe e meus avós, por terem tido tanta dedicação a mim e me ensinado tanta coisa. Tudo em mim tem um pouquinho de vocês.

À minha família, real e inventada, por ser o lugar para onde sempre quero voltar, seja lá de onde for.

Às PriMas (Priscila, Marcela, Marcela e Marina), por terem sido minha casa em todos os sentidos, desde que cheguei em Belo Horizonte.

À Pós-Graduação em Bioinformatica da UFMG, em especial à Sheila e ao Vasco, por terem ido muito além de suas obrigações para contribuir com a minha formação cientifica, acadêmica e pessoal.

À CAPES, que através dos programas Reuni e Ciência sem Fronteiras abriu novas portas para a minha formação acadêmica e cientifica.

Ao LGB, por ter sido a minha segunda casa durante os últimos anos, onde aprendi muito e fiz amigos para toda a vida. Obrigada a todos os que estão ou estiveram no Laboratório e fizeram parte desta tese. Obrigada em especial à Neuza e aos professores Andréa e Carlos Renato, por terem me ensinado tanto sobre as diversas faces da pesquisa e por todas as valiosas discussões.

Aos alunos que passaram pelos cursos de Perl e Modelagem Molecular, por terem dividido comigo a experiência de aprender a ensinar, obrigada pela paciência! Obrigada ao Edgar e ao Diego em especial por toda a ajuda.

À Ceres, à Maira, à Grazi, à Dani, ao Italo e ao André, por estarem sempre dispostos a me ajudar e pelas risadas e lições ao longo deste doutorado.

Às meninas da JeMaDa, que dividiram comigo não somente a casa, mas também a vida, que me apoiaram e me acolheram sempre.

À Família Franco, que me adotou com tanto carinho que eu não sei mais onde termina a minha família e começa a de vocês. Nem sei como agradecer.

Às irmãs que a vida me deu, Ana Laura, Yasmin e Laura, com vocês eu aprendi que alguns laços são pra sempre e que a saudade é amor com fuso horário.

À Nicole, à Catarina, ao Diego e ao Ike, por serem meus anjos da guarda!

Ao Cerne, aos ANJoS e ao Ritmo por terem feito parte da minha vida e ainda estarem comigo sempre.

ACKNOWLEDGEMENTS

To Dr. John Mattick, for listening to me, for taking the time to teach me and discuss with me and for making this year at the Garvan one of the best of my life in every way.

To Dr. Martin Smith for the patience, trust and for being part of this thesis.

To Nix for teaching me so much this year and for your friendship. I will carry both things with me forever. Thank you for the enthusiasm and support.

To Boris, Dessi, Dom, Guy, Ira, Lotta and all the Mattick Lab group for having us and for everything I have learned this year. I will never forget you.

To Luis for being, above all, the best friend I have had in a long time. For the inspiration, for the understanding and for making the most out of everything. You make me believe that time and distance are very relative.

ABSTRACT

Non-coding RNAs (ncRNAs) have been studied in great detail in the last decade, culminating in the current view of widespread molecules playing important roles in virtually all cell processes in all kingdoms of life. The tropical parasites *Schistosoma mansoni* (the causative agent of schistosomiasis) and *Trypanosoma cruzi* (the causative agent of Chagas disease) have complex life-cycles involving different hosts and environments and thus requiring a refined regulation of gene expression. In this sense, ncRNAs are known to be broad regulators of gene expression at multiple levels and more detailed surveys to identify such RNAs and the mechanisms in which they are involved are of great importance in the context of these parasites. To assess the roles and features of ncRNAs in tropical parasites, I have studied the mechanism of spliced-leader trans-splicing (SLTS) in *S. mansoni* and further expanded the study to involve all species in which this mechanism is currently known and also performed a thorough survey to identify new ncRNAs in the genome of *T. cruzi*.

The SLTS mechanism is known to be present in species of seven different phyla. The molecular characteristics of the spliced-leader (SL) RNA and the set of transcripts processed by the SLTS RNA were studied both in *S. mansoni* and further in ~150 species. Main findings include the sequence of the SL RNA exon being conserved among different species of a given phylum and more divergent when species from different phyla are compared, however the overall structure of the SL RNA is more conserved in all phyla than its sequence. Additionally, when analyzing the transcripts that receive the SL sequence, I have observed that these include genes from unrelated classes and have no clear bias to any specific function.

The genome of *T. cruzi* was scanned in search of non-annotated ncRNAs, using both sequence-based and structure-based search algorithms and different databases. As a result, I report 1,595 unique ncRNA candidates, more than 700 of these representing new findings. Interestingly, the most abundant classes of new ncRNAs include snoRNAs, RNase P RNAs and miRNAs. The first are known to be widespread in the genome of *Trypanosoma brucei* (the causative agent of the sleeping sickness) and play important roles in the modification and putative stabilization of rRNAs. RNase P RNAs were not previously reported in trypanosomatids and to date only a protein-only RNase P complex was believed to be present in such organisms. The identification of miRNAs candidates in the genome of *T. cruzi* was surprising given that the parasite does not harbour a classic RNAi machinery and therefore further analyses should be conducted to prove this finding. Lastly, RNA candidates involved with regulation of gene expression that are believed to be exclusive of prokaryotes were also found, raising questions about how their function could be performed in this eukaryotic parasite.

RESUMO

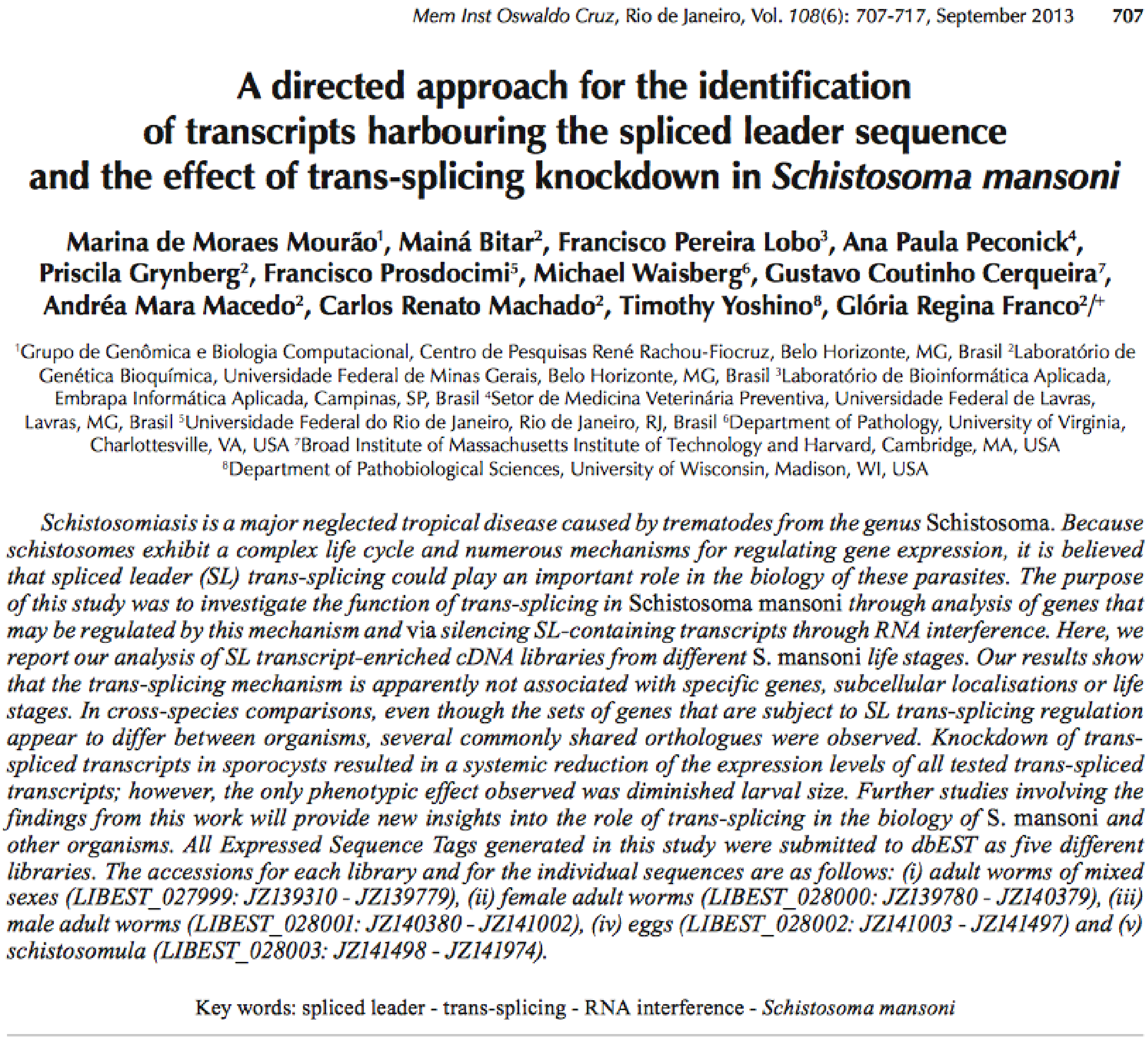
RNAs não-codificadores têm sido estudados em detalhes na última década, culminando na visão recente de que estes são moléculas ubíquas que desempenham papel importante em virtualmente todos os processos biológicos em todos os reinos da vida. Os parasitos tropicais *Schistosoma mansoni* (agente causador da esquistosomose) e *Trypanosoma cruzi* (agente causador da Doença de Chagas) têm ciclos de vida complexos envolvendo diferentes hospedeiros e ambientes, necessitando portanto de uma refinada regulação da expressão gênica. Nesse sentido, os ncRNAs são reconhecidamente reguladores da expressão gênica em vários níveis e inspeções mais detalhadas para identificar tais RNAs e os mecanismos nos quais estes estão envolvidos são de grande importância no contexto destes parasitos. Para avaliar as funções e as características dos ncRNAs em parasitos tropicais, o mecanismo de *spliced-leader trans-splicing* (SLTS) foi estudado em *S. mansoni* e, posteriormente, o estudo foi expandido para incluir todas as espécies nas quais o mecanismo é atualmente conhecido. Além disso, uma inspeção minuciosa foi realizada para identificar novos ncRNAs no genoma de *T. cruzi.*

O mecanismo de SLTS já foi descrito em espécies de sete diferentes filos. As características moleculares do RNA *spliced-leader* (SL) e o conjunto de transcritos processados pelo mecanismo de SLTS foram estudados em ambos em *S. mansoni* e posteriormente em ~150 espécies. Os principais achados incluem a sequência do éxon do SL RNA, a qual é conservada dentre diferentes espécies de um mesmo filo e mais divergentes quando espécies de diferentes filos são comparadas. No entanto, a estrutura tridimensional geral do SL RNA é mais conservada em todos os filos do que sua sequência. Adicionalmente, quando os transcritos que recebem a sequência SL foram analisados, observou-se que estes incluem genes de classes não relacionadas e não parecem estar enviesados para nenhuma função especifica.

O genoma do *T. cruzi* foi escaneado em busca de ncRNAs não anotados, utilizando tanto algoritmos de busca baseados em estrutura quanto em sequência e diferentes bases de dados. Como resultado foram encontrados 1.595 candidatos a ncRNA únicos, mais de 700 destes representando novos achados. Interessantemente, as classes mais abundantes de novos ncRNAs incluem snoRNAs, RNase P RNAs e miRNAs. Os primeiros são conhecidos por serem ubíquos no genoma de *Trypanosoma brucei* (o agente causador da tripanosomíase africana) e desempenham papéis importantes na modificação e putativa estabilização dos rRNAs. RNAs integrantes do complexo RNase P não haviam sido anteriormente descritos em tripanosomatídeos e atualmente apenas um complexo RNase constituído somente de proteínas era conhecido nestes organismos. A identificação de miRNAs candidatos no genoma de *T. cruzi* foi surpreendente, dada a ausência de uma maquinaria clássica de RNAi neste parasito e portanto análises subsequentes devem ser conduzidas para provar este resultado. Finalmente, candidatos a RNA envolvidos na regulação da expressão gênica que acredita-se ser exclusivos de procariotos também foram encontrados, levantando questionamentos sobre como sua função pode ser desempenhada neste parasito eucarioto.

SCIENTIFIC ARTICLES PUBLISHED IN THE SCOPUS OF THE THESIS

1) A directed approach for the identification of transcripts harbouring the spliced leader sequence and the effect of trans-splicing knockdown in *Schistosoma mansoni*



Mem Inst Oswaldo Cruz, Rio de Janeiro, Vol. *108*(6): 707-717, September 2013 **707**

A directed approach for the identification of transcripts harbouring the spliced leader sequence and the effect of trans-splicing knockdown in *Schistosoma mansoni*

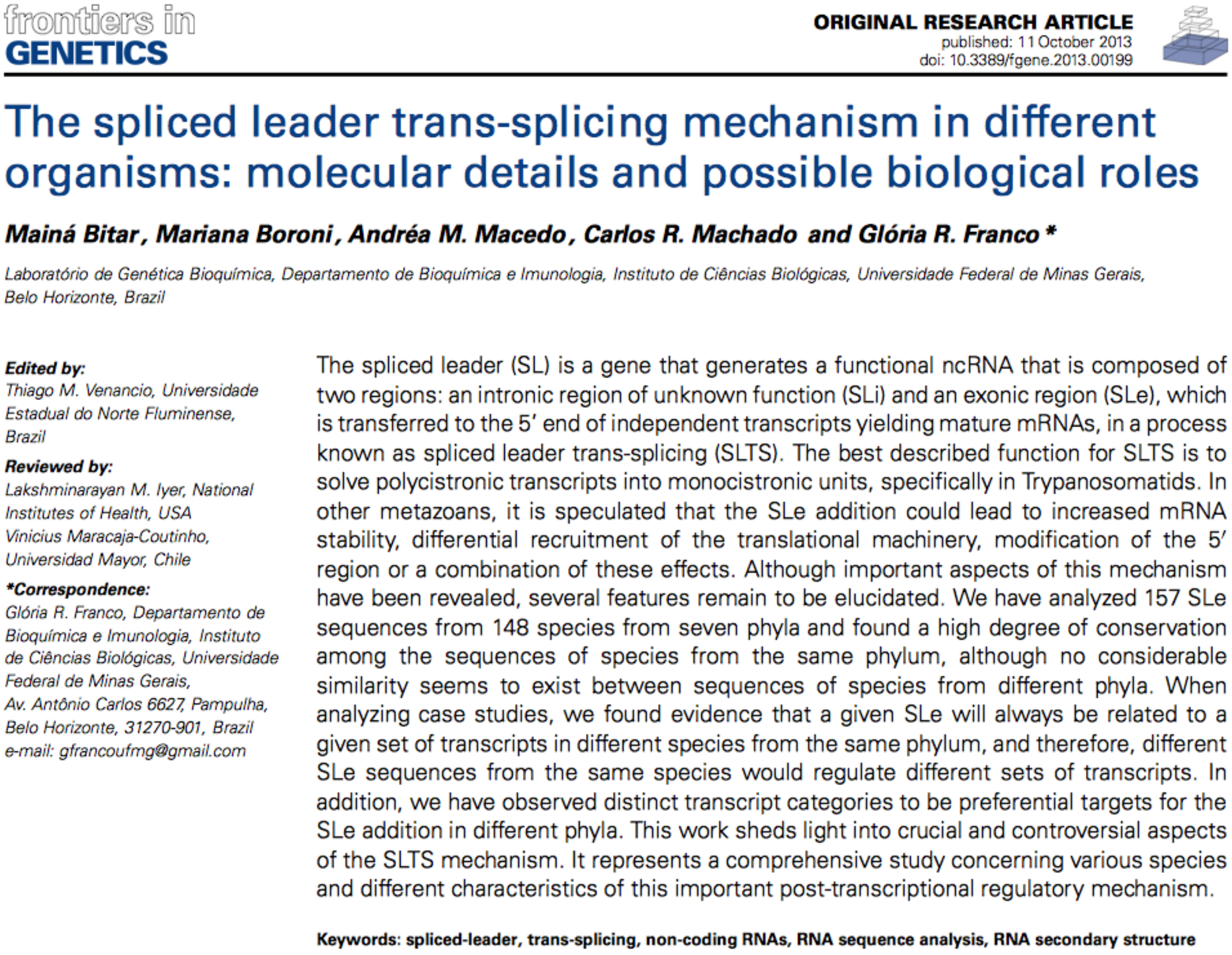
Marina de Moraes Mourão[1], Mainá Bitar[2], Francisco Pereira Lobo[3], Ana Paula Peconick[4], Priscila Grynberg[2], Francisco Prosdocimi[5], Michael Waisberg[6], Gustavo Coutinho Cerqueira[7], Andréa Mara Macedo[2], Carlos Renato Machado[2], Timothy Yoshino[8], Glória Regina Franco[2]/+

[1]Grupo de Genômica e Biologia Computacional, Centro de Pesquisas René Rachou-Fiocruz, Belo Horizonte, MG, Brasil [2]Laboratório de Genética Bioquímica, Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brasil [3]Laboratório de Bioinformática Aplicada, Embrapa Informática Aplicada, Campinas, SP, Brasil [4]Setor de Medicina Veterinária Preventiva, Universidade Federal de Lavras, Lavras, MG, Brasil [5]Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, Brasil [6]Department of Pathology, University of Virginia, Charlottesville, VA, USA [7]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, MA, USA [8]Department of Pathobiological Sciences, University of Wisconsin, Madison, WI, USA

Schistosomiasis is a major neglected tropical disease caused by trematodes from the genus Schistosoma. *Because schistosomes exhibit a complex life cycle and numerous mechanisms for regulating gene expression, it is believed that spliced leader (SL) trans-splicing could play an important role in the biology of these parasites. The purpose of this study was to investigate the function of trans-splicing in* Schistosoma mansoni *through analysis of genes that may be regulated by this mechanism and* via *silencing SL-containing transcripts through RNA interference. Here, we report our analysis of SL transcript-enriched cDNA libraries from different* S. mansoni *life stages. Our results show that the trans-splicing mechanism is apparently not associated with specific genes, subcellular localisations or life stages. In cross-species comparisons, even though the sets of genes that are subject to SL trans-splicing regulation appear to differ between organisms, several commonly shared orthologues were observed. Knockdown of trans-spliced transcripts in sporocysts resulted in a systemic reduction of the expression levels of all tested trans-spliced transcripts; however, the only phenotypic effect observed was diminished larval size. Further studies involving the findings from this work will provide new insights into the role of trans-splicing in the biology of* S. mansoni *and other organisms. All Expressed Sequence Tags generated in this study were submitted to dbEST as five different libraries. The accessions for each library and for the individual sequences are as follows: (i) adult worms of mixed sexes (LIBEST_027999: JZ139310 - JZ139779), (ii) female adult worms (LIBEST_028000: JZ139780 - JZ140379), (iii) male adult worms (LIBEST_028001: JZ140380 - JZ141002), (iv) eggs (LIBEST_028002: JZ141003 - JZ141497) and (v) schistosomula (LIBEST_028003: JZ141498 - JZ141974).*

Key words: spliced leader - trans-splicing - RNA interference - *Schistosoma mansoni*

2) The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles



frontiers in
GENETICS

ORIGINAL RESEARCH ARTICLE
published: 11 October 2013
doi: 10.3389/fgene.2013.00199

The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles

Mainá Bitar, Mariana Boroni, Andréa M. Macedo, Carlos R. Machado and Glória R. Franco *

Laboratório de Genética Bioquímica, Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Edited by:
Thiago M. Venancio, Universidade Estadual do Norte Fluminense, Brazil

Reviewed by:
Lakshminarayan M. Iyer, National Institutes of Health, USA
Vinicius Maracaja-Coutinho, Universidad Mayor, Chile

***Correspondence:**
Glória R. Franco, Departamento de Bioquímica e Imunologia, Instituto de Ciências Biológicas, Universidade Federal de Minas Gerais,
Av. Antônio Carlos 6627, Pampulha, Belo Horizonte, 31270-901, Brazil
e-mail: gfrancoufmg@gmail.com

The spliced leader (SL) is a gene that generates a functional ncRNA that is composed of two regions: an intronic region of unknown function (SLi) and an exonic region (SLe), which is transferred to the 5′ end of independent transcripts yielding mature mRNAs, in a process known as spliced leader trans-splicing (SLTS). The best described function for SLTS is to solve polycistronic transcripts into monocistronic units, specifically in Trypanosomatids. In other metazoans, it is speculated that the SLe addition could lead to increased mRNA stability, differential recruitment of the translational machinery, modification of the 5′ region or a combination of these effects. Although important aspects of this mechanism have been revealed, several features remain to be elucidated. We have analyzed 157 SLe sequences from 148 species from seven phyla and found a high degree of conservation among the sequences of species from the same phylum, although no considerable similarity seems to exist between sequences of species from different phyla. When analyzing case studies, we found evidence that a given SLe will always be related to a given set of transcripts in different species from the same phylum, and therefore, different SLe sequences from the same species would regulate different sets of transcripts. In addition, we have observed distinct transcript categories to be preferential targets for the SLe addition in different phyla. This work sheds light into crucial and controversial aspects of the SLTS mechanism. It represents a comprehensive study concerning various species and different characteristics of this important post-transcriptional regulatory mechanism.

Keywords: spliced-leader, trans-splicing, non-coding RNAs, RNA sequence analysis, RNA secondary structure

ADDITIONAL SCIENTIFIC ARTICLES PUBLISHED IN THE COURSE OF THIS PhD

1) Bitar M, Drummond MG, Costa MGS, Lobo FP, Calzavara-Silva CE, Bisch PM, Machado CR, Macedo AM, Pierce RJ, Franco GR 2013. **Modeling the zinc finger protein SmZF1 from *Schistosoma mansoni*: Insights into DNA binding and gene regulation**. Journal of Molecular Graphics and Modelling 39:29-38.

2) Calixto PH, Bitar M, Ferreira KA, Abrahão Jr O, Lages-Silva E, Franco GR, Ramirez LE, Pedrosa AL 2013. **Gene identification and comparative molecular modeling of a *Trypanosoma rangeli* major surface protease**. Journal of molecular modeling 19(8):3053-3064.

3) Badotti F, Barbosa AS, Reis ALM, doValle ÍF, Ambrósio L, Bitar M 2014. **Comparative modeling of proteins: A method for engaging students' interest in bioinformatics tools.** Biochemistry and Molecular Biology Education 42(1):68-78.

4) Bitar M, Franco G 2014. **A Basic Protein Comparative Three-dimensional Modeling Methodological WorkflowTheory and Practice.** IEEE/ACM Transactions on Computational Biology and Bioinformatics 1.

5) Machado CR, Vieira- da- Rocha JP, Mendes IC, Rajão MA, Marcello L, Bitar M, Drummond MG, Grynberg P, Oliveira DAA, Marques C, VanHouten B, McCulloch R 2014. **Nucleotide excision repair in Trypanosoma brucei: specialization of transcription- coupled repair due to multigenic transcription**. Molecular microbiology, 92(4):756-776.

6) Vieira HGS, Grynberg P, Bitar M, da Fonseca Pires S, Hilário HO, Macedo AM, Machado CR, Andrade HM, Franco GR 2014. **Proteomic Analysis of Trypanosoma cruzi Response to Ionizing Radiation Stress**. PloS one 9(5) e97526.

7) Leal Bernardes AF, Bitar M 2014. **Comparative modeling reveals the structure of Staphylococcus aureus Enterotoxin D.** NBC-Periódico Científico do Núcleo de Biociências 3(06).

8) Dias SRC, Boroni M, Rocha EA, Lüscher-Dias T, Laet-Souza D, Oliveira FMS, Bitar M, Macedo AM, Machado CR, Calliari MV, Franco, G. R. 2014. **Evaluation of the Schistosoma mansoni Y-box-binding protein (SMYB1) potential as a vaccine candidate against schistosomiasis.** Evolutionary and Genomic Microbiology 5:174.

LIST OF ABBREVIATIONS

BAC	Bacterial Artificial Chromosome
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-like Alignment Tool
CDC	Center for Disease Control
CDS	Coding Sequence
CETEA	Comitê de Ética em Experimentação Animal
CLBe	CL Brener esmeraldo-like sub-genome
CLBne	CL Brener non-esmeraldo-like sub-genome
CLBmod	CL Brener unassembled sub-genome
CM	Covariance Models
dbEST	Database of Expressed Sequence Tags
DDBJ	DNA Database of Japan
dsRNA	Double-Stranded RNA
DTU	Discrete Typing Units
EBI	European Bioinformatics Institute
ENCODE	Encyclopedia of DNA Elements
eRNA	Enhancer RNAs
EST	Expressed Sequence Tags
FIOCRUZ	Fundação Oswaldo Cruz
fRNA	Functional RNAs
GAPDH	Glyceraldehyde 3-Phosphate Dehydrogenase
GFF	General Feature Format (or Gene-Finding Format)
GO	Gene Ontology
GWAS	Genome-Wide Association Studies
IACUC	Institutional Animal Care and Use Committee
INSDC	International Nucleotide Sequence Database Collaboration
lincRNA	Long Intergenic Non-coding RNAs
lncRNAs	Long Non-Coding RNAs
MASP	Mucin-Associated Proteins
meRNA	Multiexonic RNAs
miRNA	Micro RNAs
mRNAs	Messenger RNAs
NCBI	National Center for Biotechnology Information
ncRNA	Non-Coding RNA
ORF	Open Reading Frame
PAHO	Pan-American Health Organization
PARS	RNA Associated to Promoters
piRNA	Piwi-interacting RNAs
PROPP	Proteinaceous RNase P

pseudoRNAs	RNAs Transcribed from Pseudogenes
qRT-PCR	Quantitative Reverse Transcription Polymerase Chain Reaction
rasiRNA	Small Interfering RNA derived from a Repetitive Element Transcript
RFAM	RNA Families Database
RMST	Rhabdomyosarcoma 2 Associated Transcript
RNA-Seq	High Throughput RNA Sequencing
RNAi	RNA interference
RT-PCR	Reverse Transcription Polymerase Chain Reaction
SCFG	Stochastic Context-Free Grammar
scaRNAs	Small Cajal body-specific RNAs
scRNAs	Small Cytoplasmic RNAs
SL	Spliced Leader RNA
SLA1	Spliced Leader Associated RNA
SLTS	Spliced Leader Trans-Splicing
SLe	Spliced Leader Exon
SLi	Spliced Leader Intron
SNPs	Single Nucleotide Polymorphisms
snRNA	Small Nuclear RNA
snoRNA	Small Nucleolar RNAs
spliRNAs	Splicing Site RNAs
SRA	Sequence Read Archive
SRP	Signal Recognition Particle
TIGR	The Institute for Genomic Research
tiRNA	Tiny RNAs
TMG	Trimethylguanosine
tmRNAs	Transfer Messenger RNAs
tRNA-Sec	Selenocysteine-Carrying tRNA
trypRNAs	Trypanosomatids ncRNAs
UbCRBP	Ubiquinol-Cytochrome C Reductase Complex Ubiquinol Binding Protein
UFMG	Universidade Federal de Minas Gerais
uORF	Upstream Open Reading Frame
uRNAs	U Spliceosomal RNAs
USA	United States of America
UTR	Untranslated Region
VSG	Varying Surface Glycoprotein
WGS	Whole Genome Shotgun
WHA	World Health Assembly
WHO	World Health Organization

LIST OF FIGURES

CHAPTER 1

ANALYSES OF THE SPLICED-LEADER MEDIATED TRANS-SPLICING IN *SCHISTOSOMA MANSONI* AND OTHER ORGANISMS

CHAPTER 2

ANNOTATION OF NEW NCRNAS IN THE GENOME OF *TRYPANOSOMA CRUZI* CL BRENER STRAIN

LIST OF TABLES

CHAPTER 1

ANALYSES OF THE SPLICED-LEADER MEDIATED TRANS-SPLICING IN *SCHISTOSOMA MANSONI* AND OTHER ORGANISMS

CHAPTER 2

ANNOTATION OF NEW NCRNAS IN THE GENOME OF *TRYPANOSOMA CRUZI* CL BRENER STRAIN

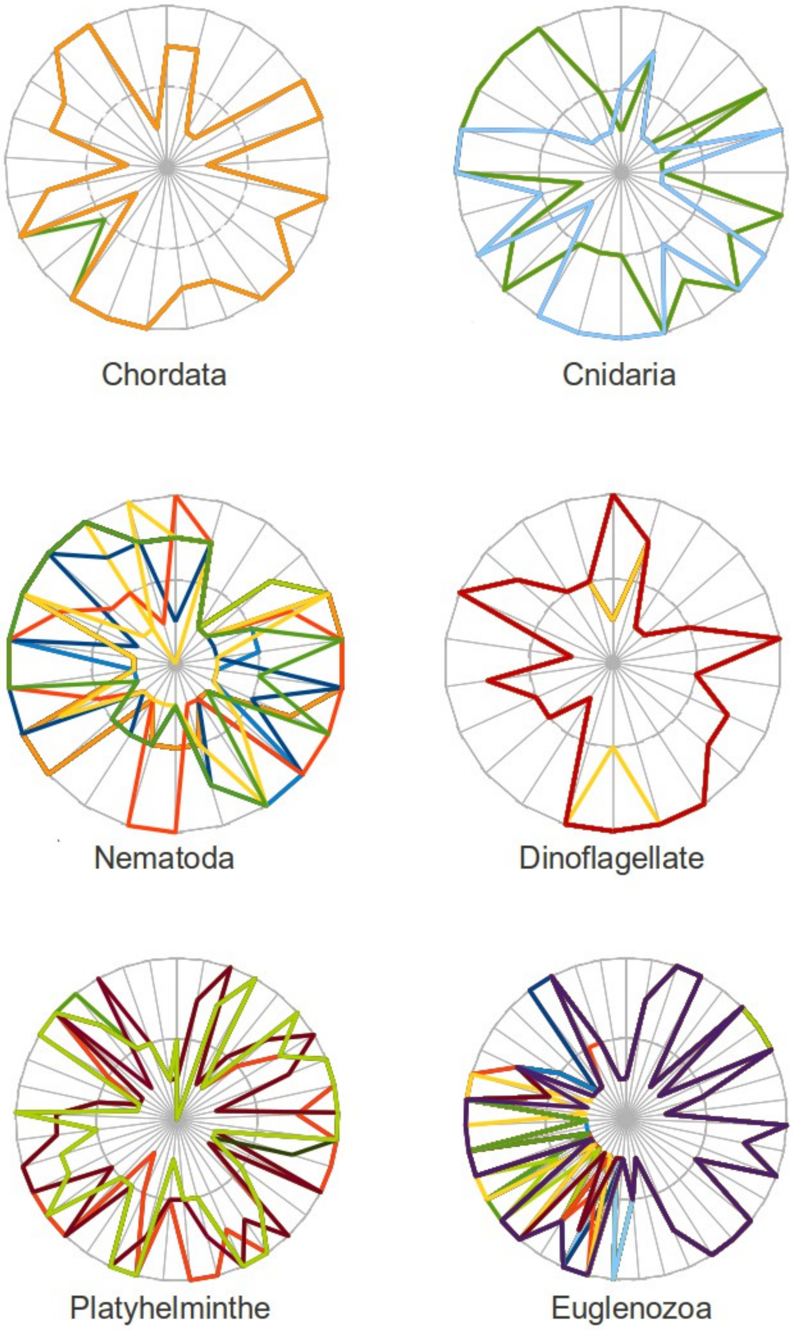
INDEX

CHAPTER 2

ANNOTATION OF NEW NCRNAS IN THE GENOME OF *TRYPANOSOMA CRUZI* CL BRENER STRAIN

CHAPTER 1

ANALYSES OF THE SPLICED-LEADER MEDIATED TRANS-SPLICING IN *SCHISTOSOMA MANSONI* AND OTHER ORGANISMS



The conservation of the spliced-leader exonic sequence in each of the seven different phyla that compose the dataset presented here.

I. INTRODUCTION

A. SCHISTOSOMA MANSONI: A CAUSATIVE AGENT OF SCHISTOSOMOSIS

Schistosomiasis is the second most important neglected tropical disease caused by species of the digenetic flatworm *Schistosoma*. In fact, according to the World Health Organization (WHO) recent reports [WHO weekly epidemiological record N° 8, 2013; WHO Fact Sheet N° 115, 2014], the number of treated individuals has increased yearly from 2006, when 12 million were treated to 2012 when over 40 million required immediate treatment and nearly 250 million required preventive treatment. This disease is responsible for 280 thousand deaths per year, with the resulting morbidity compromising local economies and child development [Fenwick & Webster, 2006; Hotez *et al*., 2008; King *et al*., 2006; Steinmann *et al*., 2006]. The transmission of schistosomiasis has been documented in 78 countries, with the great majority of cases occurring in the African continent alone [WHO Fact Sheet N° 115, 2014] (Figure 1). Of special epidemiological importance is the State of Minas Gerais, in Brazil, where recent reports pointed out that over 60% of the municipalities have reported active transmission of the parasite and over 10 million people are at risk of infection [Drummond *et al*., 2010] (detailed in Figure 1). Although the current drug of choice, praziquantel, was essential to reduce morbidity and mortality due to schistosomiasis, the disease remains endemic in several developing countries, including Brazil. The emergence of parasite resistant strains has been reported, raising concerns about the long term effectiveness of this worldwide available drug [Doenhoff *et al*., 2002; Hotez *et al.,* 2010]. Therefore, the development of new drugs and additional control measures are essential to halt schistosomiasis dissemination. In this sense, the study of parasite-specific features are of great interest, as these could subside the development of new treatment strategies targeting the parasite unique biology.

The three main species of parasites causing schistosomiasis are *Schistosoma* *mansoni*, *S. japonicum* and *S. hematobium*, although a second group of species may be of local relevance in specific areas and a third group can cause cercarial dermatitis in humans [Center for Disease Control and Prevention (CDC) website]. Among the mentioned species, *S. mansoni*, my object of study, is mainly found in Africa and Brazil and represents the major cause of intestinal schistosomiasis.

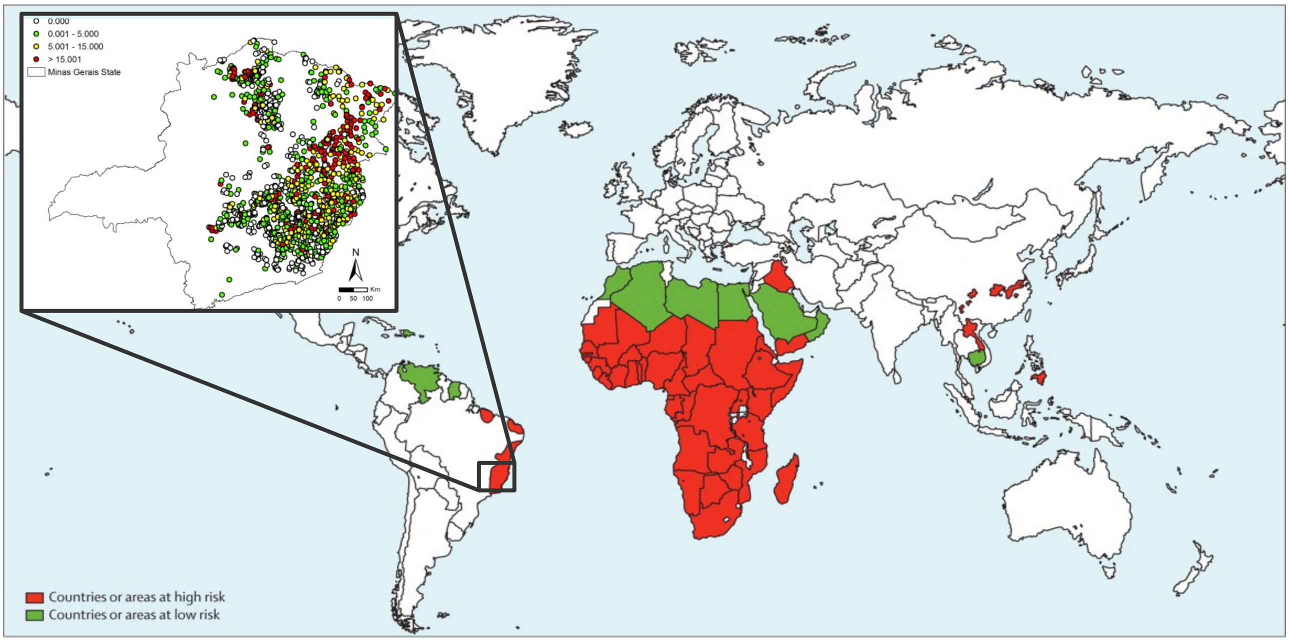


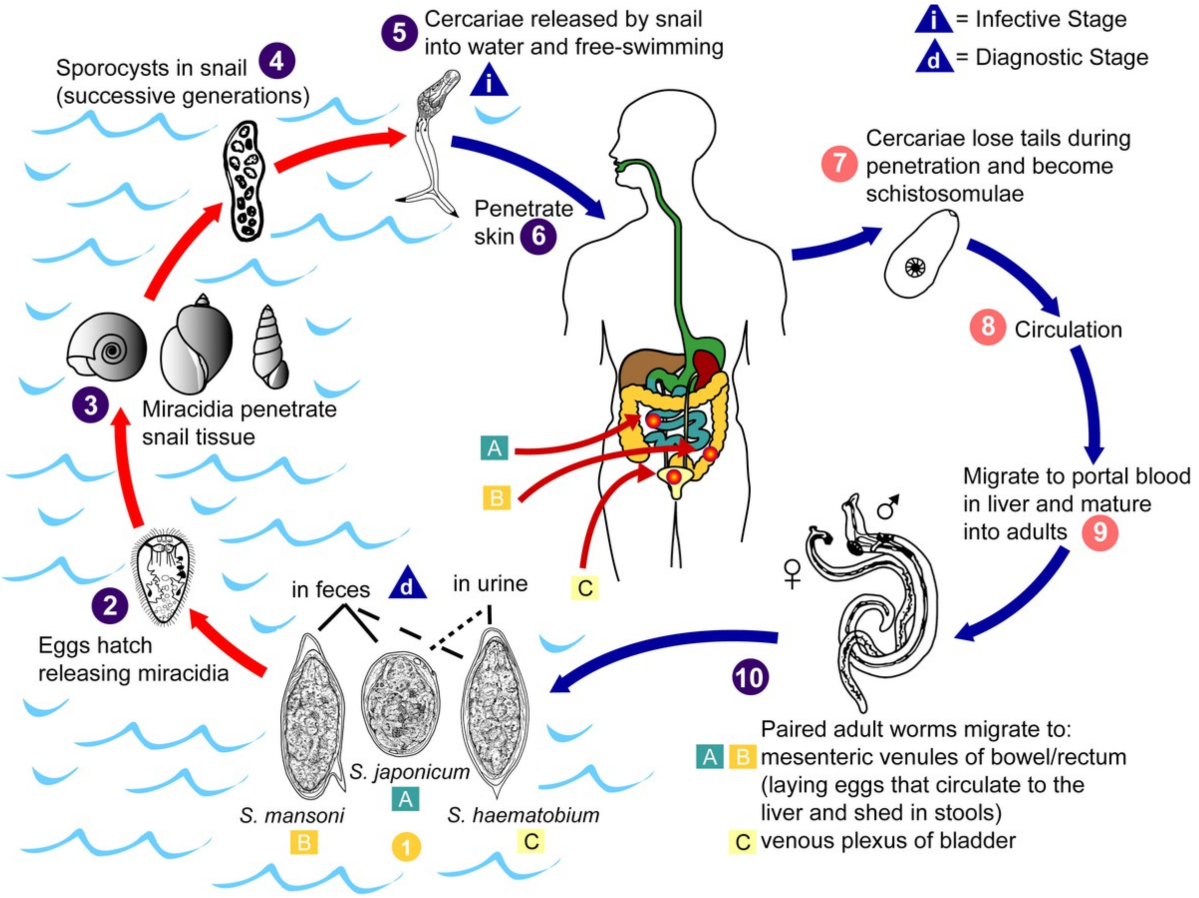
Figure 1: Spatial distribution of schistosomiasis in Minas Gerais State and worldwide. The map presents areas at risk of schistosomiasis [reproduced from Gray *et al*., 2010] and a detailed view of the spatial distribution of the disease in the Minas Gerais State in Brazil [modified from Guimaraes *et al*., 2013].

The parasite has a complex life-cycle involving a snail intermediate host, and a mammalian definitive host [Pessôa & Martin, 1982] (Figure 2). The disease cycle begins in water reservoirs, where eggs eliminated in the feces of infected individuals hatch, releasing miracidia. These larval forms of parasites swim and infect specific snail hosts, by penetrating their soft tissue. While inside the intermediate host, parasites differentiate to sporocysts (unicellular forms with asexual reproduction) and reach the cercarial stage, which is infective to humans. Cercariae released from the snail are able to survive and swim for up to two days before eventually penetrate human skin in an active mechanism culminating with the lost of their tails, thus originating schistosomulae. In the human definitive host schistosomulae migrate through different tissues before stablishing in veins. The exact location of adult worms seems to vary according to species-specific tropism, although all species can be eventually found in non-preferred locations. *S. mansoni* worms tend to be located in superior mesenteric veins, which drain the intestine. Regarding adult worms there is a marked sexual dimorphism, with females typically measuring 7 to 20 millimetres and males being slightly shorter. The thinner and longer female worms are constantly located inside a ventral cleft present in male adult worms, named the gynecophoric canal and this close interaction drives female growth and maturation. The constant copula lasts for approximately 5 years and yields on average 300 eggs per couple per day, which are mainly laid in the region of the portal veins. While part of the eggs may remain in the infected tissue causing inflammation and scarring, the rest of the eggs further progresses to the intestine to be eliminated in feces and thus recommence the cycle.

From the description of the disease cycle, it is clear that infection of humans by *S. mansoni* occurs when there is contact with contaminated freshwater. As a single couple of adult worms can lay up to 300 eggs per day it is not surprising that the main symptoms of schistosomiasis are not derived from the infection itself but from the host organism response to the presence of high quantities of parasite eggs. Most individuals are asymptomatic when first infected and further develop a rash or itchy skin within days. After one or two months the infected individual can present symptoms including fever, chills, cough and muscle aches.

Without treatment the disease persists for years leading to abdominal pain, hepatomegalia and the presence of blood in feces. More rarely, eggs can migrate to the brain or spinal cord, causing seizures, paralysis and inflammation of spinal cords. Diagnosis can come from the microscopic examination of feces and the identification of *S.*

mansoni eggs or (since egg elimination may be intermittent and in low amounts) by serologic testing for antischistosomal antibodies in individuals that were not previously treated.



i = Infective Stage
d = Diagnostic Stage
5 Cercariae released by snail into water and free-swimming
i
4 Sporocysts in snail (successive generations)
Penetrate skin 6
7 Cercariae lose tails during penetration and become schistosomulae
8 Circulation
3 Miracidia penetrate snail tissue
A
B
C
Migrate to portal blood in liver and mature into adults 9
♂
♀
2 Eggs hatch releasing miracidia
in feces
d
in urine
10
Paired adult worms migrate to:
A B mesenteric venules of bowel/rectum (laying eggs that circulate to the liver and shed in stools)
C venous plexus of bladder
S. japonicum
A
S. mansoni
B
1
S. haematobium
C

Figure 2: The schistosomiasis cycle and *S. mansoni* life-cycle stages. Figure reproduced from the CDC (http://www.cdc.gov/dpdx/schistosomiasis).

B. THE SPLICED LEADER TRANS-SPLICING MECHANISM

Since RNA splicing was first observed in the late 1980s [Green, 1986; Breitbart *et al.*, 1987], several subtypes of this mechanism have been described (Figure 3A). Of particular interest for us is the non-coding RNA (ncRNA) mediated trans-splicing mechanism named spliced leader trans-splicing (SLTS), in which the exonic portion of a specialized ncRNA transcript, the spliced leader (SL), is transferred to the 5' end of unrelated transcripts yielding mature messenger RNAs [Boothroyd & Cross, 1982 and reviewed by Liang *et al*., 2003] (Figure 3B). SL RNAs are small non-coding RNAs of 40 to 140 nucleotides in length, carrying a donor splice site and a hyper-modified cap [Nilsen, 1993]. The donor splice site divides the SL RNA in two segments: the previously mentioned 5' leader sequence (hereby SLe) and an intron-like sequence at the 3' end (hereby SLi). All organisms that undergo trans-splicing have one or more SL RNA, which are products of tandemly repeated small intronless genes transcribed by DNA polymerase II [Hastings, 2005]. Despite the lack of sequence similarity, SL RNAs from different organisms exhibit an impressive secondary structure similarity to small nuclear RNAs (snRNA), which are components of the spliceosome and actively participate in all splicing mechanisms [Hastings, 2005].

As first observed in trypanosomatids, its best described function is to resolve polycistronic transcripts into monocistronic units [Sather & Agabian, 1985 and reviewed by Preußer *et al*., 2012]. Subsequently, the SLTS was demonstrated to occur in other euglenozoans [Tessier *et al*., 1991] and several organisms, such as rotifers [Pouchkina-Stantcheva & Tunnacliffe, 2005], cnidarians [Stover & Steele, 2001], chordata [Vanderberghe *et al*., 2001], nematoda [Krause & Hirsh, 1987], platyhelminthes [Rajkovic *et al*., 1990] and dinoflagellates [Lidie & van Dolah, 2007]. Different biological roles have been proposed for this mechanism, such as: (i) enhancing translation of trans-spliced transcripts by providing a specialized (hypermethylated) 5' cap structure for trans-spliced transcripts, (ii) stabilizing the messenger RNAs (mRNA) and (iii) removing regulatory elements from the outron (the pre-mRNA region from the 5' end to the trans-splice site) in a process called 5' untranslated region (UTR) sanitization [Hastings, 2005; Matsumoto *et al*., 2010].

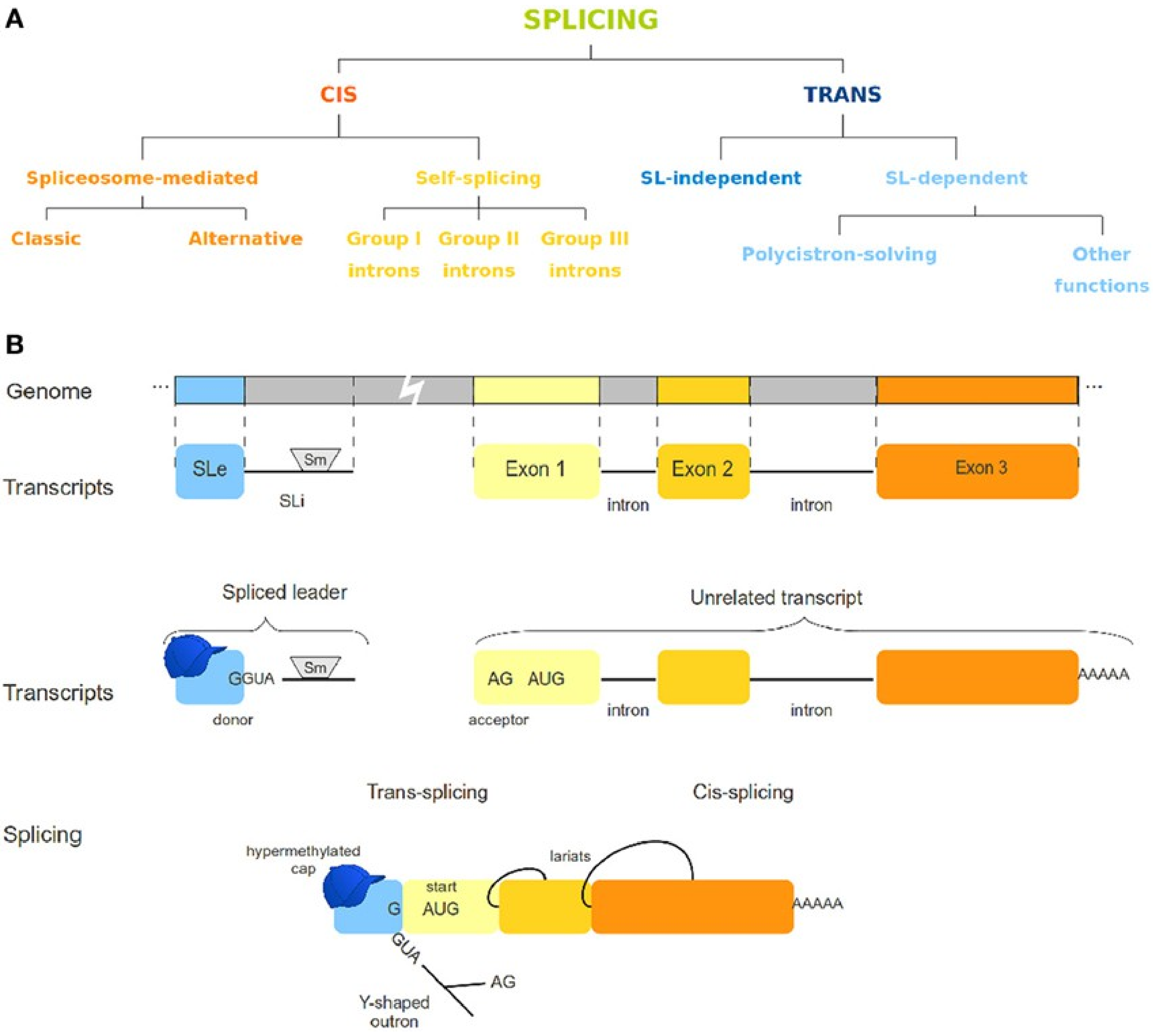


Figure 3: Different splicing mechanisms. **(A)** Variations of cis- and trans-splicing processes. **(B)** The spliced leader trans-splicing mechanism. This panel depicts the overall SLTS mechanism, in which an invariant exon (the SLe) is added to an unrelated, independently produced transcript, yielding a mature mRNA. Cis-splicing is also represented for comparison. The SLe molecule contains a hypermodified cap, whereas the SLi usually harbours an Sm-protein binding site.

In a recent work Nilsson and collaborators [Nilsson *et al*., 2010] discuss possible functions for the SLTS mechanism in *Trypanosoma brucei*. According to the authors, the differential insertion of the SLe sequence in alternative acceptor sites of genes could lead to: (i) translation blockage when the SLe insertion is downstream the protein start codon; (ii) alteration of protein subcellular location, when signalling sequences are eliminated by SLe insertion; (iii) translation of alternative open reading frames (ORF) and (iv) inclusion or exclusion of upstream open reading frames (uORF) or other regulatory elements from 5' end of transcripts. It has also been speculated that SLe addition could lead to increased mRNA stability, differential recruitment of the translational machinery, modification of the 5' UTR or a combination of those effects [reviewed by Hastings, 2005; Lasda & Blumenthal, 2011; Stover *et al*., 2006].

S. mansoni has numerous and complex transcriptional and post-transcriptional gene regulatory mechanisms to maintain its complex life-cycle. Because of the prominence of the SL sequence on a number of *S. mansoni* mRNAs, it is presumed that SLTS represents an important form of post-transcriptional regulation, and could be one of the potential targets for impairing *S. mansoni* development [Davis *et al*., 1995]. Over two decades ago, a first assessment of the parasite transcriptome has predicted that around 30% of the mRNAs would undergo trans-splicing [Rajkovic *et al*., 1990]. More recently, using undirected high throughput RNA sequencing (RNA-Seq) data, a new assessment suggested that only approximately 10% of *S. mansoni* protein-coding transcripts would be trans-spliced [Berriman *et al*., 2012]. Additionally, our group has generated directed RNA-Seq data (using the SL sequence as sequencing primer) for *S. mansoni* to analyze the addition of the SL sequence to the parasite transcripts and unpublished results indicate that over 60% of the protein-coding genes undergo trans-splicing. Although no consensus was reached, this study place the SLTS mechanism in an important position as a post-transcriptional regulatory mechanism. Interestingly, in platyhelminthes such as *S. mansoni*, the SLe sequence ends in an AUG triplet, raising the possibility that the addition of the SLe to transcripts could itself grant its translation and therefore open new reading frames for protein production. This hypothesis is currently under investigation by our group and, if proved to be true, will represent an even more crucial role for the SLTS mechanism in species of this phylum.

Although important aspects of the SLTS mechanism in this parasite were elucidated and are presented herein, several other questions regarding this mechanism in *S. mansoni* and other organisms remained unanswered. These questions surround the existence of conserved motifs within SL sequences from different species, the possible emergence of SLTS in more complex taxa as plants and mammals, the role of the SLTS mechanism in different organisms, the peculiarities of the set of transcripts under control of a given SLe and the structural aspects of the SL molecule.

Searching for answers to such questions, I have turned to the molecular constitution of the SL RNA itself and performed analyses on a great number of SL sequences. All of the answers lead to one final question, which has been debated in the literature for a long time: “what is the origin of the SLTS mechanism and how has it evolved?” This section of the present work is devoted to the proposition of hypotheses based on observed features of this mechanism that may guide the composition of a future final answer to this question.

As for now, although there is no conclusive statement, there are important observations that can help to characterize the mechanism, its biological role, phylogenetic features and molecular details. In the course of this study, I have first assessed the set of trans-spliced transcripts in *S. mansoni* and further compiled a comprehensive dataset of SL sequences from several species of various phyla. This was the starting point for several computational analyses that explored different aspects of the SLTS mechanism.

II. AIMS

A. ANALYSES OF THE SLTS MECHANISM IN SCHISTOSOMA MANSONI

Main Goal

To better describe the mechanism of SLTS in *S. mansoni* and to assess the set of transcripts regulated by such mechanism.

Specific Goals

- To identify transcripts bearing the SL sequence in the parasite transcriptome by assessing a SL-enriched expressed sequence tags (EST) library (which was generated by Dr. Marina de Moraes Mourão).
- To assess the characteristics of SL-containing transcripts using Bioinformatics tools in order to generate a molecular profile of trans-spliced transcripts.
- To functionally characterize these transcripts and further search for gene classes that may be enriched in the set of trans-spliced transcripts.
- To assess the biological role of the SLTS mechanism in *S. mansoni* by performing the knockdown of the SL RNA and observing the related phenotypical effects.

B. ASSESSMENT OF THE SL RNA SEQUENCES AND SL-CONTAINING TRANSCRIPTS OF DIFFERENT ORGANISMS

Main Goal

To compare the SL sequences and SL-containing transcripts of multiple species in order to propose general characteristics for the SLTS mechanism.

Specific Goals

- To generate an internal database of SL RNA, SLe and SLi sequences from multiple species and multiple phyla using publicly available sequence data.
- To assess the similarities and differences between the SL sequences of different species and species from different phyla using sequence comparison tools searching for general characteristics of the SLTS mechanism.
- To analyze and compare the sets of SL-containing transcripts from all species in the internal database using sequence alignment and search tools.

III. MATERIALS AND METHODS

A. ANALYSES OF THE SLTS MECHANISM IN SCHISTOSOMA MANSONI

A.1 *In vitro* and *in vivo* studies (performed by Dr. Marina de Moraes Mourão)

Biological Samples

The *S. mansoni* life-cycle was maintained at Centro de Pesquisas René Rachou, (Fundação Oswaldo Cruz, FIOCRUZ) and at the Department of Parasitology (Universidade Federal de Minas Gerais, UFMG). The LE strain of *S. mansoni* was maintained in the snail intermediate host *Biomphalaria glabrata* (Barreiro de Cima strain). Outbred SWISS Webster mice were housed conventionally and received lab mouse chow and water *ad libitum*. Adult worms were obtained by portal perfusion of 5 weeks infected mice as previously described [Smithers & Terry, 1965], washed with cold saline solution, carefully separated by gender under the microscope and immediately frozen at -80°C until further processing. Seven-day cultured Schistosomula [Basch, 1981] were provided by

Centro de Pesquisas René Rachou and eggs were recovered from the intestinal homogenates of 48-day infected mice. The collected tissues and eggs were filtered to remove debris, and allowed to settle with the resulting pellet washed with 1.7% saline and frozen at -80°C for further processing.

RNA isolation, reverse transcription and SL transcript enrichment

RNA from *S. mansoni* male and female adult worms and cultured schistosomulae were extracted using the RNAgents kit (Promega, Madison) and total RNA from eggs was extracted using TRIzol Reagent (Life Technologies, Carlbad), according to manufacturer protocol. Direct isolation of poly(A)+ mRNA from adult worms was performed using Dynabeads Oligo (dT) 25 magnetic beads (Dynal - Life Technologies). Briefly, following the extraction step, the beads containing bound mRNA were re-suspended in 20 µL of 1x reverse transcriptase buffer (Life technologies, Carlsbad) and used directly in reverse transcription polymerase chain reaction (RT-PCR) reactions. The cDNA synthesis for adult worms was performed by directly adding 1 µL of beads containing mRNA to the reaction, while for all other samples it was carried out using SuperScript II Reverse Transcriptase (Life Technologies, Carlsbad), according to procedures outlined by the manufacturer. The oligo dT anchored primer used for cDNA first strand synthesis (5' CGGTATTTCAGTCGGTGTTCAAACCT19V 3' - V=A, G, C) contains a 5' tail that was further used in a PCR reaction for selective amplification of trans-spliced transcripts. The strategy followed to enrich cDNA libraries in trans-spliced transcripts included a step-down PCR step that used a 5' tailed oligo dT [Brehm *et al*., 2000] and the *S. mansoni* SL sequence [Davis *et al*., 1995].

Size selection of cDNAs

Fragment size selection was performed to prevent over-representation of small PCR products using either of this two methodologies: (i) PCR amplicon separation by electrophoresis in a 1% agarose gel and further isolation using the enzyme β-agarase according to Franco *et al*. (1995); (ii) precipitation with 15% polyethyleneglycol 8000 to obtain fragments of 400 base pairs or longer**.** Purified amplicons were cloned into pGEM and pCR2.1 vectors, using the T-easy System Vector kit (Promega, Madison) and TA cloning kit (Life Technologies, Carlsbad), respectively, according to manufacturer specifications. The recombinant plasmids were used to transform competent *Escherichia coli*. In order to select clones with large inserts and test the library quality, recombinant bacteria clones were subjected to colony PCR using M13 forward and reverse primers. Amplification and insert size estimates were confirmed by electrophoresis in 1% agarose gel. Selected clones were grown overnight in 2X YT medium (16g of bacto tryptone, 10g bacto yeast extract, 5g NaCl per liter, pH 7.0). Recombinant plasmids were purified by standard protocol. Template preparation and DNA sequencing reactions were performed using the DYEnamic ET dye terminator cycle sequencing (GE Healthcare), following manufacturer protocol and the MegaBACE 1000 capillary sequencer (GE Healthcare). All steps for this methodological subsection are summarized in the figure below (Figure 4).

Schistosoma mansoni *siRNA treatment and phenotypic screening*

S. mansoni NMRI strain eggs were obtained from 7-8 weeks infected mouse livers and transformation was carried out as previously described [Yoshino & Laursen, 1995]. Larvae were counted and distributed into tissue culture plates (Costar, Corning Incorporated, NY) at concentrations of ~6000 (quantitative RT-PCR, qRT-PCR) and ~500 miracidia/well [Mourão *et al*., 2009a].

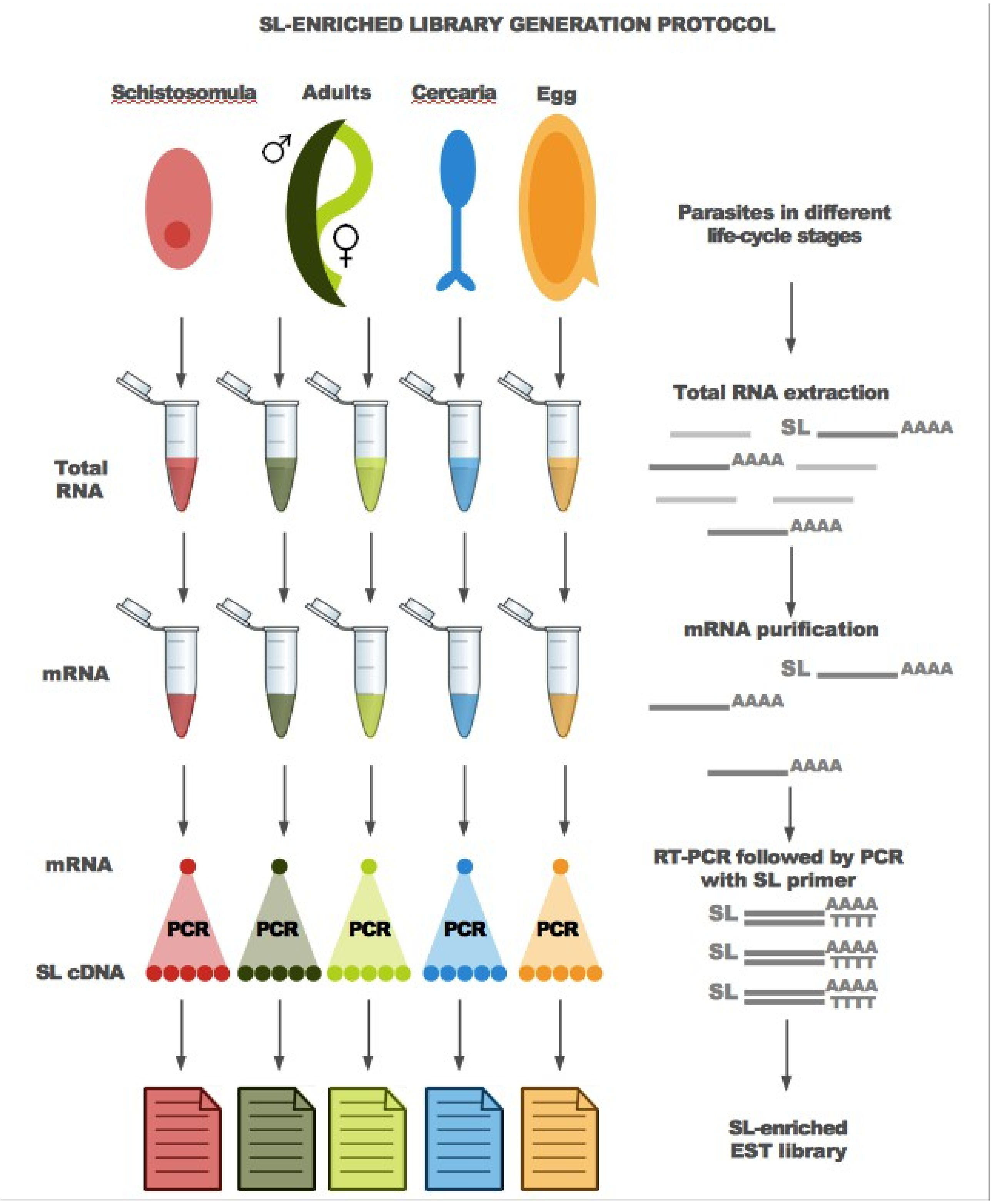


Figure 4: A schematic view of the protocol for the generation of SL-enriched EST libraries. This graphical representation summarizes the protocol from RNA extraction to SL-enriched EST library generation.

All RNA interference (RNAi) experiments involved at least two technical replicates of miracidial treatment and control groups with experiments repeated on 3 independent larval cultures. Parasites were exposed in culture on day 0 to SL-sequence small interference (si)RNAs (treatment), an irrelevant siRNA (control I, decoy) and to medium alone (control II). Cultured larvae were assessed for knockdown effect after 7 days of treatment [Mourão *et al*., 2009a]. The siRNA sequences were designed using BLOCK-iT™ RNAi Designer tools and synthesized using the StealthTM proprietary modification (Life Technologies). The decoy control was designed with similar GC content and length in comparison to the target siRNA. Cultured *S. mansoni* larvae were plated in culture plates (Costar) using approximately ~500 miracidia/well containing either 200 nM of SL-siRNA (experiment group), decoy sequence (control I) siRNA diluted in 200 µL CBSS or medium lacking siRNA (control II). Cultures were maintained at 26°C for 7 days when sporocysts were monitored for the following phenotypes: failure or delay in transformation, loss of motility, tegumental lysis and granulation (lethality) and changes in larval growth (Figure 5). Parasite viability and morphological changes were monitored daily as described elsewhere [Mourão *et al*., 2009a]. Length measurements from captured images were analyzed by Metamorph software version 7.0 (Meta Imaging series, Molecular Devices, Sunnyvale, CA). Larval growth datasets for each experimental replicate were statistically analyzed using the Mann-Whitney U-test (Wilcoxon-Sum of Ranks test) with significance of p≤0.05. All treatments were performed in triplicate wells, and the experiment was independently replicated 3 times on miracidia isolated from different batches of infected mouse livers.

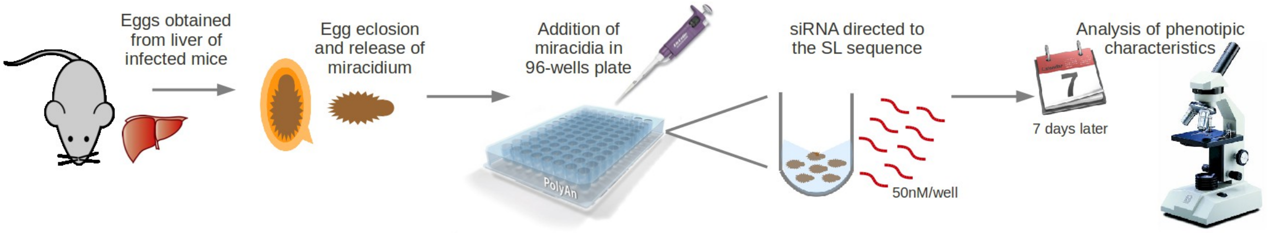

Eggs obtained from liver of infected mice
Egg eclosion and release of miracidium
Addition of miracidia in 96-wells plate
PolyAn
siRNA directed to the SL sequence
50nM/well
Analysis of phenotipic characteristics
7
7 days later

Figure 5: Analysis of the sporocyst phenotypic characteristics after knockdown of SL-containing transcripts.

Effect of double-stranded (ds)RNA treatment on larval gene expression and qRT-PCR analysis

Quantitative RT-PCR was used to determine steady-state transcript levels in specific ds-siRNA-treated sporocysts. In these experiments ~6000 miracidia were distributed into a 48-well plate (Costar), treated with 200 nM siRNA diluted in CBSS (500 µL/well) and maintained at 26°C for 7 days and washed prior to RNA extraction using the TRIzol Reagent (Life Technologies) [Mourão *et al*., 2009a]. Isolated total RNA was resuspended in diethylpyrocarbonate-treated water and subjected to DNAse treatment using the Turbo DNA-Free kit (Ambion, Austin, TX) to eliminate contaminating genomic DNA. RNA samples were quantified and their purity assessed on a Nanodrop Spectometer ND-1000 (NanoDrop Technologies, Inc., Wilmington, DE). In order to evaluate transcript levels between SL-siRNA-treated sporocysts and control treatments (decoy-siRNA), Dr. Marina de Moraes Mourão performed quantitative PCR analysis with 0.5 µg of total RNA derived from at least three different extractions which were used for cDNA synthesis with the Superscript III cDNA Synthesis kit (Invitrogen) following the manufacturer protocol. The qRT-PCR reaction mixtures consisted of 2.5 µL of cDNA, 12.5 µL of Sybr Green PCR Master Mix (Applied Biosystems, Foster City, CA), 10 µL of 600 or 900 nM primers, determined after primer concentration optimization following MIQE guidelines [Bustin *et al*., 2009] to 96-Well Optical Reaction Plates (ABI PRISM, Applied Biosystems) and the reactions were performed on an AB7500 Real Time PCR System (Applied Biosystems). The qRT-PCR validations were performed using the SL forward primer, 5'-GTCACGGTTTTACTCTTGT-3' and gene-specific reverse primers designed for: (i) five previously known trans-spliced transcripts [Davis *et al*., 1995]; (ii) *S. mansoni* α-tubulin, used as endogenous normalization control in all tested samples and (iii) *S. mansoni* genes of three known non-trans-spliced transcripts used as negative controls. Each qRT-PCR run was performed with two internal controls assessing both, potential genomic DNA contaminations (absence of reverse transcriptase) and purity of the reagents used (no complementary DNA, cDNA). For each specific set of primers, all individual treatments (including specificity controls) were run in three technical replicates. Each experiment was repeated three times as independent biological replicates and analyzed by the ΔΔCt method [Livak & Schmittgen, 2001], using the Mann-Whitney U-test with significance set at $p \leq 0.05$.

Ethical statement

The experimental protocols described herein were reviewed and approved by the Comitê de Ética em Experimentação Animal (CETEA) of UFMG (permission number 185/2006). The RNAi experiments involving mice were pre-approved by the Institutional Animal Care and Use Committee (IACUC) of the University of Wisconsin-Madison, where experiments were conducted, under assurance number A3368-01.

A.2 *In silico* analysis

Sequence processing

The output files generated from sequencing reactions (SL-enriched EST library) and from publicly available *S. mansoni* data from the EST database (dbEST) [Boguski *et al*., 1993], were submitted to a Bioinformatics pipeline, which contained algorithms for base-calling, poly-A and vector decontamination, motif search, similarity-based characterization, gene ontology (GO) assignment and manually curated annotation and analysis (Figure 6). Unless otherwise stated, all sequence and exon content information were retrieved from the public database SchistoDB [Zerlotini *et al*., 2009].

Phred [Ewing *et al*., 1998] was used for base-calling and only sequences with quality scores higher than 10 were accepted. Phred was requested to generate PHD files, sequence files and quality files for the inputs (-pd, -sd and -qd options, respectively). Bash commands were used to filter results and exclude sequences shorter than 150 nucleotides and to generate a multi-fasta file from all resulting sequences. This multi-fasta file was then used throughout the subsequent analysis.

Command Line 1: Base-calling and quality filter using Phred.

```
phred -id InputFolder -pd InputFolder/PHDFiles -sd InputFolder/SeqFiles -qd InputFolder/QualFiles -trim_alt “” -trim_cutoff 0.10 -trim_fasta
```

After Phred processing, the sequences from the SL-enriched EST library were subjected to the SeqClean program (https://sourceforge.net/projects/seqclean), from The Institute for Genomic Research (TIGR), which consists of a script capable of analysing EST data. SeqClean was used to trim ESTs based on informational content, length and vector contamination. The UniVec (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html) vector database was used with two additional vectors (pGEM and pCR2.1) to scan the sequences with SeqClean for vector contamination. Poly-A tails and vector adaptors were also removed in this step as a default feature of SeqClean. Only ESTs longer than 50 nucleotides (shorter sequences removed using a set of BASH commands after vector cleaning) were further evaluated. *S. mansoni* dbEST data were also submitted to the SeqClean cleaning pipeline and analyzed together with SL-enriched EST library data.

Command Line 2: Sequence cleaning using SeqClean.

```
seqclean MultiFasta.fasta -v UnivecFolder -o MultiFasta.sc
```

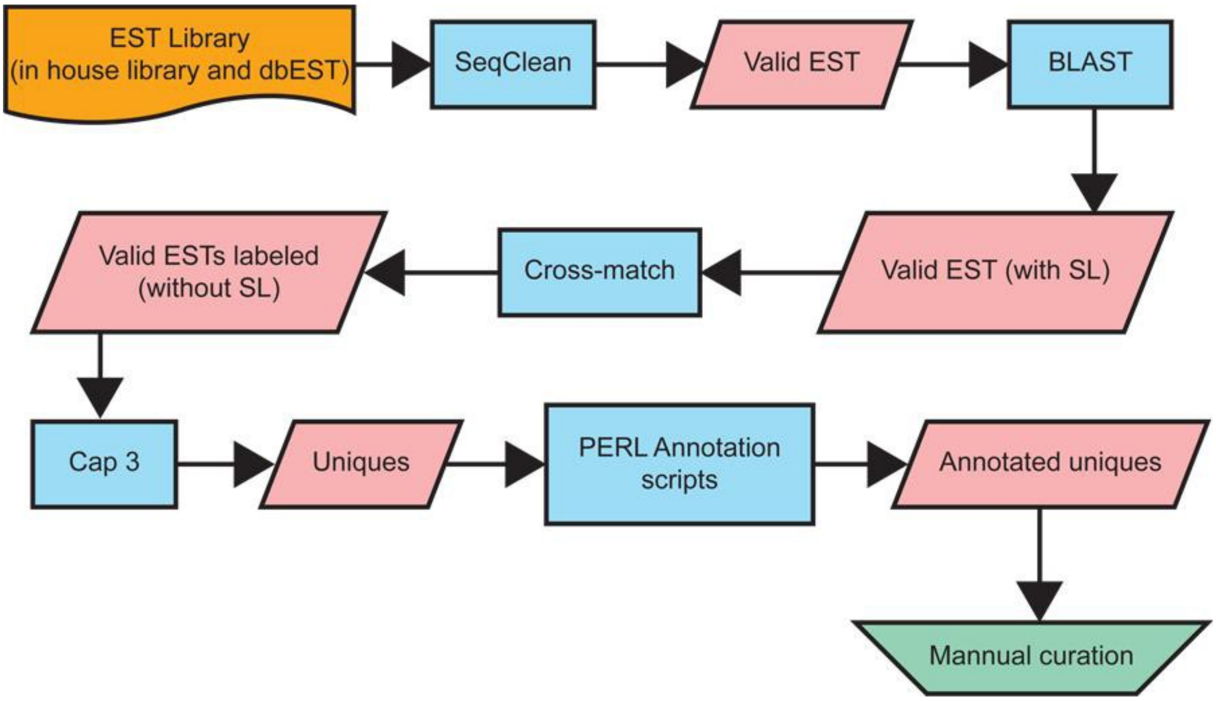


Figure 6: A graphical summary of the Bioinformatics pipeline used to perform this part of the study.

SL detection and cleaning

In order to identify SL-containing sequences in both SL-enriched EST library and dbEST, I used the Basic Local Alignment Search Tool (BLAST) and its comprising algorithms. To classify ESTs as SL-containing an e-value cutoff of 0.00005 was defined and BASH commands were further used to consider only sequences with at least 95% similarity and 25 contiguous nucleotides when aligned to the *S. mansoni* SL sequence. After identifying the SL-containing sequences, the local alignment software CrossMatch [Ewing *et al*., 1998] was used to remove the SL from all SL-containing sequences.

Command Line 3: Using BLAST to identify SL-containing sequences.

```
blastall -p blastn -i MultiFasta.sc -d SplicedLeader -e 0.00005 -o Alignments.blast
```

Command Line 4: Masking the SL sequence using CrossMatch.

```
cross_match MultiFasta.cm SplicedLeader.fasta -minmatch 10 -minscore 20
```

Sequence clustering and mapping

Both the SL-enriched EST library and *S. mansoni* dbEST were assessed using the CAP3 program [Huang & Madan, 1999] in order to generate sequence contigs. CAP3 was employed with default parameters, except for the -o and -p flags, which were set to 40 and 95, respectively. Therefore, two ESTs were only grouped into one contig when they had at least 40 aligned nucleotides with a minimum sequence identity of 95%.

Command Line 5: Contigs generation using CAP3.

```
cap3 MultiFasta.cm -o 40 -p 95 >> MultiFasta.cap3
```

After identifying the SL-containing sequences within the SL-enriched EST library and assembling the corresponding contigs, the sequences were mapped to the *S. mansoni* genome. For this purpose, I used BLAST to search for the previously assembled uniques in the parasite genome. The defined cutoff for the e-value was 10^{-10} and the minimum sequence similarity to be accepted was of 90%.

Command Line 6: Using BLAST to assemble contigs.

```
blastall -p blastn -i MultiFasta.cap3 -d SmansoniGenome -e 10e-10 -o
GenomeAlignments.blast
```

Gene Ontology (GO) assignment and manual annotation

GO assignment was performed for the set of sequences generated from the previous steps. GO categories were associated with each transcript using the web tool GoAnna [McCarthy *et al*., 2006]. GO Slim terms [McCarthy *et al*., 2006] were also retrieved for all sequences to allow a more general overview of gene ontology among the dataset. Manual annotation was performed for all known proteins (excluded sequences characterized as “expressed protein” and “hypothetical protein”). Literature search and homology information was carried out to assure the correct annotation of gene products. All proteins were clustered according to biological processes categories to provide a better understanding of trans-splicing function in the cell.

B. ASSESSMENT OF THE SL RNA SEQUENCES AND SL-CONTAINING TRANSCRIPTS OF DIFFERENT ORGANISMS

B.1 Sequence Retrieval and Database Generation

Manual searches were performed in the National Center for Biotechnology Information (NCBI http://www.ncbi.nlm.nih.gov) database to identify previously annotated SLe sequences. The searches were guided by sequence features, especifically considering entries under the “miscellaneous RNA” sobriquet. The exact expression informed in the search field was “misc_RNA[Feature key]spliced leader”. Once sequences were retrieved, a manual curation was carefully performed to exclude false positives and reduce redundancy. A consistent preliminary database (hereafter named SEED database) was then compiled containing only the manually curated and annotated sequences. Notably, only sequences from species in which the SLTS mechanism was previously described were included.

Using the sequences from the former mentioned SEED database as queries, I performed searches using BLAST [Altschul *et al*., 1990] to expand the set of SLe sequences, and thereby generate a secondary dataset (hereafter named EXTENDED database). Searches were performed in the nucleotide collections (nt) database from NCBI using the blastn program with parameters automatically adjusted to address short sequences and allow for the retrieval of up to 500 matches for each query sequence. The results were then analyzed to identify SLe sequences based on information from the SEED database.

Several criteria were used to characterize a sequence as SLe to generate the EXTENDED dataset. As objective criteria, I only considered annotated sequences that displayed 90% or higher nucleotide identity and a query coverage exceeding 90% when compared to the respective query. More subjectively, for matches meeting the objective criteria, I have analyzed the presence of such sequences in the 5' end of transcripts from these species. As an additional step, all uncharacterized sequences from organisms in which the presence of the SLTS mechanism was not previously demonstrated were not included in the datasets at first. Uncharacterized sequences were further analyzed to confirm or disprove those as SLe based on literature and sequence annotation.

Along with the retrieval of SLe sequences from the NCBI database, SL gene candidates were also retrieved to compose a separate database. SL gene candidates were identified based on sequence annotation and further analyzed for exclusion of false positives and redundancy reduction. Whenever duplicated sequences from a given species were identified, only one copy was kept to eliminate redundancy. After manual curation, the sequences were compared to the SLe from the same species present in the EXTENDED database and only candidates containing the entire SLe sequence followed by an intronic region were further considered to be SL genes.

B.2 Sequence Similarity Assessment

During the construction of the previously mentioned databases, when more than one SLe sequence was retrieved for a given species, Clustal [Larkin *et al*., 2007] alignments were performed to guide redundancy reduction and false positive discovery. Subsequent manual alignments were performed to cluster sequences from the EXTENDED dataset. Alignments were visually analyzed for the identification of duplicated sequences, such as completely identical sequences, partially identical sequences with missing residues and similar sequences presenting up to 5% (1 in every 20 positions, which is around the average size of the SLe sequences of most phyla) substitutions, deletions or insertions. Notably, missing information at the 5' end of SLe sequences is most likely to occur due to incomplete sequencing. For this reason, whenever data were missing for multiple 5' end positions (more than 5%) in a sequence for a given species, conservation of such positions in the phylum was measured considering only the remaining sequences. To better observe consensus regions within SLe sequences, alignments were also used as input for the generation of sequence logos using the WebLogo suite [Crooks *et al*., 2004].

B.3 Trans-Spliced Transcripts Retrieval

Once the SLe EXTENDED database was built, its sequences were used as queries to search for transcripts from the respective species bearing the SLe in the 5' end. BLAST searches were performed using the same parameters as described above, the same program (blastn) and reference database (nt). The results were manually curated to yield lists of transcripts that undergo SLTS in a given species. Those lists were joined together in a comprehensive set to allow further inspection of transcript conservation among species. All transcripts annotated as “hypothetical protein” or “unknown protein” were excluded from further analysis. Transcripts coding for identical proteins in different species were clustered together in a database to allow the assessment of conservation between species and across phyla of transcripts under SLTS regulation.

In parallel, data from the recently published work of Protasio and collaborators [Protasio *et al*., 2012] were used to yield a preliminary database of *S. mansoni* SLe-containing sequences. Fasta-format sequences for proteins coded by SLe-containing transcripts were retrieved from GeneDB [Logan-Klumpler *et al*., 2011], in which transcripts that undergo trans-splicing were associated with the GO ID number 0000870. In both cases described above, after retrieval, GO terms were assigned to each transcript sequence using the GoAnna and GoSlim [McCarthy *et al*., 2006], tools from the AgBase web portal [McCarthy *et al*., 2010]. Subsequently, manual annotation and classification according to the main biological function was performed for all transcripts (except those retrieved from the *S. mansoni* database) based on literature data.

B.4 Further Analyses

To analyze structural conservation and topological features of the entire SL RNA molecule, all formerly mentioned SL gene sequences were submitted to RNAfold [Hofacker, 2003], a program from the Vienna package for RNA structure generation and energy assessment. Structures were then visually analyzed to search for conserved features in sequences from different species.

The genomic location and copy number of the SLe sequences from two *Caenorhabditis* species were retrieved using the BLAST-like Alignment Tool (BLAT) [Kent, 2002] results and further analyzed. One hundred nucleotides downstream of each sequence were also retrieved to putatively represent the entire SL gene and allow for sequence comparison.

All steps were aided by simple SHELL and/or PERL scripts to automatically manipulate files, organize and categorize data, recognize specific patterns within the text and perform searches in a file. A methodological workflow is presented in the figure below (Figure 7) to illustrate each step of this study.

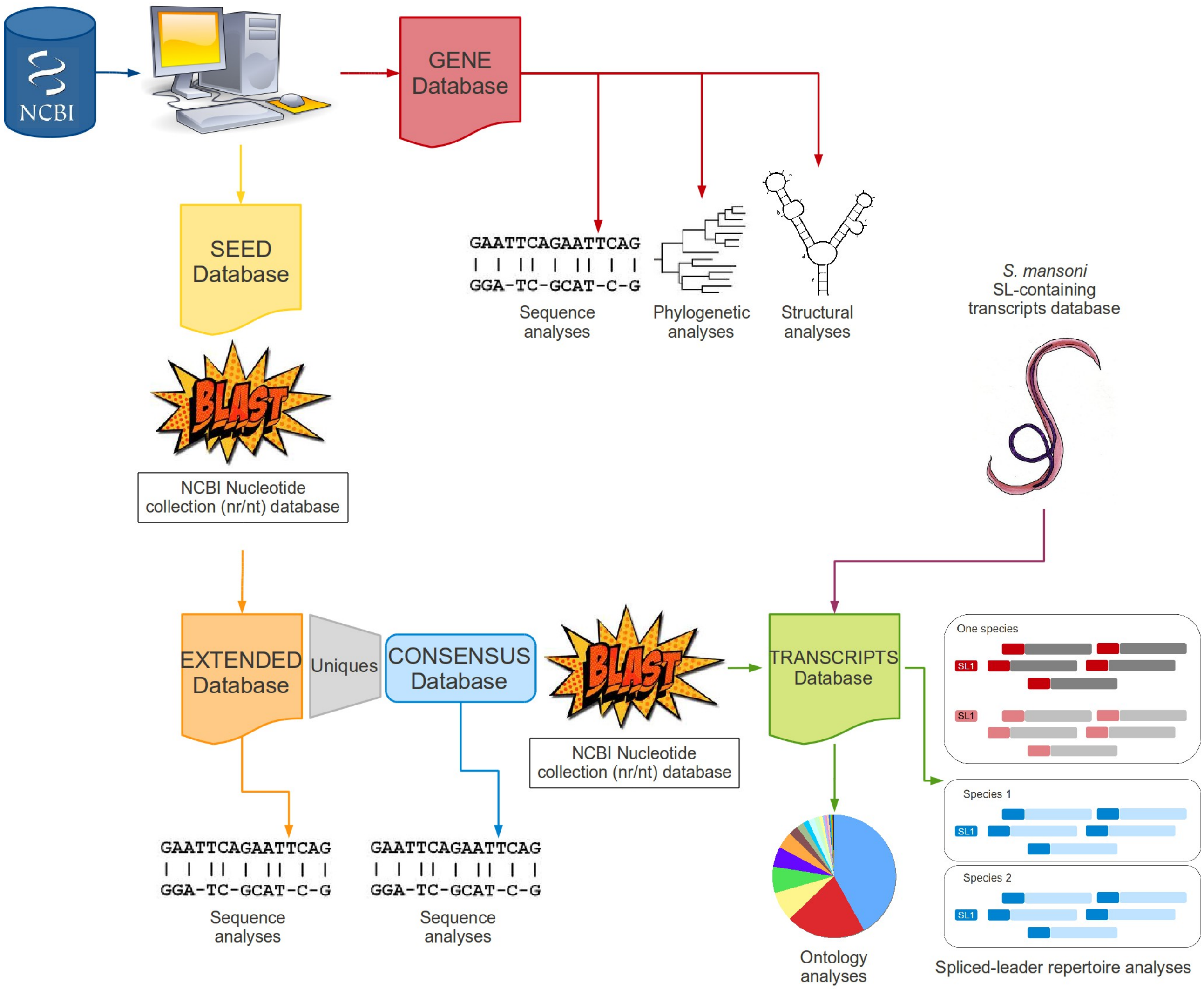


Figure 7: Methodological workflow describing the main steps performed in this work as a fluxogram.

IV. Results

A. ANALYSES OF THE SLTS MECHANISM IN SCHISTOSOMA MANSONI

A.1 Dataset Generation and Data Analysis

To determine whether the SLTS mechanism could target specific functional gene categories in *S. mansoni*, a SL-enriched EST dataset was generated by Dr. Marina de Moraes Mourão with subsets from diverse parasite life-cycle stages. The enriched EST dataset yielded a total of 3,087 sequences, from which 481 were from schistosomulae, 502 from eggs, 600 from females, 623 from males and 881 from adult worms of mixed gender. After removing spurious sequences and vector contamination by SeqClean, the database was left with 2,781 valid sequences. Of those, 1,665 were classified as SL-containing sequences according to previously described criteria. When calculating the ratio of valid ESTs in the SL-enriched dataset containing the SL sequence versus the total number of ESTs, I found that 59.8% of the ESTs carry the SLe sequence, according to the stringency parameters. This high percentage of SL-sequences confirms the enrichment of the SL dataset, since the same percentage is of only 0.1% when the entire dbEST dataset is analyzed with the same protocol (Table 1). Furthermore, recent observations from undirected RNA-Seq data [Protasio *et al*., 2012] report that ~11% (1,178 SL-transcripts) of all *S. mansoni* transcripts are being processed by SLTS and our group has generated directed RNA-Seq libraries and identified 60% of the protein-coding transcripts as SL-containing [Boroni *et al*., in preparation]. Taken together, these data clearly point to a high enrichment of SL-containing sequences in the dataset, confirming that the methodology described here represents a suitable alternative for similar surveys in other organisms.

All 1,665 SL-containing sequences were retrieved from the SchistoDB through BLAST searches and mapped to the *S. mansoni* genome. A total of 989 sequences were mapped and further clusterized by CAP3, resulting in 258 uniques (102 singlets and 156 contigs) These 258 uniques were mapped into a set of 162 different protein-coding sequences from the parasite genome. A total of 64 uniques were classified as *conserved hypothetical proteins* (7), *hypothetical proteins* (10) or *expressed proteins* (47) and therefore were not included in the functional analysis. A final set of 98 unique sequences was defined and used in subsequent analysis (see further section, Figure 8 and Table 2).

EST Library (Library dbEST ID)	Detected ESTs	Valid ESTs after sequence trimming	Valid ESTs with SL sequence	Valid ESTs with SL versus total valid ESTs	Valid ESTs mapped in *S. mansoni* genome and annotated	Valid annotated ESTs versus total valid ESTs
Egg (028002)	502	499	263	0.527	153	0.581
Schistosomula (028003)	481	481	258	0.536	157	0.608
Female worms (028000)	600	600	407	0.678	244	0.599
Male worms (028001)	623	623	356	0.571	256	0.719
Adult worms (027999)	881	578	381	0.659	179	0.469
All	3087	2781	1665	0.598	989	0.594
dbEST	205892	201694	276	0.001	242	0.876

Table 1: Enrichment of the SL-enriched EST library when compared to the dbEST dataset. The table presents the number of sequences for each life-cycle stage contained in the SL-enriched EST library and derived from dbEST after each step of processing and curation.

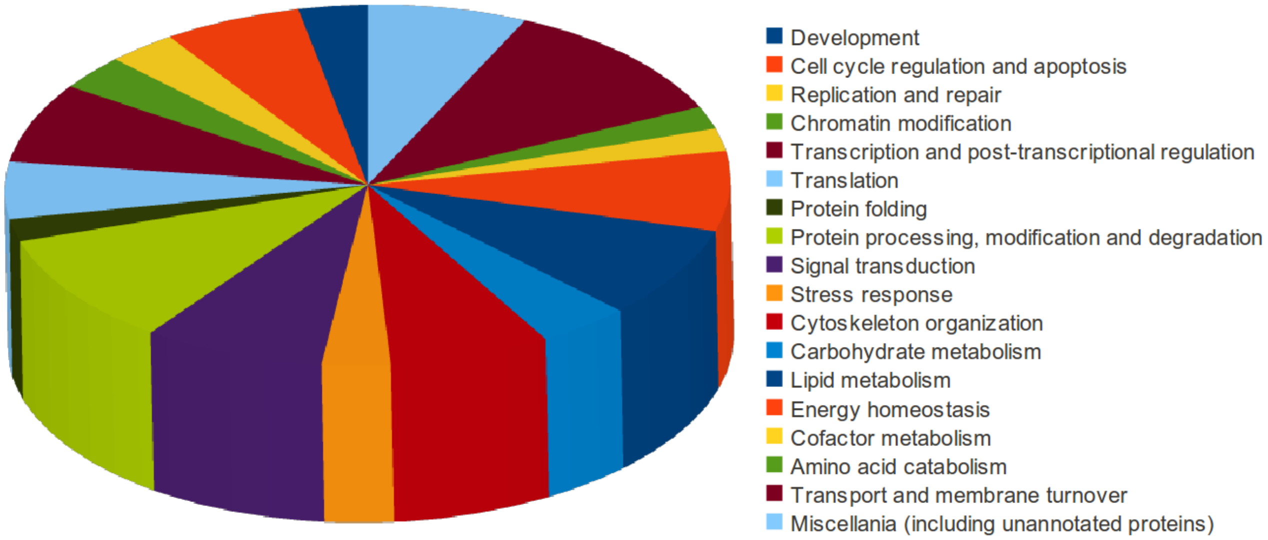


Figure 8: Pie chart illustrating the distribution of annotated proteins in different functional classes. The protein distribution was generated after manual curation and classification of transcripts regarding functional characterization.

A.2 GO and Manual Annotation

To identify biological processes, subcellular localizations and molecular functions represented by trans-spliced transcripts, GO terms were assigned to 78 of the 98 protein-coding uniques (the remaining 21 could not be associated to any GO term by GOAnna). From all 78 proteins assessed, only 30 were assigned a particular subcellular localization from GOSlim results. Among those, 8 were proteins from the membrane, 8 were cytoplasmic (including 2 cytoskeleton-associated proteins), 5 were nuclear proteins, 5 were mitochondrial proteins and another 5 were classified merely as intracellular proteins with no specific localization. When analyzing molecular functions assigned to the sequences through GO annotation, 24 proteins were classified as metal-binding (among calcium, magnesium, iron, zinc and other metal ions), 12 as nucleotide-binding proteins, 10 as nucleic acid-binding proteins (1 specifically interacts with RNA and 4 with DNA) and 5 were identified as ATP binding proteins. Also, four proteins were classified as glycolytic enzymes. In the biological processes category, 21 proteins were identified as functioning in metabolic processes, of which 8 were associated with biosynthesis, while 5 proteins were classified as related to redox mechanisms. The remainders were not specified.

Although several tendencies were observed from the GO annotation, there were no clear biases within the analyzed trans-spliced protein dataset regarding cellular location, biological process or molecular function. This is in agreement with the current opinion that the trans-splicing mechanism is not associated to any specific gene category or protein feature.

In addition to the GO annotation, manual annotation of all 98 protein-coding uniques was performed. This was an important step, from which additional information about protein function was retrieved from literature. The entire set of annotated proteins was subdivided into 19 classes of biological processes, namely: Development, Cell cycle regulation and apoptosis, Replication and repair, Chromatin modification, Transcription and post-transcriptional regulation, Translation, Protein folding, Protein processing, modification and degradation, Signal transduction, Stress response, Cytoskeleton organization, Carbohydrate metabolism, Lipid metabolism, Energy homeostasis, Cofactor metabolism, Amino acid catabolism, Transport and membrane turnover and Miscellania (Table 2 and Supplementary Table 1).

CLASS	PROTEIN NAME (SchistoDB ID)	PROTEIN FUNCTION
Development	bola-like protein my016 (012050)	Morphogenesis, cell growth
	cactin (038640)	Morphogenesis, cell growth and survival
	chibby protein pkd2 interactor (035010)	Morphogenesis, cell growth
Cell cycle regulation and apoptosis	chl1 helicase (130210)	Chromossome segregation
	enhancer of rudimentary protein (175210)	Pyrimidine synthesis and cell cycle
	jun activation domain binding protein (190100)	Activation of mitotic checkpoint
	programmed cell death protein (067490)	Cell regulation and apoptosis
	pten-related phosphatase (012400)	Cell regulation and apoptosis
	regulator of chromosome condensation-related (074990)	Chromossome condensation
Replication and repair	cript-related (132860)	DNA double strand break repair
	endonuclease III (006710)	DNA repair of damaged pyrimidine
	pol-related (186960.2)	DNA replication
Chromatin modification	histone h1/h5 (003770)	Nucleosome formation
	histone deacetylase hda2 (138770)	Histone deacetylation
	mrg-binding protein (130630)	Histone acetylation
Transcription and post-transcriptional regulation	DNA-directed rna polymerase I (072800)	rRNA synthesis
	lsm1, putative (175550)	RNA splicing (snRNP formation)
	nucleotide binding protein 2 (154750)	RNA processing, traffic and stability (snRNP formation)
	pre-mRNA splicing factor (080110)	RNA splicing
	ribonucleic acid binding protein S1(070200)	RNA traffic and surveillance

	small nuclear ribonucleoprotein U1a, U2b (069880)	RNA splicing (spliceosome protein)
	serine/arginine rich splicing factor (113620.4)	RNA splicing (spliceosome protein)
Translation	eukaryotic translation initiation factor 4e- binding protein (074390)	Inhibition of protein translation
	eukaryotic translation initiation factor 3 subunit (158500)	mRNA-ribosome interaction
	mitochondrial ribosomal protein L24 (035410)	Mitochondrial protein synthesis
	mitochondrial ribosomal protein S33 (039980)	Mitochondrial protein synthesis
	translation machinery-associated protein (169920)	Translation
Protein folding	peptidyl-prolyl cis-trans isomerase-like 3 (080150)	Acceleration of protein folding
	rotamase (160790)	Acceleration of protein folding
Protein processing, modification and degradation	26S proteasome non-atpase regulatory subunit (181380)	Protein degradation via proteasome
	ammd-like (peptidase) (074940)	Protein degradation
	aspartic proteinase (013040.1)	Protein degradation
	beta-1,4-galactosyltransferase (153110)	Glycosyltransferase activity
	glycosyltransferase (046880.2)	Glycosyltransferase activity
	methionine aminopeptidase (011120)	Removal of initial methionine
	o-sialoglycoprotein endopeptidase (012450)	Sialoglycoprotein degradation
	protein-l-isoaspartate o-methyltransferase (058140.2)	Proteins repair and degradation
	peptidase (165910)	Protein degradation
	ring finger (028430)	Protein ubiquitination
Signal transduction	hybrid protein kinase, other group,WNK family (154440)	Chloride ion traffic
	protein phosphatase 2C (133060.3)	Serine/Threonine phosphatase
	ral guanine nucleotide dissociation stimulator (154920)	GTPase activation
	rab 15, 13, 10, 1, 35, 5 (072290)	Small GTPase
	rho GDP-dissociation inhibitor-related (045610)	Interaction with Rho GTPases
	serine/threonine kinase, CAMK group,	Signal transduction pathways for

	CAMKL family, AMPK subfamily (142990)	cellular energy homeostasis
	serine/threonine kinase, CMGC group, RCK family, MAK subfamily (132890)	Signal transduction pathways for cell cycle regulation
	tbc1 domain family member (152820)	Small GTPase activation
Stress response	glutaredoxin (006550.2)	Antioxidant defence
	hsp70-interacting protein (064860)	Protection against physiologic stress
	thioredoxin mitochondrial type (037530)	Antioxidant defence
Cytoskeleton organization	beta-parvin-related (133200)	Regulation of cell adhesion and cytoskeleton organization
	cofilin, actophorin (120700.2)	Actin binding
	dynactin subunit 3 (dynactin light chain p24) (136990)	Microtubule and dynein binding
	nuclear movement protein nudc (103320)	Dynein binding
	tektin (059600)	Structural component of microtubules
	tropomodulin (003710)	Actin growth regulation
	tropomyosin, putative (031770.4)	Actin mechanics regulation
Carbohydrate metabolism	enolase (024110)	Glycolysis
	glyceraldehyde-3-phosphate dehydrogenase (056970.1)	Glycolysis
	l-lactate dehydrogenase (038950)	Glycolysis
	ribulose-5-phosphate-3-epimerase (104580.2)	Pentose phosphate pathway
Lipid metabolism	1-acylglycerol-3-phosphate o-acyltransferase (000070)	Lipid metabolism
	acetyl-CoA C-acyltransferase (129330)	Lipid catabolism
	acyl-coenzyme A binding domain containing (147930)	Lipid metabolism
	arv1 (021280)	Sterol homeostase
	serine palmitoyltransferase I (028080)	Sphingolipids biosynthesis
	3-dehydroecdysone 3alpha-reductase-related (042680)	Ecdysteroids inactivation
	isopentenyl-diphosphate delta isomerase (130430)	Biosynthesis of isoprenoids
	sterol reductase-related (124300)	Biosynthesis of steroids
Energy homeostasis	atpase inhibitor (023010)	Inhibition of ATP synthesis
	ATP synthase delta chain (082120)	Catalysis of ATP synthesis
	cytochrome C oxidase copper chaperone	Involved in ATP synthesis

	(029150)	
	cytochrome C subunit I	Involved in ATP synthesis
	ectonucleotide pyrophosphatase/phosphodiesterase (104270)	ATP hydrolysis
	hexaprenyldihydroxybenzoate methyltransferase (029060)	Ubiquinone and ATP synthesis
	ubiquinone biosynthesis protein (106190)	Ubiquinone and ATP synthesis
	ubiquinol-cytochrome C reductase complex ubiquinol binding protein (024110)	Involved in ATP synthesis
Cofactor metabolism	molybdopterin-binding (126650)	Molybdopterin cofactor synthesis
	uroporphyrinogen-III synthase (079840)	Heme synthesis
Amino acid catabolism	4-hydroxyphenylpyruvate dioxygenase (007960)	Tyrosine catabolism
	valacyclovir hydrolase (193750)	Alpha-amino acid catabolism
Transport and membrane turnover	fatty acid binding protein (174440.4)	Fatty acids transport
	mitochondrial carrier protein-related (058130)	Transport across mitochondrial membrane
	nuclear transport factor (037700)	Proteins transport
	peripheral-type benzodiazepine receptor, putative (102510)	Transport of cholesterol, porphyrin and anions
	solute carrier family 35 member C1, putative (155830)	Transport of GDP-fucose
	B-cell receptor-associated protein-like protein (175660)	Vesicular transport of proteins between the ER and the Golgi
	dynamin-associated protein (135790)	Vesicular transport
	golgi snare bet1-related (130940)	Vesicular transport between the ER and the Golgi
	rabb and c (169460)	Vesicular transport of proteins between the membrane and the Golgi
	small VCP/p97-interacting protein (194560)	ER integrity maintainance
	vacuolar ATP synthase proteolipid subunit 1, 2, 3 (004310.1)	Transport across membranes and regulation of internal pH
Miscellania **including unannotated proteins**	amyloid beta A4 protein related (144990)	Involved in several physiological processes
	autophagy protein 16-like (135430)	Involved in cellular autophagy
	metal dependent hydrolase-related (024490)	Presents hydrolase activity

	protein C14orf153 precursor (147820)	Unannotated protein
	Sj-Ts1 (123270)	Unannotated transmembrane protein
	tetraspanin D76 (041460)	Involved in cell adhesion
	WD repeat protein SL1-17 (049520)	Involved in several physiological processes

Table 2: Functional classification and description of trans-spliced transcripts in the final dataset. Transcripts were classified according to their biological function after manual annotation.

A.3 Protein Length and Exon Composition of Trans-Spliced Transcripts

In order to verify if the SLTS mechanism could be associated to genes containing small exons, as suggested by Davis *et al*. (1995), I have calculated the length of all protein-coding sequences from SchistoDB and of all 98 annotated protein-coding genes in the dataset. Based on this survey I found that proteins derived from trans-spliced transcripts averaged ~400 amino acids in length. By comparison, the average length of all *S. mansoni* proteins from SchistoDB was only slightly higher (~450 amino acids). I have also estimated the number of exons per gene and the average exon length for all protein-coding sequences of the parasite in both SchistoDB and the dataset produced herein. Again the analysis showed a conserved exon composition in both sets of proteins, with an average number of 6 exons per protein, each and an average length of ~73 amino acids per exonic region.

A.4 SL Knockdown in *S. mansoni S*porocyts

In an attempt to disrupt the trans-splicing mechanism, Dr. Marina de Moraes Mourão, from our group has designed siRNAs to the *S. mansoni* SL sequence. Over a 7 day period of cultivation in the presence of siRNAs, she monitored cultured sporocysts for various phenotypes including interruption of miracidial/sporocyst transformation rate, mortality during the *in vitro* cultivation period and larval motility and length. Visual monitoring revealed that SL-siRNA treatment altered only the larval length phenotype, resulting in sporocysts with reduced size compared to control (Figure 9A and B). To verify that the length alteration was significant, she measured captured images of live sporocysts and analyzed the data using the Metamorph software. Average length measurements of sporocysts from the SL-siRNA treated-groups were significantly reduced when compared to larvae of the control decoy-siRNA-treated and blank groups (Figure 9C).

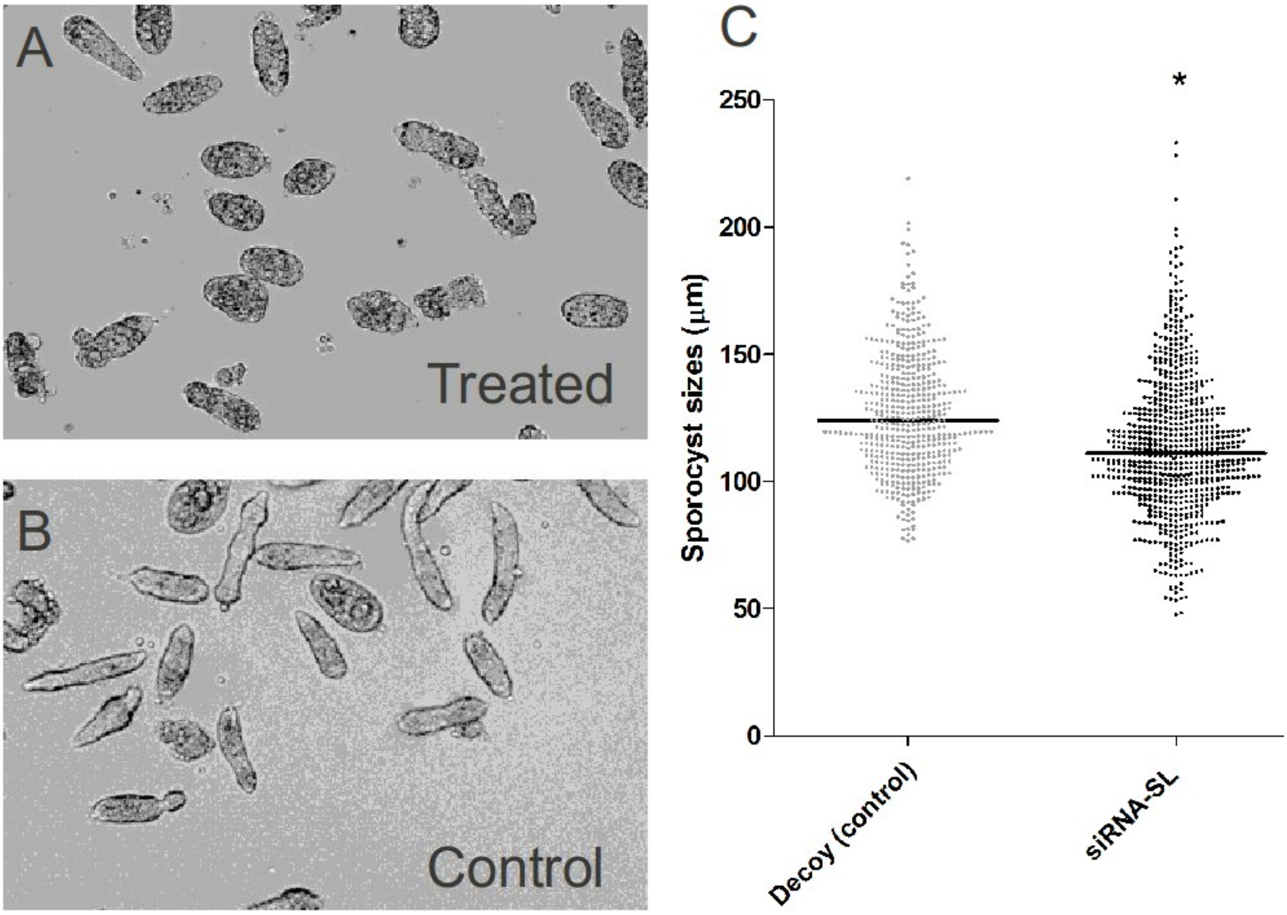


Figure 9: *In vitro* cultured *S. mansoni* larvae seven days post dsRNA treatments. **(A and B)**: brightfield photomicrographs of in vitro cultured *S. mansoni* sporocysts after seven days of treatments with SL-siRNA **(A)** compared to the control decoy-siRNA **(B)**, illustrating the effects of the exposure to SL-siRNA on sporocyst lengths; **(C)**: graphic representation of sporocyst length measurements after seven days of siRNA treatment by scatter plots with the calculated median values indicated by the horizontal bars. The median values for siRNA treatments were compared to decoy-siRNA (grey plots) treatment control. All mesurements were statistically analysed using Mann-Whitney U-test within each experiment. Asterisk means $p<10^{-4}$.

Quantitative PCR was performed in order to correlate the observed phenotype and gene expression. Notably, the five target transcripts randomly chosen from known SL-sequence-containing genes in the enriched EST dataset exhibited significant reductions of at least 50% when compared to the decoy-siRNA treatment control. Transcripts for a calcium channel, ATPase inhibitor, phosphoserine-phosphohydrolase, thioredoxin and enolase demonstrated knockdown levels of 52%, 48%, 50%, 68% and 55%, respectively (data not shown). In addition, to check for unspecific (off-target) knockdown, non-trans-spliced transcripts were assessed for transcript levels after SL-siRNA treatment. No significant alteration in transcript levels was observed for the proteins SmZF1, SmRBx and SOD after SL-siRNA treatment. Thus, all tested trans-spliced transcripts analyzed by qRT-PCR showed a similar decrease on transcript level after SL-siRNA treatment, suggesting a systemic trans-splicing knockdown effect.

B. ASSESSMENT OF THE SL RNA SEQUENCES AND SL-CONTAINING TRANSCRIPTS OF DIFFERENT ORGANISMS

B.1 Sequence Retrieval and Database Generation

In the course of this work I have assembled a comprehensive database of SL RNA sequences from several phyla. When a simple pattern-matching search was performed in the NCBI database to retrieve SLe sequences, 1,161 matches were found. Among those, the majority (757) were from kinetoplastids and almost half (544) specifically from *Trypanosoma spp*. The rest included sequences from nematodes (182), flatworms (98), dinoflagellates (95), cnidarians (4), rotifers (2) and chordates (1), phyla in which the SLTS mechanism was previously described. Sequences from organisms of other phyla were also retrieved but not included in the datasets given the lack of consistent evidence supporting the presence of SLTS and/or because no transcripts were found bearing the sequence in the 5' end. After redundancy reduction, false positive exclusion and validation by manual curation, a SEED database was defined containing only sequences with consistent evidence to be considered as SLe. This initial database was comprised 69 sequences from 7 different eukaryotic phyla (Supplementary Table 2): rotifera (2), chordata (1), cnidaria (2), dinoflagellate (8), euglenozoa (33), nematoda (18) and platyhelminthe (5).

These sequences were subsequently used as queries to extend the database based on BLAST similarity searches. The retrieved sequences were analyzed according to previously described criteria and a final EXTENDED database was generated, comprising 157 sequences from 148 different species (representing 81 genera) from the same seven phyla (Supplementary Table 3). Notably, all 157 sequences are, in fact, replicates of only 48 unique sequences, which were further clustered into 30 groups of highly similar sequences. This result indicates the high degree of sequence conservation, particularly between species from the same phylum (as will be further discussed). These 30 SLe sequences originated a third database, named the CONSENSUS database (Table 3).

Consensus Name	**Consensus Sequence**	**Length (composition)**	**Plenty***
Rotifera 1	GGCTTATTACAACTTACCAAGAG	23 (8A/5C/4G/6T)	3/3
Chordata 1	GATTGGAGTATTTGGTTGTATTAAG	25 (6A/8G/11T)	7/7
Cnidaria 1	ACTTTTTAGTCCCTGTGTAATAAG	24 (6A/4C/4G/10T)	5/7
Cnidaria 2	CAAACTTCTATTTTCTTAATAAAG	24 (9A/4C/1G/10T)	2/7
Dinoflagellate 1	WCCGTAGCCATTTTGGCTCAAG	22 (4A/6C/5G/6T/1W)	23/23
Nematoda 1	GGTTTAATTACCCAAGTTTGAGGG	22 (6A/3C/5G/8T)	45/53
Nematoda 2	GGTTTAATTACCCAAGTTTAAG	22 (7A/3C/4G/8T)	2/53
Nematoda 3	GGTTTTAACCCAGTTAACCAAG	22 (7A/5C/4G/6T)	2/53
Nematoda 4	AGGTATTTACCAGATCTAAAAG	22 (9A/3C/4G/6T)	1/53
Nematoda 5	TACCGTTCAATTAATTTTGAAG	22 (7A/3C/3G/9T)	1/53
Nematoda 6	GTAATAAGAAAACTCAAATAAG	22 (13A/2C/3G/4T)	1/53
Nematoda 7	GGTTTTTACCCAGTATCTCAAG	22 (5A/5C/4G/8T)	1/53
Platyhelminthe 1	AACCGTCACGGTTTTACTCTTGTGATTTGTTGCATG	36 (6A/7C/8G/15T)	3/10
Platyhelminthe 2	AACCTTAACGGTTCTCTGCCCTGTATATTAGTGCATG	37 (8A/9C/7G/13T)	2/10
Platyhelminthe 3	AACTATAACGGYTCTCTGCCGTGTATATTAGTGCATG	37 (9A/7C/8G/12T/1Y)	2/10
Platyhelminthe 4	CACCGTTAATCGGTCCTTACCTTGCARTTTTGTATG	36 (6A/9C/6G/14T/1R)	3/10
Euglenozoa 1	AACTAACGCTATATAAGTATCAGTTTCTGTACTTTATTG	39 (12A/6C/5G/16T)	21/54
Euglenozoa 2	AACTAACGCTATTATTGATACAGTTTCTGTACTATATTG	39 (12A/6C/5G/16T)	12/54
Euglenozoa 3	AACTAACGCTATTATTAGAACAGTTTCTGTACTATATTG	39 (13A/6C/5G/15T)	4/54
Euglenozoa 4	AACTAAAGTTATTATTGATACAGTTTCTGTACTATATTG	39 (13A/4C/5G/17T)	2/54
Euglenozoa 5	AACTAAAGCTWTTATTAGAACAGTTTCTGTACTATATTG	39 (13A/5C/5G/15T/1W)	2/54
Euglenozoa 6	AACTAAAATTATTTATAATACAGTTTCTGTACTATATTG	39 (15A/4C/3G/17T)	1/54
Euglenozoa 7	AACTAAAGATTTTATTGTTACAGTTTCTGTACTATATTG	39 (12A/4C/5G/18T)	1/54
Euglenozoa 8	AACTTACGCTATAAAAGTCACAGTTTCTGTACTTTATTG	39 (12A/7C/5G/15T)	2/54
Euglenozoa 9	AACTAACGCTATTATTGTTACAGTTTCTGTACTTTATTG	39 (10A/6C/5G/18T)	3/54
Euglenozoa 10	AACTAACGCTAWAAAAGWTACAGTTTCTGTACTTTATTG	39 (13A/6C/5G/13T/2W)	2/54
Euglenozoa 11	AACTAACGCATTTTTTGTTACAGTTTCTGTACTTTATTG	39 (9A/6C/5G/19T)	1/54
Euglenozoa 12	AACTAACGCTATATTTGTTACAGTTTCTGTACTWTATTG	39 (10A/6C/5G/17T/1W)	1/54
Euglenozoa 13	AACTAACGCTATTCTAGATACAGTTTCTGTACTTTATTG	39 (11A/7C/5G/16T)	1/54
Euglenozoa 14	AACCAACGATTTAAAAGCTACAGTTTCTGTACTTTATTG	39 (13A/7C/5G/14T)	1/54
* Total number of sequences in the phylum / Number of sequences matching consensus			

Table 3: The CONSENSUS database of SLe sequences. This table presents the 30 unique sequences that comprise the CONSENSUS database with their related abundance within the phylum, length and nucleotide composition.

In addition to the phyla previously represented within the SEED database, sequences from other phyla have also met the requirements to be considered SLe but were not included in the EXTENDED database. Species of arthropods, mollusca, mycetozoa, apicomplexa, ciliophora, echinodermata, plants and even a bacterial species presented putative SLe sequences, although most of these phyla have never been proven to harbour the SLTS mechanism. No obvious decision on whether these are real SLe sequences could be reached, and those were therefore excluded from the dataset (Supplementary Table 4).

B.2 Description of the EXTENDED Database

In this section, I present a short description of sequences in the EXTENDED database according to phyla, which is fully available as Supplementary Table 3 and graphically summarized in Figure 10 below. There are 3 species of rotifera in the dataset that share an identical adenine rich 23 nucleotide SLe sequence. All 7 chordata species share an identical SLe sequence, apart from missing residues. The consensus is a 25 nucleotide sequence enriched in thymine nucleotides. Cnidarians are also represented by 7 sequences, but from 5 species of the same genus (*Hydra*). All species share the exact same SLe sequence and, among those, 2 of the species present an additional SLe sequence, which is identical in both species. The 2 consensus sequences are composed of 24 nucleotides from which 10 are thymines. All 23 species of dinoflagellates in the database share an identical 22 nucleotide SLe sequence that is only degenerated in the first position (either A or T), with a balanced nucleotide composition. Notably, this is the only phylum in which the SLe sequence itself carries the Sm-protein binding site (a T-rich element, that like in *C. elegans* and many other species is a $AT_{4-6}G$ motif). There are 53 sequences from the nematoda phylum within the database. Among these, 45 are identical apart from a repetition of the nucleotide G of variable length at the 3' end. Among the 8 remaining sequences, 2 are identical pairs and 4 are unique, one of which harbours an Sm binding site (the other 2 sequences present a putative inverted Sm binding motif). Most sequences have 22 nucleotides (apart from the variable repetition of guanines), and the majority is slightly richer in A and T nucleotides. Among the platyhelminthes, there are 10 species from 7 different genera represented in the dataset. All SLe have a high percentage of thymines and are 36-37 nucleotides long. The sequences can be divided into four different groups of two or three virtually identical sequences each. Notably, each group contains species of a unique order. A classical Sm-protein binding site was found in one of such groups, and another group presents a putative inverted binding site. Euglenozoans are unique organisms in this context because in these species all transcripts are processed by SLTS. This phylum is represented by 54 sequences of 39 nucleotides from 14 different genera. The sequences can be clustered together in 8 groups of identical sequences plus 6 isolated sequences that are not identical to any other sequence. In summary, the terminal regions (first 6 and last 20 nucleotides) are conserved in all sequences, whereas the central region is variable. All 16 *Leishmania* species share the exact same sequence, whereas the 22 *Trypanosoma* species were divided into 4 groups with identical sequences and 2 isolated sequences. Alternatively, with less stringency, SLe from this phyla could be clustered into three groups according to specific signatures within the last 20 nucleotides: (i) *Trypanosoma* species; (ii) *Leishmania*, *Leptomonas*, *Wallaceina* and *Chritidia* species (with identical sequences); and (iii) the remaining genus, which presents more diverse sequences (the data presented in this section are presented in

Table 3 and Supplementary Table 3 and graphically represented in Figure 10).

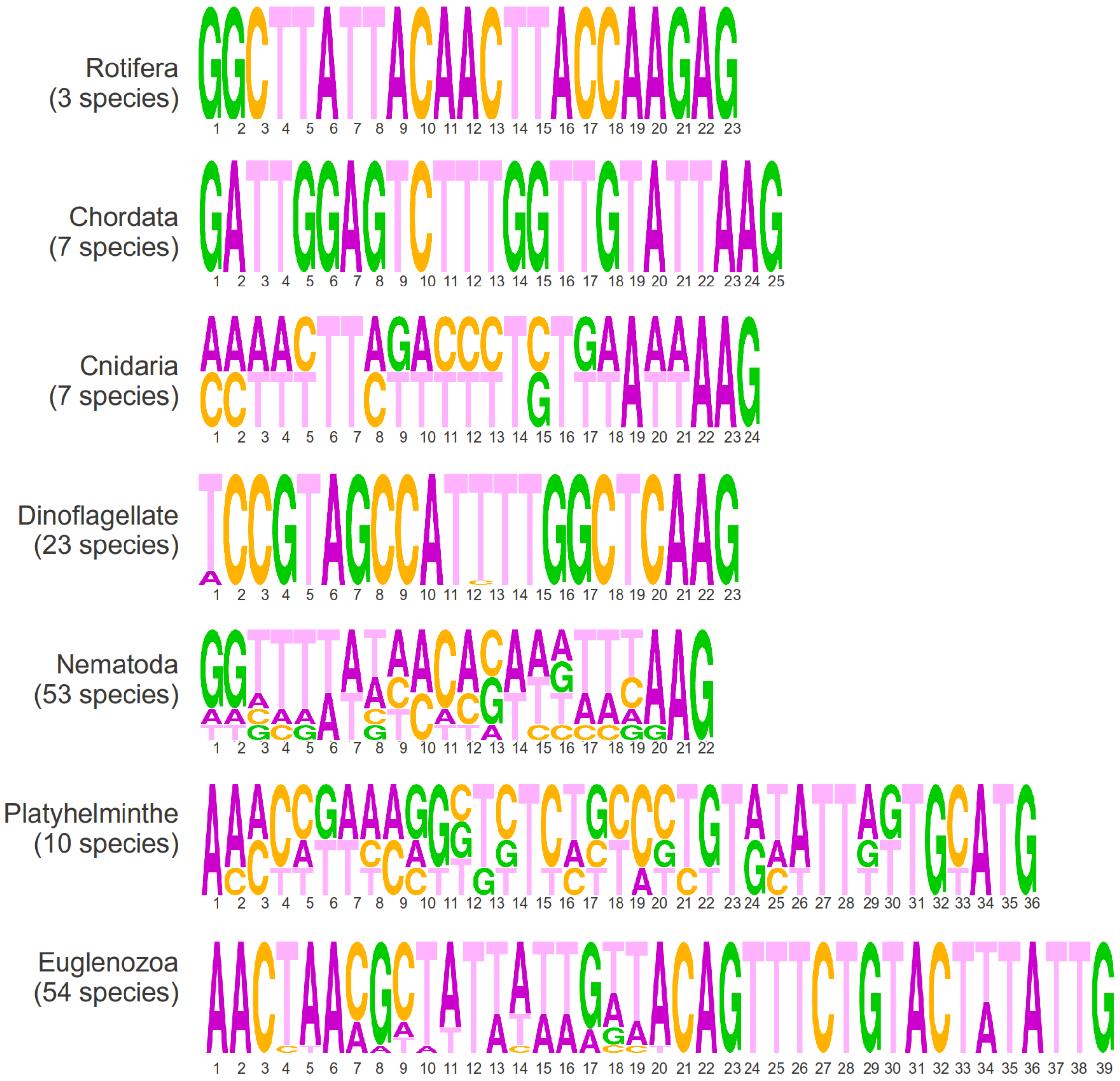


Figure 10: Sequence logos for SLe sequences. Logos were generated from the alignment of all SLe sequences from each phylum showing sequence lengths in numbers and the frequency of the four nucleotides in each position.

B.3 SLe Sequence Comparison Among Phyla

When analyzing the EXTENDED dataset of SLe sequences, a high sequence conservation within each phyla was revealed, but a very low conservation among different phyla was found. One interesting feature I highlight as a general tendency is that sequence length is more conserved than sequence composition itself in any given phyla. There are some clear differences between sequences from euglenozoa and platyhelminthes and those from other phyla. Such differences include the sequence length and terminal residue identity. Whereas sequences from all other phyla range from 22 to 25 nucleotides in length, sequences from euglenozoans and platyhelminthes are 39 and 36-37 nucleotides long, respectively. As for the nucleotides at the 3' end, in which SL exon-intron cleavage occurs and the SLe is incorporated into mRNAs, euglenozoans and platyhelminthes have TTG and ATG patterns, respectively, whereas sequences from other phyla have an AAG pattern (except for one nematoda consensus sequence and the sequences from rotifera that end in GAG). Regarding the 5' end, 17 of the 18 consensus SLe sequences from euglenozoa and platyhelminthe present a conserved AAC pattern as the first nucleotide triad. For all other phyla, there is no conservation of nucleotides in the 5' end. Notably, all but five sequences from the CONSENSUS dataset present a TTT triplet, which in all dinoflagellates and three species from the other phyla, are part of the Sm binding site (in five other species it is part of a hypothetical inverted Sm binding motif). When considering all sequences in the CONSENSUS database, there is an evident enrichment in adenines (~30% of all nucleotides) and in thymines (~40% of all nucleotides) in comparison to guanines and cytosines (which together comprise only ~30% of all nucleotides).

B.4 SL Gene Sequence Comparison Among all Phyla

Several candidate SL gene sequences were identified through manually analyzing the retrieved sequences from the NCBI database. Within the preliminary dataset, I performed a manual curation to reduce redundancy and exclude false positives, and I also mapped the available SLe sequence to the initial portion of the putative genes of the same species. As a result, 30 sequences remained and were further separated according to phyla and analyzed. Gene length was relatively variable, from 75 to 123 nucleotides; although most sequences were approximately 100 nucleotides long (mean length is 107). The most evident patterns within sequences are the cleavage site in the exon-intron boarder and the presence of the Sm-protein binding site, which is usually in the intronic portion of the gene. Nevertheless, there is some degree of SLi sequence conservation among species from a given phylum, although it is much lower than for the SLe sequence alone. Differing from the exonic portion, the intronic nucleotide composition has an almost even distribution of nucleotides with 24% adenine, 24% cytosine, 26% guanine and 26% thymine.

B.5 SL Genomic Location and Composition in two *Caenorhabditis* Species

Regarding the genomic position, it is already known that most SL genes are located near the 5S ribosomal RNA gene and comprises multiple copies in tandem (apud Hastings, 2005). I have used BLAT to map SLe sequences in the genomes of *C. elegans* and *C. remanei*. Both species have two different SLe sequences in the database, and these sequences are identical between species. One sequence (hereafter named as SLeI and identical to the SL1 sequence from Ross, 1995) is shared among the majority of the nematoda species (45 of the 53 sequences from this phylum), and the other (hereafter named SLeII, closely related to the SL2 first described by Huang and Hirsh, 1989 and identical to SLf as described by Ross, 1995) is only common to these two *Caenorhabditis* species in the database.

In *C. elegans*, the SLeI sequence is repeated 13 times in chromosome V, twice in chromosome I and once in chromosome III (thus comprising 16 copies). The SLeII sequence is repeated twice in chromosome I and has only one copy in chromosomes III and IV (thus comprising 4 copies). The genome of *C. remanei* has a much higher concentration of SLeI sequences, which are represented by 135 copies. The SLeII sequence on the other hand seems to only have 9 copies. The karyotype is not available in BLAT for the latter species and therefore, it was not possible to map sequences onto chromosomes. Notably, the SLeI sequence, which is widespread among nematoda species, is always more represented in the genomes of these two species (16 versus 4 copies in *C. elegans* and 135 versus 9 copies in *C. remanei*).

Regarding the intronic portion of the putative SL gene sequences, I have analyzed all occurrences of the *C. elegans* SLeI in chromosomes V, III and I and observed that 10 repetitions have identical introns (all in chromosome V) and that the other putative SLeI gene sequences are very divergent from one another. Notably, these 10 identical sequences are the only ones to present classical Sm-protein binding site. When evaluating the intronic portion of the 4 putative genes with the SLeII sequence in *C. elegans*, a high similarity is observed, although the sequences are not identical. These sequences also present Sm binding sites, although three of these have five thymine within the repetition, whereas the additional sequence has four. I have noticed a conserved GTTAG pattern in the 4 putative SLeII gene sequences that is also present at a different position in the 10 identical sequences and is absent in all other 6 putative SLeI gene sequences (two other short patterns - ACAA and GGAA - are also present in the 14 sequences, but are not exclusive to these sequences). Among the 9 sequences of the putative SLeII genes in *C. remanei*, eight are identical (apart from one substitution in two sequences) and the other contains approximately 10% substitutions. All nine sequences contain classical Sm-protein binding sites. In addition, these sequences are closely related to the putative SLeII gene sequences of *C. elegans*, although not identical. When analyzing the alignment of 135 putative SLeII gene sequences in *C. remanei*, several clusters of highly similar or identical sequences are observed. One important observation is the lack of Sm binding sites in most of the sequences. Only 27 sequences bear Sm binding site motifs. These can be divided into groups of identical or nearly identical sequence; one of which is also closely related to the main group of *C. elegans* putative SLeI gene sequences. I have reconstructed a phylogenetic tree with all the SL sequences from both species (data not shown) and observed a more randomized distribution of SL genes not bearing Sm binding sites in comparison to sequences that contain this motif. I then narrowed the analysis to consider only Sm binding site-containing sequences, and the corresponding tree is shown in the supplementary material (Supplementary Figure 1). The tree is divided into three main groups: (i) a group of SLeI gene sequences containing all *C. elegans* and approximately half the *C. remanei* SLeI gene sequences, (ii) a group of all SLeII gene sequences and (iii) a more distant group of *C. remainei* SLeI gene sequences. Each group can be further clustered into smaller groups with closely related sequences.

B.6 SL Gene Structural Comparison

I have assigned secondary structures to all 30 SL gene sequences with RNAfold from the Vienna package. As a result of this structural analysis, a tendency for SL RNAs to form a Y shaped molecule was observed (Figure 11). The topology is the result of three stem-loops and a branch point and seems to be conserved in nearly all analyzed species, although the branch point position and stem-loop length are variable. The average free energy (ΔG) value for gene structures in fixed secondary structures was -30.6 Kcal/mol (ranging from -14.4 to -51.80 Kcal/mol).

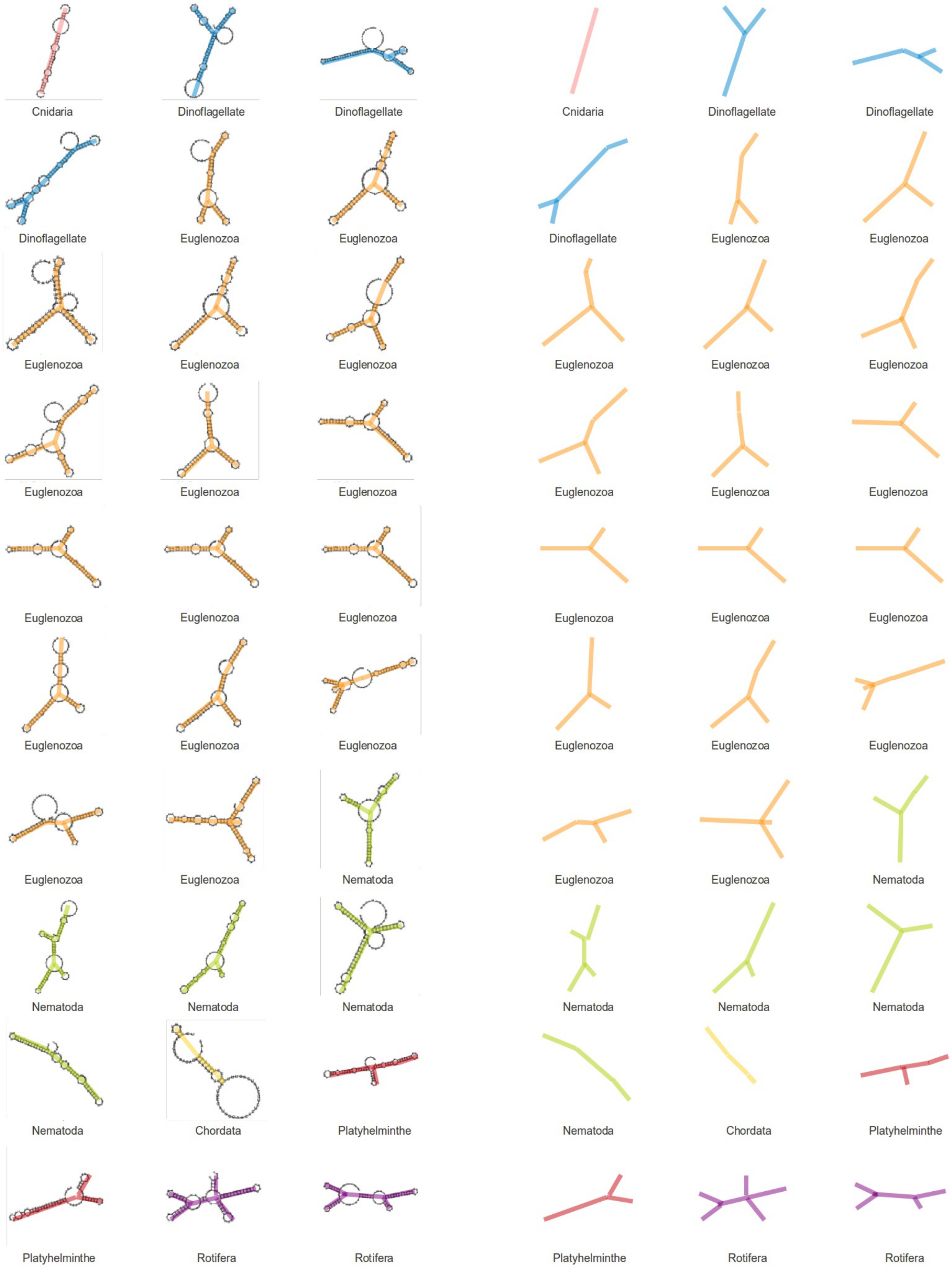


Figure 11: Structures of SL genes. Secondary structures were generated for SL genes of different phyla and a visual analysis of these structures point to a conserved Y-shape for the SL molecule. The backbones of the structures are displayed to the right for a simpler view.

B.7 Analysis of Transcripts Trans-Spliced to Specific SLe

To assess whether the sets of trans-spliced transcripts from different organisms harboring the same SLe sequence are similar, I conducted BLAST searches in the nr/nt database using two different sequences as queries: (i) a sequence shared by different dinoflagellate species and (ii) a sequence shared by different nematoda species (other than *C. elegans*). The retrieval of up to 500 transcripts was allowed for each SLe from the different species of each phylum (dinoflagellate and nematoda). I have excluded *C. elegans* from the survey because in this organism the addition of different SLe sequences in different transcripts has been investigated by high throughput sequencing (Allen *et al.*, 2010) and could not be included here without introducing a substantial bias to the analyzed data. Among all retrieved dinoflagellate sequences, 133 remained after false positive exclusion and redundancy reduction. From these, 63 (over 47%) transcripts were shared by more than one species (representing 30 different transcripts, from which many code for ribosomal proteins). The remaining 70 are specific to only one species. From the nematoda transcripts, after manual curation, I analyzed 158 transcripts from which 54 (over 34%) are shared by different species. I then decided to analyze whether in one given species two different SLe sequences would regulate different sets of transcripts. BLAST was used to search the nr/nt database for transcripts from *H. vulgaris* bearing any of the two different SLe sequences. As a result I have identified 30 transcripts trans-spliced with one sequence and 13 with the other, none of which were related to both SLe sequences.

B.8 Analyses of Trans-Spliced Transcripts from All Species

In a similar context, a more overall analysis was performed with the unique sequences of the EXTENDED database of SLe sequences as queries for BLAST searches (with the blastn program) within the nr/nt database. Search results and manual curation resulted in a final set of 455 transcripts (Figure 12 and Supplementary Table 4), among which 5 were annotated as “alternatively spliced”. From this set, I have previously excluded transcripts from euglenozoa species because in these organisms, SLe addition is ubiquitously used to solve polycistronic transcripts. Among these 455 transcripts, 237 are present in only one species and 218 are shared by more than one species (totaling 60 unique sequences). Additionally, 138 are common to species from different phyla (totaling 32 unique sequences). Enolase, calmodulin, gluthatione S-transferase, ATP synthase subunits, cyclins, eukaryotic translation initiation factors, superoxide dismutase, ras-related proteins and ribosomal proteins are among the most ubiquitous trans-spliced proteins, as these are found in species of at least three different phyla.

I have automatically assigned GoSlim terms to all transcripts and the results revealed no clear bias to any specific gene category. This was true for the entire set of transcripts and also for each individual phylum (data not shown). Because this protocol was not curated and has classified transcripts into generic and less informative classes, I have decided to manually annotate and classify all transcripts aiming to achieve a more specific, reliable and informative result. I have therefore separated transcripts according to their main biological function, generating a total of 30 different functional classes with different representations. The result was unexpected and revealed unique classes composed of transcripts that seem to more frequently undergo trans-spliced in each phylum, although this was not true for the entire set of transcripts, in which no specific category seemed to be dominant (Figure 12 and Supplementary Table 4).

In parallel, using data available from the work of Protasio and colleagues (Protasio *et al*., 2012), I have generated a set of 1,411 SLe-containing transcripts from *S. mansoni*. Preliminary analyses after GO assignment and GoSlim retrieval revealed no clear bias regarding the biological processes under the control of the SLTS mechanism, although the number of transcripts in a few biological processes was higher, such as whole-cell functions, metabolic processes and organismal development (data not shown).

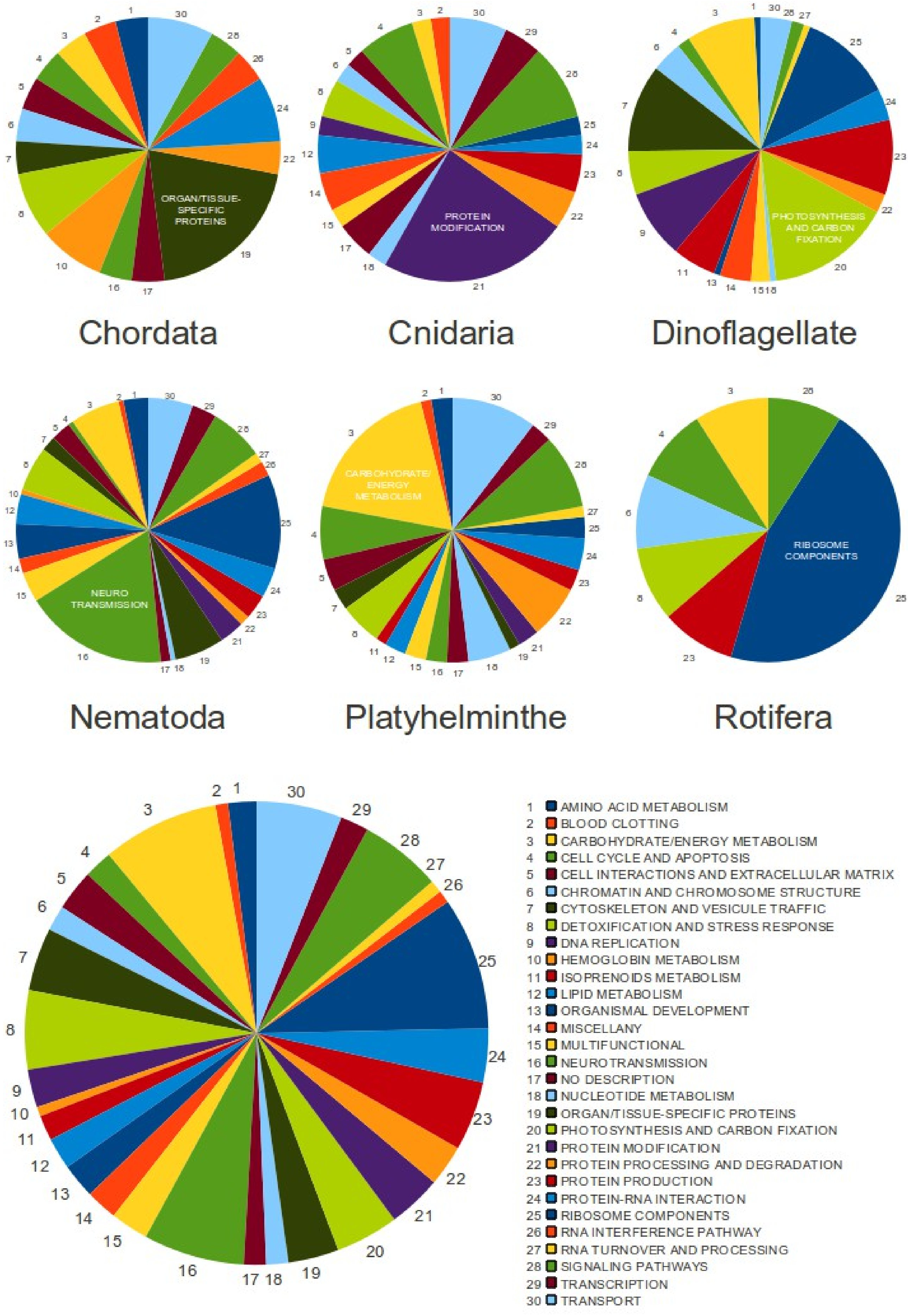

1
30
28
26
24
22
ORGAN/TISSUE-SPECIFIC PROTEINS
19
17
16
10
8
7
6
5
4
3
2
Chordata
3
2
30
29
28
25
24
23
22
PROTEIN MODIFICATION
21
18
17
15
14
12
9
8
6
5
4
Cnidaria
1
30
28
27
25
24
23
22
PHOTOSYNTHESIS AND CARBON FIXATION
20
1518
14
13
11
9
8
7
6
4
3
Dinoflagellate
2
1
30
29
28
27
26
25
24
23
22
21
19
17
18
NEURO TRANSMISSION
16
15
14
13
12
10
8
7
5
4
3
Nematoda
2
1
30
29
28
27
25
24
23
22
21
19
18
17
16
15
12
11
8
7
5
4
CARBOHYDRATE/ ENERGY METABOLISM
3
Platyhelminthe
3
28
RIBOSOME COMPONENTS
25
23
8
6
4
Rotifera
2
1
30
29
28
27
26
25
24
23
22
21
20
19
17
18
16
15
14
13
12
11
10
9
8
7
6
5
4
3
1 AMINO ACID METABOLISM
2 BLOOD CLOTTING
3 CARBOHYDRATE/ENERGY METABOLISM
4 CELL CYCLE AND APOPTOSIS
5 CELL INTERACTIONS AND EXTRACELLULAR MATRIX
6 CHROMATIN AND CHROMOSOME STRUCTURE
7 CYTOSKELETON AND VESICULE TRAFFIC
8 DETOXIFICATION AND STRESS RESPONSE
9 DNA REPLICATION
10 HEMOGLOBIN METABOLISM
11 ISOPRENOIDS METABOLISM
12 LIPID METABOLISM
13 ORGANISMAL DEVELOPMENT
14 MISCELLANY
15 MULTIFUNCTIONAL
16 NEUROTRANSMISSION
17 NO DESCRIPTION
18 NUCLEOTIDE METABOLISM
19 ORGAN/TISSUE-SPECIFIC PROTEINS
20 PHOTOSYNTHESIS AND CARBON FIXATION
21 PROTEIN MODIFICATION
22 PROTEIN PROCESSING AND DEGRADATION
23 PROTEIN PRODUCTION
24 PROTEIN-RNA INTERACTION
25 RIBOSOME COMPONENTS
26 RNA INTERFERENCE PATHWAY
27 RNA TURNOVER AND PROCESSING
28 SIGNALING PATHWAYS
29 TRANSCRIPTION
30 TRANSPORT

Figure 12: The classification of SLe-containing transcripts from all analyzed species. Pie charts are presented for each phylum individually (upper charts) and for the entire set of 455 transcripts (bottom chart). Functional classes are represented by color, and the arrow indicates the direction that corresponds to the legend. The most populated class for each phylum is labeled by name.

V. DISCUSSION

A. ANALYSES OF THE SLTS MECHANISM IN SCHISTOSOMA MANSONI

Over 30 years ago, scientists studying the expression of varying surface glycoproteins (VSG) in trypanosomatids have found that all transcripts, from different VSGs presented the same 39-nucleotides sequence on their 5' region, which was later named spliced leader [Booth-royd & Cross, 1982]. Few years after, HeLa cells studies demonstrated a previously unknown type of splicing, in which exons from otherwise unrelated genes were transcribed and fused together in a process then named trans-splicing [Solnick, 1985; Konarska *et al.*, 1985]. Both findings have contributed to the definition of a new mechanism, the SLTS. Therefore, the mechanism was first described as a post-transcriptional processing strategy to process polycistronic transcripts in trypanosomatids [Agabian, 1990]. In subsequent years, the SLTS was observed in several other organisms, but its functional role remains poorly defined outside the context of polycistronic transcription. One of the first hypotheses was that the SLTS could be functionally associated with specific genes or gene categories. For example, in *Ciona intestinalis* trans-splicing was suggested to more prominently regulate the expression of specific functional gene categories, such as plasma and endomembrane system, Ca2+ homeostasis and actin cytoskeleton [Matsumoto *et al.*, 2010]. However, in *S. mansoni* this hypothesis was never supported, as there was no clear evidence to link the SLTS to any particular gene category, biological process, molecular function, life-stage, gender, tissue or subcellular localization of protein-coding transcripts [Davis *et al.*, 1995].

In the present study, I have analyzed a large set of transcripts undergoing SLTS within different *S. mansoni* life-cycle stages. Since the percentage of protein-coding transcripts that would undergo trans-splicing in *S. mansoni* was estimated by our group as ~60% [Boroni *et al*., in preparation], and given the complexity of its life-cycle, it is plausible that this process is important for the regulation of gene expression involved on the parasite development and/or adaptation to different environments. Nevertheless, the fraction of SL-containing transcripts reported so far in *S. mansoni* is even lower than the percentage of genes that undergo SLTS in organisms such as *C. elegans* and *Ascaris* spp., which can be up to 70% and 90%, respectively [Allen *et al*., 2011].

As previously stated, the generated EST dataset was highly enriched in SL-containing sequences, given that the dataset contained 6,000-fold increase in SL-containing sequences when compared to the total number of *S. mansoni* ESTs from dbEST (60% of SL-containing sequences in the dataset versus only 0.01% in the dbEST). This group of SL-sequence-enriched transcripts represents a highly informative set of genes that could potentially inform several features of the SLTS mechanism. An interesting result comes from comparisons between the set of annotated trans-spliced sequences and those reported by Protasio *et al*. (2012). Approximately half of the protein-coding transcripts from the annotated dataset of unique sequences are also present in the larger set generated by these authors. Two observations can be made from this comparison: (1) the strategy used in the present study was appropriated for investigation of SL-transcripts, as demonstrated by the fact that 50 SL-transcripts were identified in both datasets, and (2) since the dataset contained unique protein-coding genes not found in the broader Protasio dataset, it is inferred that the approach allowed for the retrieval of different SL-transcripts. Differences in the content of these datasets could be explained by the methodologies, since Protasio and collaborators used a “whole transcriptome” approach, whereas our group has developed a more selective protocol involving the capture and enrichment of SL-transcripts prior to cloning and sequencing.

Some of the transcripts in the *S. mansoni* dataset generated here were previously described as SL-containing sequences, such as enolase and an ATPase inhibitor [Davis, 1996; Davis *et al*., 1995]. Although other protein-coding sequences described in the 1995 study by Davis *et al*., such as synaptobrevin, a guanine nucleotide-binding protein and HMG-CoA reductase, were not identified within the dataset, sequences related to these genes or their pathways (e.g., Golgi Snare bet1, small GTPases or mevalonate pathway enzymes) were represented herein. Although a trans-spliced form of the glycolytic enzyme, glyceraldehyde 3-phosphate dehydrogenase (GAPDH), was not previously observed in *S. mansoni*, this enzyme was found to undergo the SLTS process in *Caenorhabditis* spp. As I report here, I also captured a trans-spliced GAPDH transcript. One hypothesis to explain this discrepancy among different studies is that a transcript regulated by SLTS is not necessarily always processed by this mechanism and thus any given transcript may be or not be subjected to trans-splicing at a particular time and cell state. This could account for the suggested role of SLTS as a mechanism for gene expression modulation and coordination [Davis *et al*., 1995].

Bachvaroff and Place (2008) showed that the SLTS of dinoflagellate transcripts correlates with their expression levels, suggesting that highly expressed genes are more likely to be trans-spliced. This correlation was made by comparing trans-spliced transcript levels with their corresponding protein abundance as estimated by proteomic analyses [Beranova-Giorgianni, 2003]. Accordingly, I have observed that at least 25% of protein-coding transcripts that I classified as undergoing trans-splicing in *S. mansoni* code for proteins that are present in previously reported proteomic studies [Cass *et al*., 2007; Castro-Borges *et al*., 2011; Curwen *et al*., 2004; Knudsen *et al*., 2005; Mathieson & Wilson 2010; Wu *et al*., 2009], including some glycolytic enzymes and many ribosomal proteins I identified as trans-spliced transcripts in the present dataset. This is an indication of the importance of the SLTS mechanism for different organisms, since it could contribute for an increase in protein abundance.

One of the trans-spliced transcripts of particular interest identified in this study is the ubiquinol-cytochrome C reductase complex ubiquinol binding protein (UbCRBP), which had been previously described as the first cistron of a trans-spliced resolved polycistronic transcript, in which only the second gene (enolase) would undergo trans-splicing [Agabian, 1990]. UbCRBP transcript has also been described as undergoing trans-splicing in *E. multiloculares* [Brehm *et al*., 2000], reinforcing the idea that the SLTS is a conserved mechanism among selected orthologous genes. These findings also revealed that the *S. mansoni* UbCRBP sequence insertion of the SL occurred before the second exon of the gene, which also contains an upstream AG acceptor splicing signal. This result suggests that the SLTS may act to generate alternatively spliced products in *S. mansoni*, with different exons alternatively receiving the SL sequence. Alternative splice-sites were previously observed in the 3-hydroxy-3-methyl-glutaryl CoA reductase transcript of *S. mansoni,* in which the third exon accepts the SL sequence [Rajkovic *et al*., 1990]. Thus, alternative trans-splicing appears to be a conserved mechanism in the parasite, and suggests a unique way to expand the protein repertoire in this organism. I have further being involved in an even larger study conducted by our group using RNA-Seq data, in which the alternative splicing was observed to be a very important facet of the SLTS mechanism [Boroni *et al*., in preparation].

Previous results have suggested that short exons on pre-mRNAs are more prone to undergo trans-splicing [Davis *et al*., 1995]. Based on the data, however, I have not observed any differences in exon size when comparing the dataset to the whole set of *S. mansoni* genes. Additionally, protein length and the number of exons per sequence were also equivalent when the dataset was compared to the set of all protein-coding sequences from SchistoDB.

Interestingly, many transcripts coding for proteins from the spliceosome machinery seem to undergo trans-splicing. This observation indicates that the SLTS mechanism may be self-regulated and, if so, represents a unique characteristic of this mechanism. In addition to transcripts encoding spliceosome proteins, the eukaryotic translation initiation factor 4e-binding protein and a subunit of the eukaryotic translation initiation factor 3 transcripts also were shown to undergo trans-splicing. These proteins are part of a mechanism that enables efficient translation of the trimethylguanosine (TMG)-capped mRNAs in nematodes [Wallace *et al*., 2010] and are of great importance for the SLTS, since one of the functions attributed to this mechanism is to facilitate translation of transcripts containing this modified cap.

To my knowledge, the experiments performed by Dr. Mourão from our group represent the first (and only) report of an attempt to disrupt the SLTS mechanism in a metazoan using RNAi in order to assess its regulatory function. Introduction of SL-siRNA to *in vitro* cultured sporocysts resulted in a phenotype characterized by a reduction in larvae size. Because a large variety of SL-containing genes may have been affected by knockdown through RNAi, it is difficult to predict how this “size reduction” phenotype came about. In general, this phenotype may have resulted from a metabolic imbalance caused by a decrease in a large number of different trans-spliced transcripts. Clearly, proteins associated with crucial metabolic processes may have been affected by the knockdown of the trans-splicing mechanism, thereby resulting in a systemic decrease in metabolism, leading to possible parasite starvation and decrease in larvae length. Apart from the previous discussion for the occurrence of trans-splicing in glycolytic transcripts, other affected processes could also account for the diminished size of sporocysts after knockdown, for example, proteins from the cell cycle, metabolic pathways other than glycolysis and morpho-proteins as the ones described in my results. The phenotype may also reflect a stress caused by a lower activity of the trans-splicing mechanism. Taken this together, I can infer that the knockdown parasites are away from physiological equilibrium, and growth impairment is a common consequence of systemic stress and starvation, which could be caused by the reduced expression of transcripts under trans-splicing control in *S. mansoni*.

Although lethality was not observed after seven days of SL knockdown, it is worth noticing that the attempt to silence the trans-splicing machinery lowered by 60% the expression of SL-containing transcripts. It is likely that this partial knockdown at the transcript level may had exerted only a minor effect on the intact larvae, not only because transcript levels, although lower, were still present in these parasites, but also their encoded proteins may have continued to persist for extended time, depending on their turnover rate. Thus, with the remnant transcript levels and residual protein pools available, this was probably sufficient to maintain larvae viability, even though they appeared morphologically stunted. Other possible explanations could be that only a fraction of molecules from the same transcript would undergo trans-splicing. Interestingly, all tested trans-spliced transcripts demonstrated a similar decrease on transcript level, suggesting a systemic trans-splicing knockdown effect after SL-siRNA exposure. Since a large percentage of the *S. mansoni* transcripts population seems to be trans-spliced, a hypothesis for the limited level of transcript knockdown could include a saturation of the components of the RNAi machinery.

In this work I found *S. mansoni* orthologues of genes from different categories that had been described to undergo SLTS as well. They correspond to genes coding for ribosomal proteins, small nuclear ribonucleoproteins, members of the solute carrier family, glyceraldehyde-3-phosphate dehydrogenase, thioredoxin, a mitochondrial ribosomal protein component, a WD-repeat containing protein, a peptidyl prolyl cis-trans isomerase, serine/threonine kinases and a cAMP dependent protein kinase [Davis *et al*., 1995; Davis 1996]. These findings support the hypothesis that there is some conservation among genes regulated by SLTS, indicating that some genes may have maintained trans-splicing as a form of post-transcriptional regulation throughout evolution. This data could therefore suggest that the SLTS mechanism originated in a common eukaryotic ancestor rather than independently in multiple organisms. However, there are also previous findings that favour an evolutionary scenario in which the SLTS mechanism had a polyphyletic origin during eumetazoan evolution, rather than arising from a common ancestor [Hastings, 2005]. These contrasting proposals were yet to be assessed, but required a more comprehensive study, taking into account a larger number of representative species and constructing an evolutionary sequence-based molecular profile for the trans-splicing mechanism. Next I discuss my findings in such a study.

B. ASSESSMENT OF THE SL RNA SEQUENCES AND SL-CONTAINING TRANSCRIPTS OF DIFFERENT ORGANISMS

One of the most discussed topics related to the SLTS mechanism is its evolutionary origin. There is previous evidence for a unique origin in a common ancestor but there is also evidence for multiple unrelated origins. From the 157 sequences comprising the described EXTENDED database, it is possible to observe few features that could universally define a SLe. Almost all sequences have a conserved WWG at the 3' end and a TTT pattern. As for the latter, I suggest it may be the vestige of a Sm binding site motif that was once present within the SLe and was further lost or transferred to the intronic portion. Notably, there is a high sequence conservation rate within phyla, specifically to the level of subfamilies, although only the two aforementioned features are conserved among unrelated phyla. In addition to the sequence itself, the SLe length is even more conserved within phyla, indicating this may be a crucial characteristic of the molecule.

There are important characteristics that differentiate euglenozoa and platyhelminthe consensus SLe sequences from other sequences, such as the length and composition of the 5' and 3' ends. This could reflect a distinct role of the SLTS mechanism in such species. Regarding trypanosomatids, the SLTS mechanism has a crucial role because in these organisms large regions of the genome are transcribed as polycistronic units and SLe incorporation is crucial for their resolution into monocistronic units. In platyhelminthes, the mechanism is supposed to be acting as an expression regulator for a more restricted subset of genes. A unique feature of the SLe sequence in this phylum is the presence of an ATG as the last nucleotide triad. This could account for an intrinsic start codon, that could generate an alternative open reading frame in which the SLe insertion occurs within the transcript sequence, thus giving rise to alternative forms of the resulting protein. Nevertheless, the relationship between sequences from platyhelminthes and euglenozoans is neither clear nor it is phylogenetically expected because these two phyla come from different eukaryotic kingdoms (animalia and excavata, respectively).

Similarity between SL gene sequences from different phyla is less evident, but some features are conserved. Specifically, the last nucleotide of all SLe sequences is a guanine. This is because the cleavage site between the exon and intron is a highly conserved GGTA motif, from which the first nucleotide (G) is the last nucleotide of the exon and the other three (GTA) are in the beginning of the intron. Another conserved motif is the binding site for the Sm protein (usually a $AT_{4-6}G$ motif). This conserved binding site is usually present in the intron, with the exception of the dinoflagellate sequences (and a few sequences from other phyla), in which the Sm binding site is located in the exon. The consequences of this exonic localization of the Sm binding site are not completely understood. Notably, intronic nucleotide composition is homogeneous (not biased for a specific nucleotide type), indicating a tendency for a lack of selective pressure in this region, except for the presence of the Sm binding site. From all analyzed phyla, euglenozoa species present the longest SLe sequences, with 1.5 times the size as compared to SLe from other species (which average 23 nucleotides). This discrepancy may indicate an independent origin of SLTS in euglenozoa, which may in turn be related to the unique transcription strategy adopted by such organisms and the role of SLe insertion in the post-transcriptional processing of polycistronic transcripts.

There are two different SLe sequences in each of the two *Caenorhabditis* species in the EXTENDED dataset (I have internally named these as SLeI and SLeII). These sequences were mapped with BLAT on the genomes of the respective organisms, and the retrieved sequences of the putative SL genes (the SLe sequence plus 100 nucleotides downstream) were analyzed. The results show a differential abundance of each SLe in the genomes, with the SLeI sequence being the most abundant in both species. Notably, this SLe is shared with the majority of the nematode species represented in the database. In each species, regarding sequence similarity, putative genes of identical SLes seem to be more related to one another than to putative genes of other SLe. Although not all sequences are identical for a given SLe, sequence alignments present groups of closely related introns.

In their 1987 article, Krause and Hirsh [Krause & Hirsh, 1987] reported the existence of more than 100 SLeI genes in *C. elegans*, which is much higher than I have observed. This discrepancy can be explained by the methodology used for SL gene identification, which considered sequences that were a 90% match to the SLe (20 out of 22 nucleotides) and was performed by Southern blot analysis. By contrast, in this study, I considered only 100% matches in the genome, thus restricting the number of retrieved sequences. Unfortunately, because the *C. elegans* genome was not available in 1987, I cannot thoroughly compare my results to the previously published results.

When analyzing putative intronic sequences of different SLe, sequence conservation is more clear among species. There are only 10 SLeI gene sequences in *C.*

elegans bearing the Sm binding site, and these are closely related to one group of SLeI sequences of *C. remanei*, in which only 27 sequences contain Sm binding sites. All of the remaining sequences (not bearing the Sm binding site) do not seem to be related to one another. The scenario is simpler for the SLeII sequences because there are fewer in both species. All 4 *C. elegans* SLeII sequences contain Sm binding sites and are similar. The 9 *C. remanei* SLeII sequences are nearly identical, contain Sm binding sites and are closely related to the *C. elegans* SLeII sequences.

When analyzing the phylogenetic tree in the correspondent figure (Supplementary Figure 1), one should see two main groups: one comprising approximately half *C. remanei* SLeI gene sequences and another containing all other sequences. The latter is further divided into two groups, one with the SLeII gene sequences and another with SLeI gene sequences. This seems to indicate the existence of a more ancient SLe from which the SLeI and SLeII genes have arisen by duplication, most likely before speciation between *C. elegans* and *C. remanei* (given the similarity among intronic sequences of both species). The second group of SLeI genes from the later species may have arisen by duplication after speciation or (most likely given the divergence among sequences in this group) may be a result of duplication from a common ancestor prior to speciation and was then lost by the former species. Taken together, these results show that the orthologous SLe gene sequences (identical SLes in different species) are more related to one another than to paralogous genes (other SLe in identical species). This most likely indicates that divergence between SLeI and SLeII took place before speciation.

This study observed that a given species may have more than one SLe sequence, supporting the literature (as in the classic case of *C. elegans* first reported in Huang and Hirsh, 1989). This could account for the differential expression of transcripts or the production of different protein repertories under certain environmental conditions or developmental stages. I have identified two different SLe sequences from *Hydra vulgaris* and analyzed the set of transcripts related to each in public databases. Although this cannot be considered a fully conclusive analysis, it gives an indication of the possible roles for different SLe sequences. Considering the two distinct SLe from *H. vulgaris*, whereas one was found in 30 annotated transcripts, the other was found in less than half of these (only 13 transcripts). There was no superposition of the two sub-groups, indicating that each SLe sequence could be trans-spliced to a distinct set of transcripts. Notably, the SLe that was added to a higher number of transcripts is the most conserved when compared to other species from identical genera. This result is in agreement with the recent work of Allen and collaborators [Allen *et al*., 2011], where the authors conclude that in *C. elegans* the trans-splicing to SLeI or SLeII are mechanistically separate and distinct phenomena. On the other hand, the fact that in the consulted database both SLe sequences were related to the same number of transcripts is not expected given the higher prevalence (of >80%) of SLeI trans-splicing in *C. elegans* as reported on Allen *et al*., 2011.

The analysis of transcripts bearing SLe sequences that are shared by different organisms from a same phylum indicated that these SLe sequences may control similar transcript repertories. The first SLe sequence that was analyzed is the one conserved in all dinoflagellate species in the database. As a result, from the retrieved matches, 133 transcripts remained after curation with almost half (63) shared by more than one species. The other analyzed SLe is common to several nematode species. In this case, from the retrieved matches, 158 transcripts remained after manual curation, 54 of which are shared by different species from the same phylum. In the latter case, I have excluded *C. elegans* from the search, since in this organism it is already known that SLeI and SLeII have different roles and are therefore added to different sets of transcripts [Allen *et al*., 2010]. Taken together, these results may indicate that identical SLes in different species from the same phylum are incorporated to the same transcripts. The observation that not all transcripts are common to all species harbouring identical SLe sequences could be a result of the restricted number of annotated transcripts deposited in publicly available databases. Sets of transcripts from different species that are regulated by the same SLes, have between 34% and 47% common elements, potentially indicating an overall tendency for each SLe to regulate specific transcripts, regardless of the species. If this is true, then a given SLe will always be related to a particular set of transcripts in all species (except maybe for species-specific genes), and therefore, different SLe from identical species would regulate different sets of transcripts. Indeed, Allen and collaborators have shown that, in *C. elegans*, spliced leader sequence SLeI is added either to monocistronic genes or to the first gene of a polycistron, while the SLeII sequence is added to internal genes of a polycistron. These two trans-splicing events are therefore mechanically unrelated and may be functionally different. This observation *per se* could account for the lack of overlap between the sets of transcripts regulated by each SLe sequence in this species and maybe other species that carry genes in an operon-like structure. Additionally, the authors show that some genes can actually be trans-spliced to both SLeI and SLeII and these are internal genes in the polycistron for which there are independent promoters. Results show that, for genes preferably trans-spliced to SLeII sequences, some level of SLeI addition may still be observed. According to the authors, this may be due to the 10-fold excess of SLeI in the cells in comparison to SLeII or it may be that SLeI is added to transcripts in a constitutive frequency, while SLeII addition could be more specific. Unfortunately, there are limited data on trans-spliced transcripts of other species with multiple SLe sequences and therefore it was not possible to reproduce here the same observation in other species, apart from the previously mentioned limited analysis of *H. vulgaris* transcripts, in which I found 30 SLeI-containing transcripts and 13 SLeII-containing transcripts, with no overlap among these sets.

I analyzed over 450 transcripts bearing SLe sequences from all species in the EXTENDED database except euglenozoans (since these species add SLe to all transcripts). Almost half of the investigated species (48%) contain transcripts shared with at least one other specie and almost one third (30%) contain transcripts shared with species from different phyla (136 transcripts, representing 32 unique sequences). Among the 32 transcripts represented in multiple phyla, most code for proteins which are conserved in all eukaryotes, and that perform basic functions in the cell, including involvement in ribosomal activity, cell structure, ATP synthesis, glucose metabolism, protein folding, antioxidant defense, DNA replication and translation. This may reflect a tendency of the SLTS mechanism to regulate ancient and conserved functions.

To thoroughly investigate the relationship between trans-spliced transcripts from different species and phyla, I have manually annotated and classified all 455 transcripts according to their biological function. In a surprising result, I have identified a dominant class of transcripts in each phylum, with no overlap among phyla. This could indicate that each phylum may have a preference to add the SLe to a specific gene category. Given the limited amount of available data, I did not perform analyses to investigate if this is true at the species level, although the previously presented results from *S. mansoni* suggest the SLTS mechanism not to be directed to any specific gene category. There is no overall tendency when I observe the set of transcripts from all phyla and this is expected since each phylum has a bias to a different class. Notably, I have not identified genes related to host-parasite interactions undergoing SLTS in the parasitic organisms. I consider this to be an expected observation if one considers that the SLTS mechanism was derived early in the eukaryotic lineage. On the other hand, photosynthesis is the major class of trans-spliced transcripts for the dinoflagellate species (from which only two *Perkinsus* species are not photosynthetic) and this is a phylum-specific class in the study because no species out of this phylum perform photosynthesis, suggesting species-specific adaptations of the SLTS mechanism or an independent origin in multiple phyla.

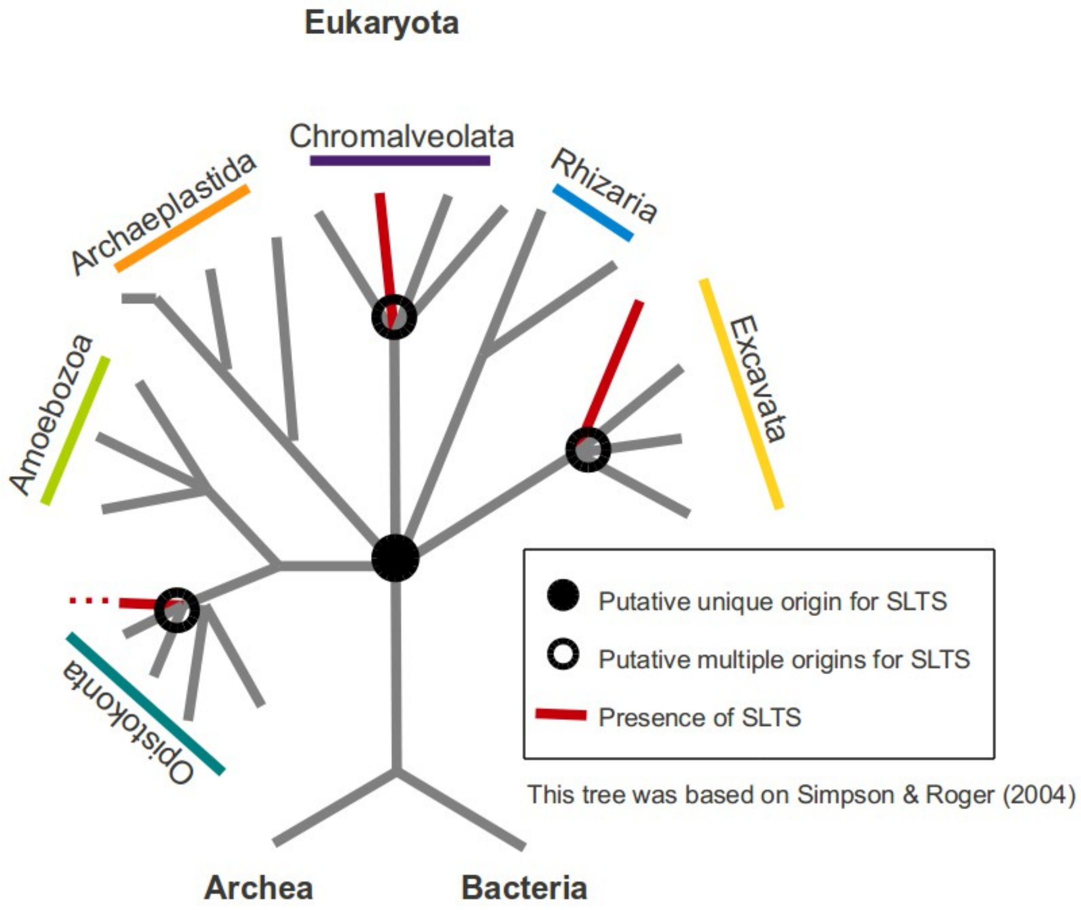
I cannot confirm that the observed bias to specific functional categories is not expected for a given phylum because I did not perform the same annotation and classification protocol for the entire set of transcripts for each species. For example, it is reasonable to expect that energy metabolism would be a broadly represented class in the transcriptome of most organisms and, accordingly, it is the major class among platyhelminthe transcripts that undergo trans-splicing. However, in nematodes there is a bias for transcripts involved with neurotransmission to undergo trans-splicing. This is not an expected result. Despite the evidence above, there remains a limited number of SLe-containing transcripts in public databases, therefore, no final conclusions can be reached.

As I have observed in the previous study regarding the SLTS mechanism in *S. mansoni*, I found no bias for any specific functional gene category to have transcripts undergoing SLTS. In a preliminary survey using a recently published *S. mansoni* SLe-containing cDNA dataset [Protasio *et al*., 2012], I have used GO annotations to assess whether specific gene categories could be found among the 1,411 SLe-containing transcripts. The most abundant biological processes are whole-cell processes (as cell differentiation, cell cycle, cell death, cell communication, cell proliferation, cell growth, cell recognition and cellular homeostasis), cellular component organization, transport, response to stimulus, signalling, protein function, protein expression, metabolic processes and organismal development, notably multicellular organismal development (which is the class with the highest number of related transcripts).

Taken these observations together, I can conclude that the SLTS mechanism does not regulate any specific biological process category. Nevertheless, some categories are slightly more represented than others and those categories are crucial processes for the metabolism of the cell and the organism. I can then speculate that the SLTS mechanism is of fundamental importance and that the disruption the mechanism should lead to serious consequences for the organism. This represents a notable result that places the SLTS mechanism in an important place for organismal development and survival.

A structural RNA analysis was performed to identify possible structural conservation despite the lack of sequence similarity across phyla. Structures were generated for the 30 SL gene sequences mentioned previously, and the results do not show a clear conservation, although some features may be observed. Because SL gene sequences vary in length, structure complexity is also diverse. An almost universal topological identity can be observed among the different SL structures, with few exceptions. SL RNA structures have three stem loops and a bifurcation point. This topology can be defined as a Y shaped structure. Although the bifurcation point location and stem-loop length may vary, the overall topology may be the only common aspect providing the SL RNA structure, which is a feature that is shared by all species.

Keeling and colleagues [Keeling *et al*., 2005] have published a consistent classification for eukaryotes that includes five different 'supergroups', namely excavates, rhizaria, unikonts, chromalveolates and plantae. As shown in the figure below (Figure 13), those phyla in which the SLTS mechanism was previously characterized are located in three of such supergroups: unikonts (rotifera, chordata, cnidaria, nematoda and platyhelminthe), excavates (euglenozoa), and chromalveolates (dinoflagellates). The widespread presence of the mechanism could place a unique origin in an ancestor common to all eukaryotes (or at least to these groups). The outcome of this hypothesis, if supported, is the previous existence of the SLTS mechanism in all eukaryotes, although the progressive loss of relative importance reduces the chance of identifying SLe-containing transcripts in more complex organisms. It may even be true that some species have completely lost the SLTS mechanism when more robust regulatory mechanisms emerged.



This tree was based on Simpson & Roger (2004)

Figure 13: A modified representation of the eukaryotic tree as published by Keeling and colleagues. The SLTS mechanism was found to be present in three out of the five different supergroups presented by the authors.

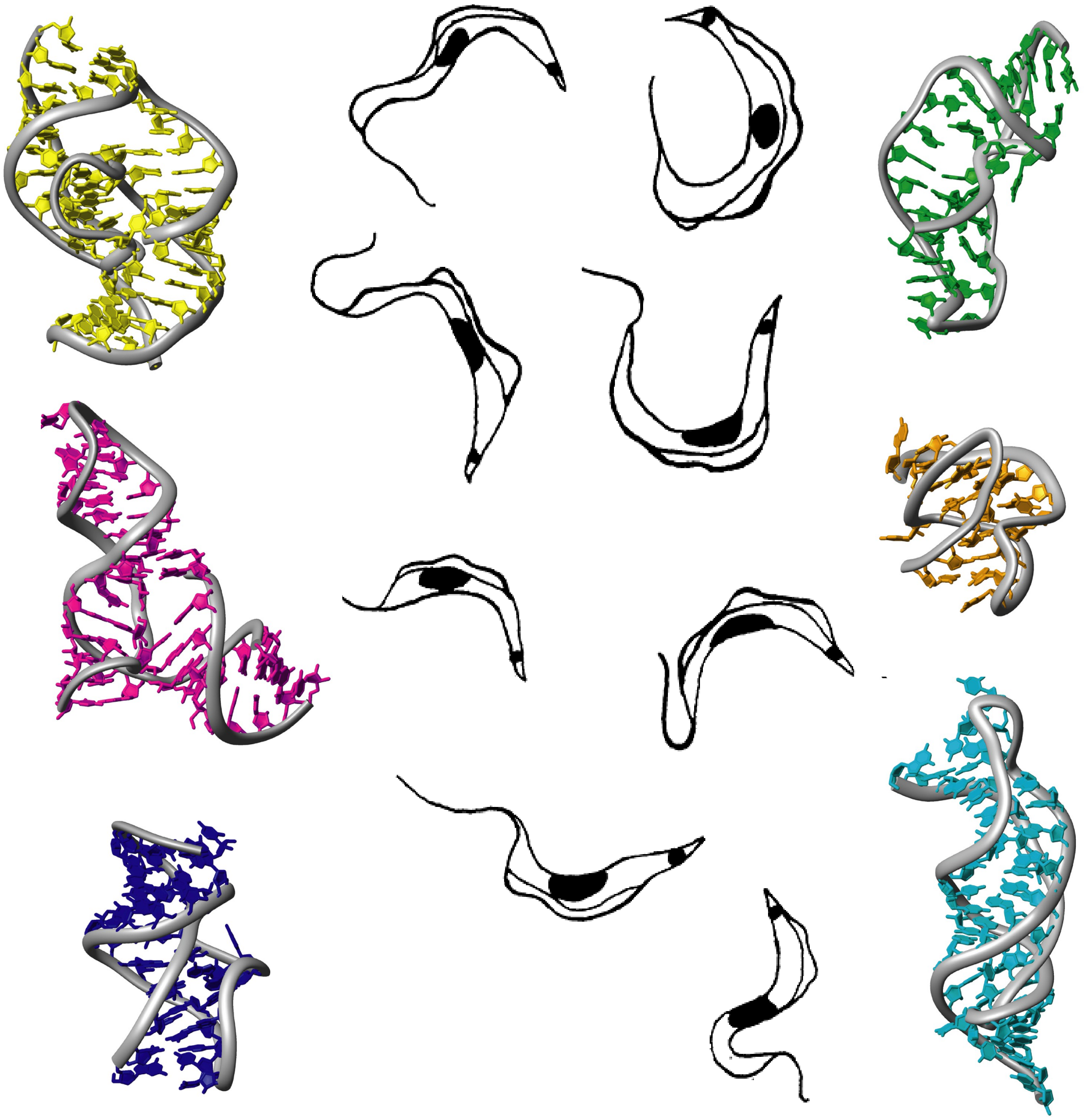
VI. CONCLUSIONS

In this work the *S. mansoni* SL-sequence containing transcripts were demonstrated not to be associated with any specific gene category, subcellular localization or life-cycle stages. The protein length, number of exons and exon length of the SL-containing transcripts were not different when compared to the entire set of *S. mansoni* transcripts. The fact that a wide range of different genes is regulated by SLTS suggests that this mechanism plays an important role in controlling the expression levels of different proteins, as well as the protein repertories present in the different life-cycle stages and under distinct environmental conditions. Disruption of the SL trans-splicing mechanism in *S. mansoni* sporocysts by RNAi produced a reduction in larval size, providing evidence for the importance of this mechanism for the organism development/growth and suggesting a crucial role in the regulation of metabolic processes by SL trans-splicing.

Analyses of conserved features in SL sequences of ~150 species have revealed a close relationship between sequences from species of the same phylum and sequences from species of different phyla tend to be more divergent. Regarding the overall features of the SL molecule, I report that (i) all SLe sequences have a WWG pattern at the 3' end; (ii) a Sm binding site is always present, either in the exon or in the intron of the SL gene; (iii) the sequences are 22-25 or 36-39 nucleotides long; (iv) a SLe TTT pattern is present in almost all sequences; and (v) the RNA structure is Y-shaped, bearing three stem-loops and a bifurcation point. Another important contribution of this study is the observation of a tendency in which different SLe sequences in a given species control different transcripts and identical SLe sequences in different species control identical transcripts. I hypothesize that each SLe sequence is always related to a given set of transcripts and, therefore, the expression of the respective protein repertories could be switched on and off according to the presence of a SLe sequence.

CHAPTER 2

Annotation of new ncRNAs in the genome of *Trypanosoma cruzi* CL Brener strain



Drawings of *T. cruzi* parasites naturally infecting a rodent host [modified from Mello and Teixeira, 1977] and structures of RNA pseudoknots [modified from Staple and Butcher, 2005].

I. INTRODUCTION

A. TRYPANOSOMA CRUZI*: THE CAUSATIVE AGENT OF CHAGAS DISEASE*

Trypanosoma cruzi, the etiologic agent of Chagas disease (also known as American trypanosomiasis), is a member of the Trypanosomatidae family and the Kinetoplastidae order. This tropical parasite was first described by the Brazilian physician Carlos Ribeiro Justiniano Chagas in 1909 [Chagas, 1909]. Along the last decade, different aspects of the parasite unique biology have been unraveled. *T. cruzi* is a digenetic parasite that differentiates into several evolutive forms along its complex life-cycle, which takes place in two different hosts, an intermediate insect host and a definitive mammalian host. In the mammalian host, during its initial phase, the Chagas disease is characterized by high parasitemia, local manifestations and variable symptoms that include fever, edema, headaches, muscle pain, difficulty to swallow and breathe, inflammation of lymph nodes (benign lymphadenopathy), hepatomegalia and splenomegalia. Indicative signs of the disease include a localized inflammatory response (chagoma) which occurs at the site of parasite entrance and can persist for up to 8 weeks [reviewed by Berger, 2014] and the unilateral edema of the eyelid (the Romaña sign) which is observed when contamination happens through the ocular mucosa. The infection then evolves to an asymptomatic chronic phase with a decrease in the levels of parasitemia, which may last for decades. The chronic phase is often unnoticed by the patient, until symptoms evolve and typically characterize either the cardiac, digestive, cardio-digestive or neurological form of the disease. Most symptoms are derived from either cardiac lesions, leading to alterations in the heart volume (cardiomegaly) and contraction rhythm, or digestive tract lesions, manifesting as esophageal and stomach swelling. The digestive form is prevalent in the Amazon region and specially found in Argentina, Brazil, Chile and Bolivia. In other countries, as Panama and Venezuela, the chagasic infection is usually manifested as the cardiac form. This geographic distribution Chagas disease clinical symptoms have been proposed to be related to the diversity of parasite strains found in each endemic area, thus suggesting a differential tropism of different strains to a specific organ or system [Rassi Jr *et al*., 2010; Macedo *et al*., 2004].

The main transmission mechanism of *T. cruzi* is through the bite of insects from the Triatominae subfamily, which comprises around 130 species, from which 12 are known to act as intermediate hosts and, therefore, vectors [Schofield, 1994]. During the day, the insects usually inhabit cracks in house walls made of clay and wood, mainly found in the poor regions of endemic areas. During the night, these haemophagus bugs leave their daylight lurking place to feed from the blood of homeothermal mammals. Given their habit of preferentially biting the face (often the lips) of their victims, *Triatoma* species are often referred to as *kissing bugs*. During or after feeding, which takes around 10 minutes, the insect usually defecates and, when infected, releases metacyclic trypomastigote parasites in feces. The insect saliva contains anesthetic substances [e.g. Dan *et al*., 1999] and, therefore the bite is usually unnoticed, but it may trigger itchiness and swelling at the place of proboscis (the tubular feeding structure) introduction. Insect feces containing active parasites then penetrate the scratched skin or even the intact mucosa of the eye or mouth. At the entry site, trypomastigote parasites invade nearby cells, multiply and differentiate into amastigotes, which are the forms responsible for the pathogenic symptoms of Chagas disease at the chronic stage. Amastigotes then successively differentiate by binary fission, bursting the cells and assuming the trypomastigote blood form, which is released to the interstitial space. Trypomastigotes have the capacity to invade other cells from virtually any tissue or organ, thus initiating a new cycle of asexual reproduction. The *T. cruzi* circulating in the blood is then ingested by the vector insects, which triggers transformation to the epimastigote stage, induced by the low nutritional state in the midgut of the triatomine bug. Lastly, the epimastigotes divide by binary fission originating infective metacyclic trypomastigotes, which migrate through the midgut and are eliminated in the insect feces during or after blood feeding [reviewed by De Souza, 2002; Duszenko *et al.*, 2011] (Figure 1).

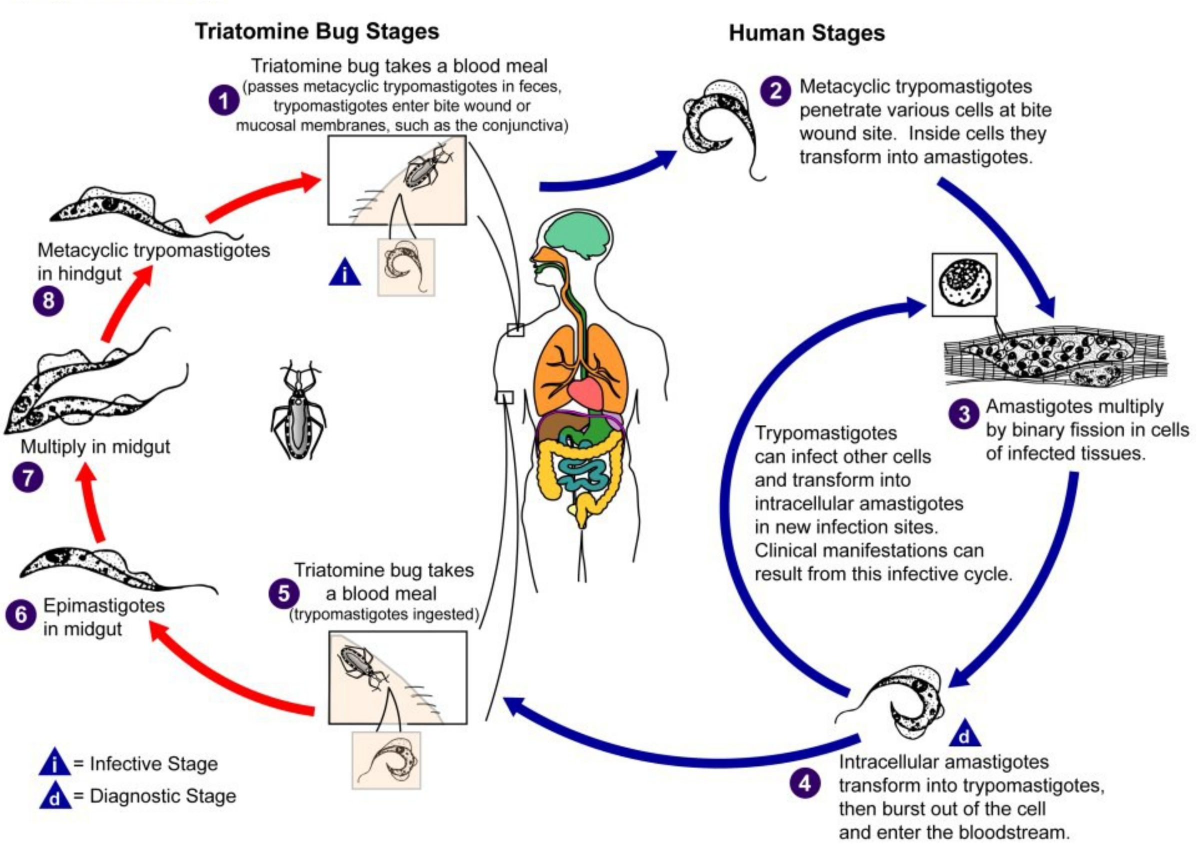

Triatomine Bug Stages
Human Stages
1 Triatomine bug takes a blood meal (passes metacyclic trypomastigotes in feces, trypomastigotes enter bite wound or mucosal membranes, such as the conjunctiva)
2 Metacyclic trypomastigotes penetrate various cells at bite wound site. Inside cells they transform into amastigotes.
3 Amastigotes multiply by binary fission in cells of infected tissues.
Trypomastigotes can infect other cells and transform into intracellular amastigotes in new infection sites. Clinical manifestations can result from this infective cycle.
4 Intracellular amastigotes transform into trypomastigotes, then burst out of the cell and enter the bloodstream.
5 Triatomine bug takes a blood meal (trypomastigotes ingested)
6 Epimastigotes in midgut
7 Multiply in midgut
8 Metacyclic trypomastigotes in hindgut
i
d
i = Infective Stage
d = Diagnostic Stage

Figure 1: The Chagas disease and *T. cruzi* life-cycle stages. Figure reproduced from the CDC (http://www.cdc.gov/parasites/chagas/biology.html).

The Chagas disease is endemic in 21 Latin American countries and, due to the increasing migration of infected individuals especially to Europe and the United States of America (USA), it is now a threat for the entire world. The World Health Organization (WHO) currently estimates that between 7 and 8 million people are infected worldwide [WHO fact sheet 340, 2014], although other reports suggest this number may be underestimated (e.g. Coura & Borges-Pereira, 2010 report up to 17 million human infections in Latin America). From the chronically infected individuals, around 30% develop cardiac alterations and up to 10% develop gastrointestinal tract diseases (neurological or mixed alterations are also reported) requiring specific treatment [WHO fact sheet N° 340, 2014; Taylor & Bestetti, 2009]. In Colombia alone, the annual cost of treatment was of approximately US$267 million in 2008 and around US$5 million are annually spent for insect control, which is currently the most effective form of prevention in Latin America [WHO fact sheet N° 340, 2014]. Chagas disease treatment has not significantly evolved since the first introduction of efficient drugs in the 1970 decade [Prata *et al*, 1975; Ferreira, 1976; Prata *et al*., 1977]. As there is no vaccine against *T. cruzi*, the treatment is mainly restricted to the use of drugs such as nifurtimox (which was discontinued) and the first-line drug benzonidazol, both with reported limited efficacy and severe side effects [Senkovich *et al*., 2005]. It is therefore of great importance to develop new therapeutic strategies to prevent and/or treat the disease.

After the undeniable breakthroughs in the last decade regarding the control of *T. cruzi* vectorial and transfusional transmission, other so far irrelevant forms of transmission have became epidemiologically relevant. The emerging pathways include oral, congenital and through organ transplants. Surprisingly, apart from being under control in endemic areas where the insect vector was heavily targeted, transmission rates are increasingly higher in non-endemic countries, which were unprepared for this parasite introduction. This modern dynamics of disease occurrence is mostly influenced by the emerging alternative courses of transmission, subsided by the previously mentioned migratory movement of infected individuals. As a result, the Chagas disease is now present worldwide, with over 300,00 cases estimated in the USA [Bern & Montgomery, 2009], between 60,000 and 120,000 in the European continent (specially in Spain and Italy) [Coura & Vinas, 2010; Angheben *et al*., 2011; Requena-Méndez *et al*., 2014], and over 5,500, 3,000 and 1,500 cases reported in Canada, Japan and Australia, respectively [Coura & Vinas, 2010] (Figure 2A). Most of these patients are at the chronic stage of the disease, but a few cases of acute phase patients were reported. In response, several

European countries have implemented strict tests to search for contaminated units in blood banks and organ transplants among other policies to contain the spread of Chagas disease [reviewed by Requena-Méndez *et al*., 2014].

In 2010, during the 63rd World Health Assembly (WHA), a resolution was approved highlighting the epidemiological and medical relevance of this disease, both in endemic and non-endemic countries [WHA resolution N° 63.20]. The resolution requires the adoption of new strategies and the reinforcement of successful strategies to decrease or eliminate transmission and improve prevention, diagnosis and treatment both in endemic and non-endemic areas. In addition, countries are encouraged to subside translational research aiming the control of transmission by domestic insect vectors, the assessment of alternative safer and cheaper treatment options, the development of a test of cure, the reduction of late complication risks in infected individuals, the establishment of early detection systems and the optimization of examination of blood for transfusion in endemic areas [WHA resolution N° 63.20].

In Brazil, after significantly reducing the vectorial and transfusional transmission of *T. cruzi*, the reported number of new cases has drastically decreased in the last decade [e.g. Ostermayer *et al*., 2011]. In fact, at June 2006, the Pan-American Health Organization (PAHO) has granted an international certificate attesting the elimination of Chagas disease transmission by *Triatoma infestans*, the main vector in the country. This was the outcome of a conjoint effort including other South American countries such as Argentina, Bolivia, Chile, Paraguay and Uruguay to eliminate this insect [Dias, 2007]. Although important, this only reflects the current interruption of transmission by this specific species and does not classify the disease as eradicated, thus requiring uninterrupted surveillance [reviewed by Dias *et al*., 2002]. Recently, non-classical transmission routes emerged and oral transmission, through contaminated food, has been the most prevalent form of Chagas disease transmission in Brazil [Dias *et al*., 2011]. This route certainly contributes to recent reports of 2 to 3 million infected individuals and 5 to 6 thousand annual deaths still observed in the country [Martins-Melo *et al*., 2012 and Martins-Melo *et al*., 2014], although this number is much lower then in earlier decades. In a less satisfactory estimative, a comprehensive literature survey reported 4.6 million cases of *T. cruzi* infection in Brazil [Martins-Melo *et al*., 2014]. The figure below (Figure 2B) is reproduced from this publication, where the authors suggest, based on the current literature, the pattern of *T. cruzi* dispersion in Brazil. Despite the attention it has been given, Chagas disease remains as a priority for the public health systems and, in the socio-political scenario of the American continent, the disease still requires close attention from the governmental organs responsible for the maintenance of health and epidemiological surveillance. The study of *T. cruzi* and its molecular aspects is therefore of extreme importance in Brazil, as an endemic area of the Chagas disease.

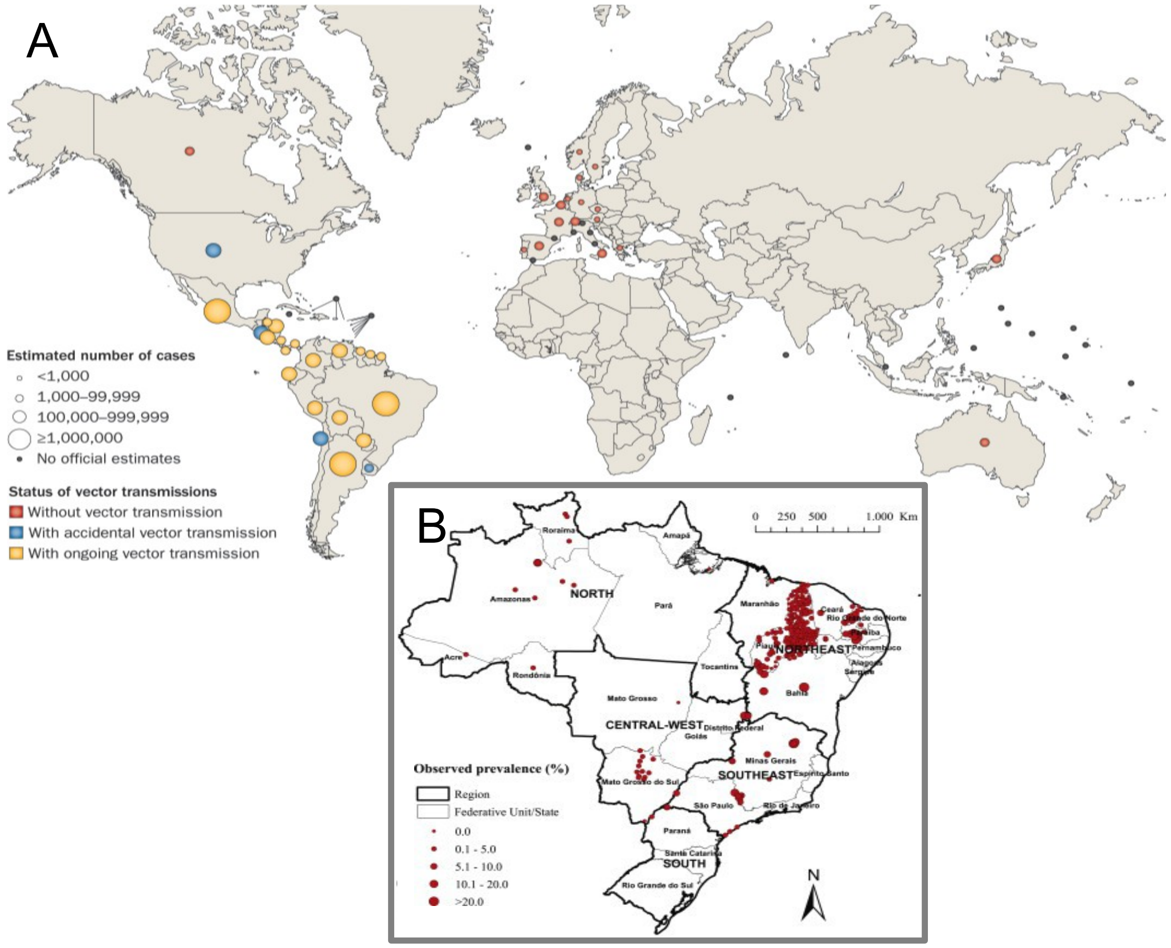


Figure 2: Distribution of *T. cruzi* infection for the entire world and for Brazil. **(A):** Worldwide distribution of *T. cruzi* infection based on official estimates from 2006-2009 [reproduced from the First WHO report on neglected tropical diseases] **(B):** Spatial distribution of observed Chagas disease prevalence in Brazil according to population-based surveys [reproduced from Martins-Melo et al. 2014].

Microscopically, all *T. cruzi* strains are flagellate protozoan with a single mitochondria and a unique additional organelle named kinetoplast, which is differentially positioned within the cell along the parasite life-cycle. From the study of parasites isolated from both humans and sylvatic hosts, nearly 80 different strains have been described (Medline 97179491), differing in geographic location, host specificity and tissue tropism, among other characteristics. *T. cruzi* is therefore a very diverse species and different criteria for classification can be used to cluster the strains in subspecies with similar features. After much divergency among the scientific community, during the Commemorative Symposium (*Simpósio Internacional Comemorativo dos 100 anos da Descoberta da Doença de Chagas* - Búzios/RJ, August of 2009) that was held 100 years after the disease was first described by Dr. Chagas, a consensus was reached to classify the known strains into six discrete typing units (DTU), numbered from TcI to TcVI, each comprising multiple strains with similar biological characteristics (tested by genetic, molecular, biochemical or immunological markers and ecoepidemiologic factors) [Zingales *et al*., 2009; revised in Zingales *et al*., 2012]. These DTUs vary in geographic location, ecological niche, host and preferential vector [Zingales *et al*., 2012] and, despite the lack of conclusive proof, evidences suggest an association between DTUs and the chronic manifestation of the disease. For example, the digestive form is more prevalent in areas were TcII, TcV and TcVI are predominant (including Argentina, Bolivia, Brazil and Chile) and rare in regions where TcI and TcIV are prevalent. [Zingales *et al*., 2012] As mentioned, *T. cruzi* main reproduction strategy is asexual. Nevertheless, isolated events involving the exchange of genetic material between different DTUs may occur and studies show high heterozygosity in TcV and TcVI isolates, thus suggesting that these may be hybrids of TcII and TcIII [Sturm *et al*., 2003; Sturm & Campbell 2010]. The current most accepted phylogenetic relation between the six DTUs is depicted in the figure below, which was based on data from Sturm & Campbell 2010 (Figure 3A) and Zingales *et al*. 2012 (Figure 3A and Figure 3B).

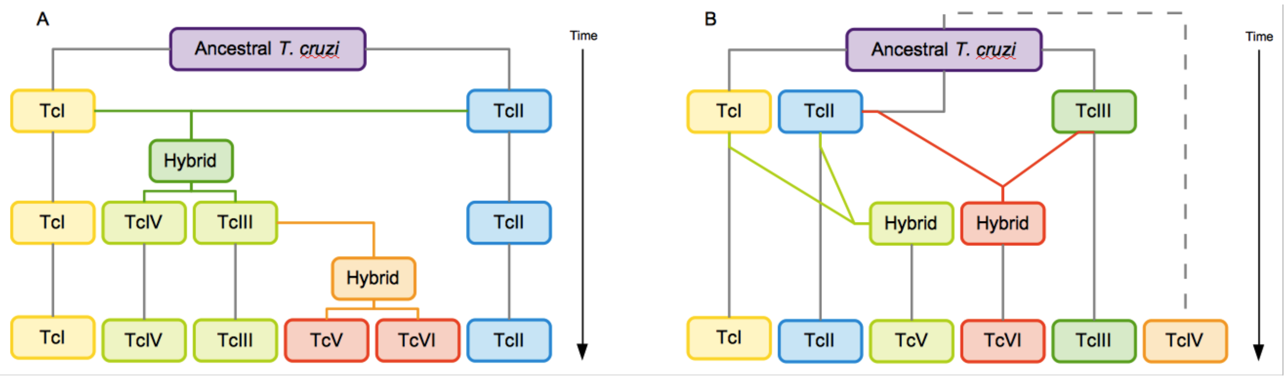


Figure 3: Two-hybridization and three ancestor models for the *T. cruzi* population structure. Two-hybridization model for the *T. cruzi* population structure as observed in the present. **(B):** Three ancestor model as presented by Zingales *et al*., 2012. Mitochondrial clades are not represented in this figure and colors are random.

In 2005 El-Sayed and collaborators published the sequenced genome of the CL Brener strain, a hybrid lineage comprising two different haplotypes [El-Sayed *et al*., 2005]. The authors have used the Whole Genome Shotgun (WGS) strategy for sequencing but, in contrast to other trypanosomatid genomes (e.g. *Leishmania* and *T. brucei*), the *T. cruzi* genome was first published as an unassembled set of 5,489 scaffolds and 8,740 contigs. Only in 2009, four years after its first release, the genome was partially assembled in 82 chromossomes (41 from each haplotype) by Weatherly and collaborators with the aid of Bacterial Artificial Chromosome (BAC) libraries and a synteny map with the *T. brucei* genome [Weatherly *et al*., 2009]. There were five main reasons for selecting CL Brener as the reference strain for the genome project: (1) it was isolated from the domestic vector *T. infestans*; (2) its infection dynamics in mice was well characterized; (3) its preferential tropism for heart and muscle was known; (4) its acute phase was well described both in mice and accidentally infected human; (5) it is susceptible to the drugs commonly used in the Chagas disease treatment [Zingales *et al*., 1997]. Additionally, this strain was well characterized both biologically and experimentally, thus enabling comparisons with data from the EST project, which was being carried out at that moment [Brandão *et al*., 1997; Verdun *et al*., 1998]. The annotation of CL Brener genome has indicated the existence of 22,570 protein-coding genes, form which 6,159 are provenient from the Esmeraldo-like haplotype (TcII ancestral lineage, hereby CLBe), 6,043 from the non-Esmeraldo-like (TcIII ancestral lineage, hereby CLBne) and 10,368 coding sequences were not related to any particular haplotype (hereby CLBmod). Based on literature data, the degree of similarity to known protein sequences and the presence of functional or characteristic domains, the authors were able to attribute function to only 50.8% of these predicted sequences [El-Sayed *et al*., 2005]. Nevertheless, nearly 6,000 proteins remain unannotated and further characterization studies will be needed. Over half the CL Brener genome corresponds to repetitive elements, retrotransposons and subtelomeric repeats, thus hampering the identification of protein-coding genes [El-Sayed *et al*., 2005]. The sequencing of this model *T. cruzi* genome and its comparison to the known genomes of *T. brucei* [Berriman *et al*., 2005] and *L. major* [Ivens *et al*., 2005] has expanded the knowledge about the genomic content and gene expression control in these organisms, although a large fraction of the molecular features of these parasites remain unknown.

Since the CL Brener hybrid genome was published, individual research groups have finished or are currently performing the sequencing of genomes from strains of different DTUs. From these, there are currently 6 different *T. cruzi* genomes available in the TriTrypDB website [Aslett *et al*., 2010], each from a different strain accounting for four out of six DTUs being represented, in addition to a TcBat representative. Among these initiatives, it is worth to highlight the recent sequencing of the complete genome of the Sylvio clone X10 [Franzén *et al*., 2011a]. Interestingly, among the genes identified in CL

Brener, only six genes were not found in the Sylvio strain genome, although multiple copy gene families as mucins, mucin-associated proteins (MASP) and GP63 are present in lower numbers. The Sylvio X10 (DTU TcI isolated from *H. sapiens* at Para, Brazil) and the CL Brener (DTU Tc VI isolated from *Triatoma infestans* at Rio Grande do Sul, Brazil) strains are currently the best annotated genomes, although this annotation is still far from covering the entire genome. Other available genomes include the Esmeraldo strain (DTU TcII isolated from *H. sapiens* at Bahia, Brazil), JR cl4 strain (DTU TcI isolated from *H. sapiens* at Anzoategui, Venezuela), Marinkellei B7 strain (DTU TcBat isolated from Phyllostomus discolor at São Felipe, Brazil) and Tulahuen cl2 strain (DTU TcVI isolated from *H. sapiens* at Tulahuen, Chile), although from these four genomes only the Marinkellei strain is annotated [strain information was obtained from Lewis *et al*., 2009; Subileau *et al*., 2009; Zingales *et al*., 2009]. The tables below were generated after a survey in the TriTrypDB in which the six sequenced strains were assessed regarding different features (Table 1 and Table 2).

***T. cruzi* strain**	**Source**	**Version**	**Mega base pairs**	**Genome annotation**	**All genes**	**Protein coding**	**Non protein coding**	**Pseudogenes**
Tulahuen (cl. 2)	GenBank	26/06/13	83.51	No	Null	Null	Null	Null
Esmeraldo	GenBank	17/01/13	38.08	No	Null	Null	Null	Null
JR (cl. 4)	GenBank	17/01/13	41.48	No	Null	Null	Null	Null
Sylvio (X10/1)	GenBank	02/10/12	38.59	Yes	10947	10876	71	30
Marinkellei (B7)	Franzén	-	38.65	Yes	10282	10228	54	0
CLBne	GeneDB	16/01/13	32.53	Yes	11109	10834	275	1448
CLBe	GeneDB	16/01/13	32.53	Yes	10600	10342	258	1300
CLBmod	GeneDB	16/01/13	36.03	Yes	3397	2135	1262	890

Table 1: General information for the six *T. cruzi* strains with sequenced genomes available at the NCBI.

Genomic Elements	**CLBmod**	**CLBe**	**CLBne**	**Sylvio**	**Marinkellei**
Proteins	2135 (62.85%)	10342 (97.57%)	10834 (97.52%)	10876 (99.35%)	10228 (99.47%)
Non-hypothetical	*1743 (81.64%)*	*4909 (47.47%)*	*5158 (47.61%)*	*5526 (50.81%)*	*5194 (50.78%)*
Hypothetical	*392 (18.36%)*	*5433 (52.53%)*	*5676 (52.39%)*	*5350 (49.19%)*	*5034 (49.22%)*
Pseudogenes	887 (41.54%)	1289 (12.46%)	1442 (13.31%)	-	-
tRNAs	5 (0.15%)	55 (0.52%)	55 (0.49%)	71 (0.64%)	54 (0.52%)
rRNAs	200 (5.89%)	9 (0.08%)	10 (0.09%)	-	-
snRNAs	3 (0.09%)	8 (0.07%)	8 (0.07%)	-	-
snoRNAs	1054 (31.03%)	186 (1.75%)	202 (1.82%)	-	-
Total	3397	10600	11109	10947	10282

Table 2: Information on the gene annotation of *T. cruzi* strains with genomes available at the NCBI.

B. NON CODING RNAs, THEIR BIOLOGICAL FUNCTIONS AND SEQUENCING

Since the idea of an RNA world was first introduced by Walter Gilbert [Gilbert, 1986] nearly three decades ago, we have revisited essential concepts of molecular biology and the so-called *central dogma* proposed by Crick in the 1950s [reviewed by Crick, 1970]. Nowadays, the non-coding RNAs (ncRNA or fRNA, from functional RNAs) are known to have crucial roles in virtually all biological processes in the cell and the constantly growing list of ncRNA classes (Figure 4) and functions related to such molecules has influenced fundamental concepts, as the classical definition of a *gene*. To illustrate the impact of ncRNAs in modern Molecular Biology, a text search on Pubmed (the database for Biomedical literature, currently containing 24 million articles) for the term “non coding rna” has returned nearly 120 thousand scientific papers, which were published between 1970 (when the first UTR was described by Cory *et al*.) and 2014. Additionally the table below (Table 3) presents some statistics regarding the number of RNA sequences deposited in each specialized database. It is worth to mention that these databases are not mutually exclusive and the broad collections such as RNA Central [RNAcentral Consortium, 2014] can encompass more specific sets.

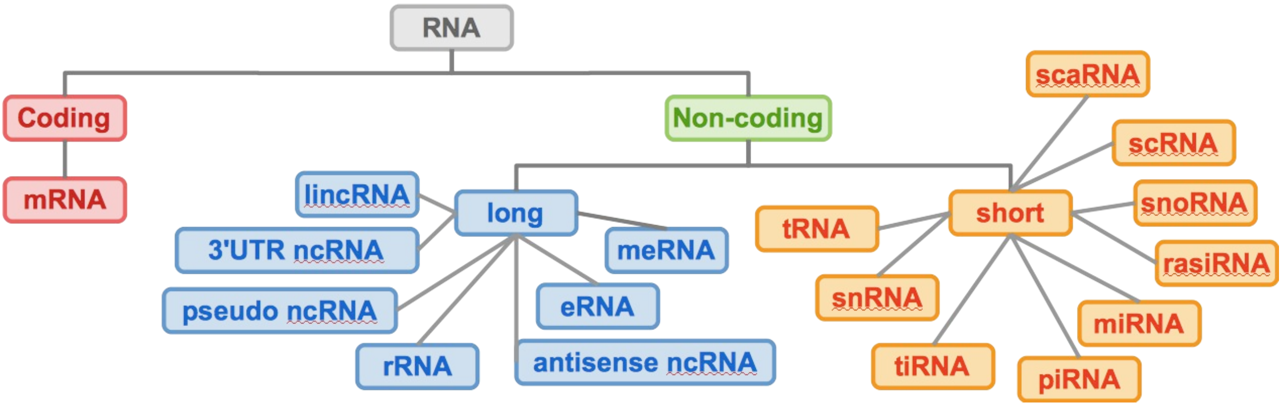


Figure 4: A proposed RNA classification scheme. Coding RNA (mRNA) is in red and non-coding RNAs are in orange (< 200 base pairs) and blue (> 200 base pairs) [data from the International Nucleotide Sequence Database Collaboration (INSDC) website and Kowalczyk et al., 2012]. The graph depicts some of the known

ncRNA classes, such as lincRNAs (long intergenic ncRNAs), 3' UTR ncRNAs, pseudo ncRNAs (RNAs transcribed from pseudo genes), rRNAs, antisense RNAs (ncRNAs complementary to protein-coding sequences), eRNAs (enhancer RNAs), meRNAs (multiexonic RNAs), tRNAs, snRNAs, tiRNAs (tiny RNAs), piRNAs (Piwi-interacting RNAs), miRNAs (micro RNAs), rasiRNAs (small interfering RNA derived from the transcript of a repetitive element), snoRNAs (small nucleolar RNAs), scRNAs (small cytoplasmic RNAs) and scaRNAs (small Cajal body-specific RNAs).

Database Name	**Number of Sequences***	**Brief Description**
RNA Central	~8 million	An unified database for ncRNA sequences
Rfam	~2.5 million	Information on ncRNA families and other structured RNAs
gtRNAdb	~75 thousand	Genomic tRNAs (whole genome searches with tRNAscan)
RefSeq	~45 thousand**	Comprehensive, non-redundant and well annotated RNAs
miRBase	~30 thousand	Published microRNA sequences and associated annotation
SSU rRNAs DB	~4 thousand	A joint set of two ribosomal databases
tmRNA website	~2 thousand	Non-redundant transfer messenger (tm)RNA sequences
SRP database	~1 thousand***	Annotated sequences of signal recognition particles (SRP)

* Unless otherwise stated, the databases contain redundant entries

** The mRNAs were excluded from the total set of transcripts and all other entries were considered

*** Only SRP RNA sequences were considered

Table 3: Number of deposited entries for some ncRNA-containing public databases. A survey was performed in November 2014 to retrieve the number of ncRNA entries in each specialized database.

NcRNAs are functional molecules that play a large variety of roles in the cell, for example orchestrating the effector protein molecules. With more recent analyses of the human genome, researches have concluded that that non-coding regions must have a significant evolutionary importance. In this context, the ENCODE project [encyclopedia of DNA elements; ENCODE Project Consortium, 2004] has revealed that ~75% of the human genome has potential to produce functional transcripts [Djebali *et al*., 2012] and only a small percentage of the transcribed RNA should be protein-coding. A significant portion of the non-coding genome of humans is actually transcribed as ncRNAs with numerous functions [see references within Rinn & Chang, 2012 and Clark *et al*., 2013]. For metazoan genomes, it is known that large sets of RNAs of different types are produced, the majority of which from intronic and intergenic regions of the DNA. While some of the resulting

RNAs are long (lncRNA, >200 base pairs), other are small (< 200 base pairs). Small ncRNAs include siRNAs, miRNAs, piRNAs associated to protein complexes involving the Piwi protein, tiRNAs derived from translation start sites, splicing sites RNAs (spliRNA), RNAs associated to promoters (PARS), and even small structural RNAs that are long known and participate on ribonucleoproteic complexes with processing functions (e.g. snRNA and snoRNA), among others (Figure 4). Considering the variety of roles that ncRNAs can perform on cell networks, it is no surprise that they have been implicated on several human syndromes and diseases. Genome-wide association studies (GWAS) revealed that only a small percentage of human single nucleotide polymorphisms (SNP) associated to diseases or specific traits are located on protein-coding regions, while the vast majority (88%) is located on either intronic (45%) or intergenic (43%) regions [Hindorff *et al*., 2009].

LncRNAs are currently defined as transcripts larger than 200 nucleotides, with no obvious protein-coding function [Rinn & Chang, 2012]. It is worth noting that the set of lncRNAs is composed of several classes of transcripts, such as: enhancer RNAs (eRNA), antisense RNAs (transcribed as the complement for a mRNA), lincRNAs (long intergenic ncRNAs), among others (Figure 4). Recent genome annotation data suggest that the number of lncRNA genes should exceed the number of protein-coding genes in the human genome [Kapranov *et al*., 2007]. Although lncRNA genes are under lower selective pressures than protein-coding genes, sequence analyses show that lncRNA genes are under higher selective pressures than ancient repetitive sequences, which are under neutral selection [Derrien *et al*., 2012]. In addition, comparisons between lncRNAs of zebrafish and mammals revealed that small portions of conserved sequences are functionally important, even in the absence of a thorough sequence conservation [Ulitsky *et al*., 2011].

In parallel to what I known for protein evolution, it is expected that functionally related RNAs should present structural conservation in a higher degree than sequence conservation. In this sense, Bioinformatics tools have been developed to assess and explore structures. One of the best known of such tools is a collection of algorithms implemented in a web server to generate and analyze RNA structures, named the Vienna package [Lorenz *et al*., 2011]. Among the programs that comprise this package, the most used is the secondary structure prediction algorithm named RNAfold. Given the evolutionary importance of structural conservation, when performing homology searches it is desirable to assess both the primary and the secondary structure of RNAs. In this context, INFERNAL [Nawrocki *et al*., 2009; Nawrocki & Eddy, 2013] is a tool that builds probabilistic profiles for both the primary and secondary structures of RNA families and further uses these profiles (or covariance models) to search for new members of the RNA families in sequence inputs. INFERNAL in association with the Rfam database is widely used to annotate RNAs in genome sequences. Additionally in this field, Dr. Martin Smith and his collaborators are currently working on the development of a program and systematic bioinformatics approach for the identification of common RNA structures within a subset of sequences, considering an ensemble of sub-optimal base pairings. When considering sub-optimal base pairings, the algorithm will significantly improve the physical realism of RNA structure predictions (unpublished results).

It has been more than a decade since Sean Eddy raised questions such as: How many ncRNAs exist on a given genome? How important are these? Which cell functions do these RNAs perform? [Eddy, 2001]. To answer these questions, large-scale genomic and transcriptomic data of several organisms should be analyzed through computational tools and biochemical techniques, in order to identify and functionally characterize the set of ncRNAs. Therefore, there is a great need for the application of *in silico* and/or *in vitro* approaches to the identify and classify ncRNAs in previously sequenced genomes. In this sense, the development of next generation DNA sequencing technologies and its application to the study of transcriptomes have yielded the high-throughput RNA sequencing (RNA-Seq) method [reviewed by Wang *et al*., 2009], a revolutionary shotgun sequencing technique that allows one to analyze whole transcriptomes. Since it was first applied to model cells and organisms in 2008 [e.g. Faulkner *et al*., 2008; Morin *et al*., 2008; Mortazavi *et al*., 2008 and Nagalakshmi *et al*., 2008] until the current days, the RNA-Seq methodology has been responsible for major increases in the number of known transcribed elements. To store and organize publicly available RNA-Seq derived sequences, the NCBI (www.ncbi.nlm.nih.gov) has created a database named Sequence Read Archive (SRA [Leinonen *et al*., 2011]), which is the primary archive for high-throughput sequencing data and is part of the international partnership of archives at the NCBI, the European Bioinformatics Institute (EBI) and the DNA Database of Japan (DDBJ) [Wheeler *et al*., 2008]. According to the authors, the SRA was specifically designed to meet the challenges presented by massively parallel sequencing technologies, such as RNA-Seq. Since the original and provisional SRA was deployed in 2007 until mid 2014 the number of deposited bases in the database has increased from $2x10^{10}$ to over $2x10^{15}$, characterizing this as a highly valuable resource from which biological data can be retrieved to supply further analyses. In recent years related *Trypanosoma* species as *Trypanosoma brucei* and *Trypanosoma vivax* have being subjected to RNA-Seq studies [e.g. Greif *et al*., 2013 and the articles reviewed in Main *et al*., 2012], mainly focused on the protein-coding portion of the transcriptome. The present work is thus devoted to the identification and characterization of ncRNAs in *T. cruzi* that may play crucial roles in the development and biology of this parasite.

II. AIMS

Main Goal

To identify known and novel ncRNA candidates in the genome of *T. cruzi* CL Brener strain via similarity search.

Specific Goals

- To survey the parasite genome in search for putative ncRNA candidates using sequence-based and structure-based similarity search tools.
- To explore the genomic location of each ncRNA candidate.
- To assess the expression of such ncRNA candidates in publicly available RNA-Seq data.

III. MATERIALS AND METHODS

A. IDENTIFYING THE NCRNA CANDIDATES IN THE COMPLETE GENOME OF T. CRUZI *(CL BRENER CLONE)*

A.1 The CL Brener Genome

As previously mentioned, the *T. cruzi* CL Brener is a hybrid strain and as a consequence its genome is divided in three sub-genomes, which were individually obtained from the data repository of the TriTrypdb webserver (http://tritrypdb.org/common/downloads/release-8.0). The specific path to each FASTA (DNA sequence) and GFF (genome annotation) file is given in the table below (Table 4). The individual sub-genome files were downloaded and concatenated in single files for all subsequent analyses.

Paths to the Genome FASTA Files
TcruziCLBrener/fasta/data/TriTrypDB-8.0_TcruziCLBrener_Genome.fasta
TcruziCLBrenerNon-Esmeraldo-like/fasta/data/TriTrypDB-8.0_TcruziCLBrenerNon-Esmeraldo-like_Genome.fasta
TcruziCLBrenerEsmeraldo-like/fasta/data/TriTrypDB-8.0_TcruziCLBrenerEsmeraldo-like_Genome.fasta
Paths to the Genome GFF Files
TcruziCLBrener/gff/data/TriTrypDB-8.0_TcruziCLBrener.gff
TcruziCLBrenerNon-Esmeraldo-like/gff/data/TriTrypDB-8.0_TcruziCLBrenerNon-Esmeraldo-like.gff
TcruziCLBrenerEsmeraldo-like/gff/data/TriTrypDB-8.0_TcruziCLBrenerEsmeraldo-like.gff

Table 4: Paths used to obtain the genome sequence files (FASTA format) and annotation files (GFF format) for the CL Brener genome.

A.2 Structure-Based Identification of ncRNA Candidates in Rfam

The RNA families database (RFAM, v.11.0) is a collection of multiple sequence alignments, consensus secondary structures and covariance models for over 2,000 RNA families [Burge *et al*., 2012], containing structure and sequence relationships. The families deposited in RFAM are from three main classes: ncRNAs, cis-regulatory elements and self-splicing RNAs. As it is to expect, all ncRNA families tipically exhibit a stronger structural conservation compared to a weaker sequence conservation and this is often observed among RFAM families [http://rfam.xfam.org]. Therefore, this database was assembled mainly as the result of a large-scale assessment of known RNAs using the INFERNAL algorithm (which stands for Inference of RNA alignments [Nawrocki & Eddy, 2013]). In brief, INFERNAL generates profiles from a multiple alignment of structurally annotated RNAs with a position-specific scoring system (for base substitutions, insertions and deletions) that takes into account not only the primary but also the secondary structure of the molecules by assessing the base position itself and its flanking elements. These structure-based profiles are in fact covariance models (CM), a specific type of stochastic context-free grammar (SCFG) [Lari & Young, 1991], which can be interpreted as an analogous of the sequence-based hidden Markov models used for protein analysis with the added complexity of the structural features that are taken into account. By investigating both sequence and structure, INFERNAL is in general more accurate and able to identify more distant homologues then other methods that rely solely on sequence comparisons [Nawrocki & Eddy, 2013].

I have used INFERNAL more recent version (version 1.1) instead of the classical previous version (version 1.0) with additional parameters. The main advantage of INFERNAL v.1.1 is the implementation of a new filter pipeline and constrains to the alignment algorithm, allowing the program to perform 100 times faster than the previous version and with virtually no loss of accuracy [Nawrocki & Eddy, 2013]. The CMCALIBRATE is a program from the INFERNAL package that estimates statistical parameters which are necessary for reporting e-values and stores these within the CM file itself. The e-value is calculated by assessing the score given to an alignment and estimating the probability that scores equal or greater than what would be observed merely by chance given the size of the database being queried. Therefore, each database should be calibrated prior to INFERNAL searches to account for such variation. The CM file used as input for INFERNAL was retrieved from the RFAM database and according to its documentation, it has been already calibrated with the CMCALIBRATE tool.

I have therefore installed the newer version of INFERNAL and decided to use the --rfam pre-determined set of thresholds, which is described by the INFERNAL manual as the most strict. By applying --rfam, INFERNAL will set all filter thresholds to the values used by a complex database like RFAM. The INFERNAL manual (http://selab.janelia.org/software/infernal/Userguide.pdf) also states that in future, when RFAM itself upgrades from the use of INFERNAL 1.1 instead of the 1.0 version, this option should be activated.

Command Line 1: INFERNAL run to identify RFAM ncRNAs with matches in the CL Brener complete genome.

```
cmsearch --rfam --cpu 8 --tblout Rfam_Genome.out Rfam.cm Genome.fasta
```

After running INFERNAL as described above I treated the output table to generate metrics about the results and further compose a FASTA format file containing the sequences of ncRNA candidates identified by the program. The output table of INFERNAL informs for each hit the chromosome, start and end position of the match, in addition to a description of the RNA class and family identifier from RFAM. These fields (and these fields only) were used to generate a summary file for each *T. cruzi* sub-genome (CLBe, CLBne and CLBmod). The summary file was then used by the in-house PERL-written script RetSeq.pl to retrieve the DNA sequence in each sub-genome which represents the ncRNA gene candidate, thus generating a FASTA file. The table output from INFERNAL was also used for analyses regarding the observed distributions of e-values, length of ncRNA candidates, classes of identified ncRNAs, distribution of ncRNA candidates across the three sub-genomes, for example.

The FASTA files were then transformed to FASTQ files (using the fasta2fastq.pl script) to serve as input for BOWTIE*2* [Langmead & Salzberg, 2012] genome mapping experiment which was performed with a very high stringency in order to (theoretically) only recover perfect matches. To achieve this, I have set BOWTIE*2* to the very sensitive end-to-end alignment mode and applied a very high penalty for mismatches.

Command Line 2: Converting FASTA files to FASTQ format.

```
fasta2fastq.pl input.fasta >> output.fastq
```

Command Line 3: BOWTIE2 genome mapping of previously identified ncRNA candidates.

```
bowtie2 --end-to-end --very-sensitive --mp 10,6
```

The generated BAM format files were converted to human-readable SAM format and sorted by read identifier using programs from the SAMTOOLS package (version 1.0) [Li *et al*., 2009] prior to submission to the htseq-count algorithm of HTSEQ (version 0.6.1) [Anders *et al*., 2014] which scans the set of mapped reads and attributes each to a genome feature. The selected feature for this work were the annotated genes contained in the combined *T. cruzi* CL Brener genome annotation file.

Command Line 4: SAMTOOLS format conversion and sorting of BAM files.

```
samtools sort -n -T temp -O sam -o SubGenome.sam SubGenome.bam
```

Command Line 5: HTSEQ annotation retrieval of identified ncRNA candidates.

```
htseq-count -o SubGenome_HTSeq.sam -i ID -r name -s no -t gene -a 0
SubGenome.sam SubGenome.gff >> HTSeq_SubGenome.log
```

This command line specifies the order in which reads were sorted (-r per read name), that the data in non-stranded (-s no), that the feature selected for read annotation is *genes* (-t gene) and that the GFF attribute for feature identification is the gene identifier (-i ID). In order to annotate mapped read pairs against the entire set of CL Brener genes, I have combined all three independent GFF files for the sub-genomes in one single file (All.gff) which was used by HTSEQ.

A.3 Sequence-Based Identification of ncRNA Candidates

While INFERNAL was used to assess structural characteristics of RFAM ncRNA molecules and search for CL Brener genomic sequences which potentially generate structurally analogous RNAs, BLAST searches were used to scan the genome and search for sequences with high similarity (likely homology) to known ncRNAs. Blastn was the program used to perform similarity searches in different databases, namely: (i) the NONCODE database of ncRNAs (which excludes tRNA and rRNA) [Liu *et al*., 2005], (ii) the miRBase database of microRNAs [Griffiths-Jones *et al*., 2006], (iii) the set of annotated *T. brucei* snoRNAs published by Liang *et al*. [Liang *et al*., 2005] and (iv) the set of trypanosomatid ncRNAs of various classes available at the TriTrypdb website (herein trypRNA). Unless otherwise stated, the e-value cutoff applied after homology searches was 0.01 and no additional BLAST filter was used. Outputs were generated in the table format (-m 8). The generic command line used while performing these BLAST searches is presented below.

Command Line 6: BLAST searches in ncRNA database

```
formatdb -p F -o T -i database.fasta

blastall -p blastn -d database.fasta -e 0.01 -m 8 -i genome.fasta -o output.blast
```

Although each result was not likely to be random (since a stringent e-value cutoff was applied), it could reflect the presence of no more than a fragment of an annotated ncRNA in the CL Brener genome. To further account for such cases and apply a selective criteria of acceptance to each BLAST hit, I have developed a PERL script to first calculate the length of individual entries in the mentioned datasets, then calculate its coverage, given the length of its alignment to the genome. An additional cutoff for sequence identity between the annotated ncRNA (subject) and the genomic region (query) was also applied using the same script and data from the BLAST output. For each of the databases a different combination of coverage and identity cutoffs was used to filter BLAST results. These are described in the respective results sections.

A.4 Scanning the Genome for Putative tRNAs

The search for tRNAs in the *T. cruzi* CL Brener complete genome was performed using the pre-calibrated TRNA.cm CM file provided within the source files of tRNAscan

[Schattner *et al*., 2005 based on the algorithm by Lowe & Eddy, 1997]. The tRNAscan algorithm has been successfully applied to over 800 entire genomes and resulted in the identification of thousands of tRNAs which are available in the Genomic tRNA Database [Chan & Lowe, 2009]. Although tRNAs have already been identified in the *T. cruzi* genome, I have decided to reassess these predictions using the refered database.

B. VALIDATING ncRNA CANDIDATES USING PUBLIC RNA-SEQ DATA

B.1 Public RNA-Seq Reads from the Sequence Read Archives

I have surveyed the SRA to search for previously deposited *T. cruzi* RNA-Seq data and found one major set of sequences from an experiment performed with parasite samples from the four different life-cycle stages. The experiment was part of a study from Sabatini and collaborators [Ekanayake *et al*., 2011] who were interested in the impact of depletion of the glucosylated thymine DNA base (base J), which was reported to have an impact in the unique transcriptional dynamics of the *T. cruzi* genome. A comprehensive review on the biosynthesis, genomic localization and possible functions of base J was published by Borst and Sabatini in 2008 [Borst & Sabatini, 2008]. The sequences deposited in SRA are from wild type parasites and parasites depleted from the enzymes that synthesize the modified base, all from the Y strain. I have retrieved the RNA-Seq results from all four life-cycle stages of the wild type parasite (SRX38883, SRX38885, SRX38887 and SRX38889 single-end unstranded libraries generated with Genome Analyzer II) and analyzed the set of sequences to identify possible ncRNAs that are expressed and are likely to have a role in the biology of *T. cruzi*.

B.2 Quality Assessment of RNA-Seq Reads

Downloaded FASTQ files were submitted to the FASTQC (version 0.11.2) program (http://www.bioinformatics.babraham.ac.uk/projects/fastqc), which performs quality control checks on deep sequencing data, generating useful reports from which one can identify problematic data, artefacts and biases that may be present in the sequenced material and have originated from either the sequencing process itself and/or the biological material extraction and preparation steps. FASTQC comprises a series of analysis modules, each capable of investigating a different feature from the set of sequences in the FASTQ file given as input. Next there is a brief description of each module (see the Program Descriptions Supplement for details):

i. *Basic Statistics* - this module presents a summary of simple statistics generated for the input dataset;

ii. *Per Base Sequence Quality* - this module presents an overview of the Phred quality range for all bases at each position within the sequence;

iii. *Per Sequence Quality Scores* - this module presents a graph of mean Phred quality values and standard deviation per number of sequences;

iv. *Per Base Sequence Content* - this module presents the content of each of the four bases per position in all reads in a single graph.

v. *Per Sequence GC Content* - this module presents a graph of GC content per number of reads plotted against a predicted normal distribution which represents the theoretically expected (ideal) distribution of GC content among sequences;

vi. *Per Base N Content* - this module presents the content of undefined bases (N) across all sequences;

vii. *Sequence Length Distribution* - this module presents the length of all sequences in the FastQ input file;

viii. *Duplicate Sequences* - in this module the first 200,000 sequences are truncated at the 50th nucleotide and analyzed regarding the presence of repetitive sequences in the entire set of sequences;

ix. *Overrepresented Sequences* - in this module any of the first 200,000 sequences that correspond to more than 0.1% of the total are reported as a warning;

x. *K-mer Content* - this module presents the number of each possible combination of 7 nucleotides (7-mer) measured at each position in the set of reads and assessed regarding the statistical significance of the observed representation;

xi. *Adapter Content* - this module is essentially identical to the previous, apart from the fact that the k-mers in this module are preset sequences that represent common adapters used by sequencing platforms;

xii. *Per Tile Sequence Quality* - this module, which is exclusive for inputs generated from sequencing on Illumina platforms where the information about the position (tile) of each sequenced read in the flowcell is retained, presents a heatmap where colours represent deviation of the sequence quality from the average, which is set as blue.

B.3 Read Counts Against the Set of Candidate ncRNAs

The alignment-free SAILFISH [Patro *et al*., 2013] method was used to calculate the expression counts of each ncRNA candidate based on the individual expression of k-mers derived from each ncRNA sequence. SAILFISH performs fast quantification of gene expression in two independent steps. First, it generates a database of k-mers based on a reference universe (in this case the set of all ncRNA candidates) and further assesses the representation of each k-mer in the RNA-Seq dataset of interest (in this case each of the four life-cycle stages) to estimate the expression levels of each ncRNA. I have used SAILFISH with a k-mer size of 20 to estimate the expression of the ncRNA candidates in *T. cruzi* RNA-Seq data from the SRA database.

SAILFISH is agnostic to any method of analysis and reports different outputs and when using differential expression tools (e.g. EDGER [Robinson *et al*., 2010], DESEQ [Anders & Huber, 2010] or similar R-implemented methods) for further analyses, it is advisable to consider the *estimatedReads* column of SAILFISH output, since these are raw counts of sequencing reads. Other counts reported by SAILFISH are normalized and could lead to meaningless results when used as input to such downstream tools, unless this normalization is taken into account. It is worth noting that, as SAILFISH counts k-mers instead of the actual reads, results are only *estimated* read counts.

Command Line 7: SAILFISH counts for expressed ncRNAs

```
sailfish index -t Sequences/Sequences_1595 -k 20 -o Sailfish/Index
sailfish quant -i INDEX -r READS -o OUT -a -l "T=SE:S=U"
```

IV. RESULTS

A. IDENTIFYING NCRNA CANDIDATES IN THE COMPLETE GENOME OF T. CRUZI *CL BRENER CLONE*

A.1 Problems with the Current Available CL Brener Genome

Since the genome content of the parasite is extremely repetitive, with nearly half of its DNA being composed of high-to-moderate repetitive sequences [Castro *et al*., 1981], its genome sequencing, assembly and annotation is not a trivial task. The currently available genome for the CL Brener clone is faulty in different ways and this is mainly due to the parasite's nature, which differs from most of the other eukaryotes in several aspects. I thus begin this session with a brief technical note to describe some of the issues with the genome which will help the understand of problems faced during the analysis reported here. First, I stress that CL Brener is a hybrid strain composed by two different sub-genomes (Esmeraldo-like, CLBe and Non-Esmeraldo, CLBne). Unfortunately, during the genome assembly process it was not possible to assign all contigs to the corresponding sub-genome [El Sayed *et al*., 2005]. As a result, the genome is available as three main subsets: the CLBe sub-genome, the CLBne sub-genome and a more fragmented subset of unassigned and mainly unassembled contigs (CLBmod). During the development of the present work, a decision had to be made on whether to treat the three sub-genomes as individual units or to join them in one combined genome. I decided to adopt the latter alternative, mainly to minimize both the number of runs for each program. To combine all three sub-genomes, CLBe, CLBne and CLBmod, the obtained FASTA files were merged by simply adding all three files together, one after the other in one final resulting file (e.g. using the *cat* command on a Unix terminal).

While investigating the organization of the genome annotation file for the *T. cruzi* CL Brener clone, I have noticed that gene annotation is not complete, in the sense that UTRs are not annotated as part of the genes. In summary, only the coding sequence (from the initial ATG codon to the most likely - or sometimes the known - stop codon) is annotated for each gene and therefore there is no distinction in the GFF annotation file between the elements named as *gene*, *coding sequence* (CDS), *mRNA*, and *exon* (given that *T. cruzi* is known to be an intronless organism). This should be kept in mind when sequences are annotated e.g. using HTSEQ, since elements identified as *genes* are in fact only the coding sequence of that gene and there are no clear definitions of regulatory elements.

A.2 An INFERNAL Search Through the RFAM Database

The chimeric complete genome of the *T. cruzi* CL Brener clone (obtained as previously described) was scanned with INFERNAL for significant matches to annotated RNAs from the RFAM database. The chosen version of RFAM (v.11.0) contains 383,004 sequences from 2,208 RNA families. By default, INFERNAL applies a very permissive e-value cutoff of 10, which roughly indicates that one would observe 10 hits in a random database of the same size. Nevertheless, once the results were available, only those matches which INFERNAL reports to satisfy the inclusion threshold were selected for further analyses. Inclusion thresholds differ from reporting thresholds and, by default, inclusion thresholds usually require an e-value of 0.01 or less (while reporting thresholds are set to 10). I have also delimited a subset of the valid matches using an e-value cutoff of 10^{-06}, which is very stringent. The table below (Table 5) depicts the number of RFAM matches retrieved when each of the thresholds is applied to the result of the INFERNAL.

RFAM hits	**Esmeraldo**	**Non-Esmeraldo**	**Not Assigned**	**Total**
e-value < 10.0	3475	3471	6605	13551
e-value < 0.01	275	294	1804	2373
e-value < 10^{-06}	157	161	771	1089

Table 5: Number of *T. cruzi* ncRNA candidates with similarity to RFAM entries retrieved by INFERNAL after different e-value cutoffs. Results are given for each sub-genome individually and for the entire genome.

I decided to work with the matches that were retrieved after the second filter (e-value < 0.01), since these are the ones considered as valid by INFERNAL itself and the third and more stringent filter (e-value < 10^{-06}) removed more than half the ncRNA candidates identified from the second filter, totalling 2,373 RNAs. The e-value distribution of these hits present a mean of 4.4^{-04} and a median of 7.8^{-05}, with standard deviation of 0.001. This is mainly (and negatively) influenced by the mode e-value of 8^{-05}, relative to matches from the tmRNA family, over-represented in the CLBmod sub-genome as will be further described (Table 6).

Metric	**CLBe**	**CLBne**	**CLBmod**	**Total**
e-value mean	1.2e-03	1.4e-03	1.6e-04	4.4e-04
e-value median	1.2e-09	1.1e-09	7.8e-05	7.8e-05
e-value SD	2.4e-03	2.7e-03	8.1e-04	1.5e-03
Score mean	54.8	52.0	52.7	52.9
Score median	86.6	74.4	22.0	21.9
Score SD	38.9	31.7	65.5	59.7
Number of hits	275	294	1804	2373

Table 6: The e-value and score statistics of INFERNAL searches in the *T. cruzi* genome against the RFAM database. Results are presented per sub-genome and for the entire genome.

Regarding the functional distribution of RFAM-matched ncRNA candidates, when considering all three sub-genomes, these 2,373 sequences are spread over a total of 10 RFAM families were found to be represented in the CL Brener complete genome, namely: 1002 tmRNAs, 628 *T. brucei* snoRNAs, 201 rRNAs, 143 RNase P, 142 tRNAs, 110 snoRNAs, 77 putative miRNAs, 11 U spliceosomal RNAs (uRNA), 10 snRNAs and 65 sequences from other families (including hammerhead ribozymes, scaRNAs, Phe Leaders and others). Although all these candidates were derived from INFERNAL with an e-value lower than 0.01 and INFERNAL has considered all these hits as significant, in some cases the alignment between the RFAM RNA and the genome does not cover the entire length of the annotated RNA. In fact, a preliminary assessment of the data has shown that for part of the hits this coverage can be very low. Considering this as an important feature I decided to further investigate the prevalence of such cases. First, I calculated the length of each RFAM entry and further the minimum, maximum, mean and median length of each RNA family. Next, I retrieved the length of each of the 2,373 matches in the CL Brener complete genome and compared with the range of length (from minimum to maximum) observed in the corresponding family. Considering only the CL Brener ncRNA candidates that had length within the range observed for the correspondent RFAM family there are 1,228 (~52%) matches. Interestingly, more than 80% of the matches with insufficient coverage were classified as tmRNAs, a class where 99% of the predicted ncRNAs have a shorter length than the shortest annotated RNA in RFAM. Additionally, this family presents a median e-value of 8.0^{-05}, while the same metric for all other families together is 1.4^{-14}. Taking these results together, I concluded that these may actually represent false positives.

I have then decided to further analyze only the candidates with length within the expected range. From the 1,228 ncRNA candidates, 242 were identified in the CLBe sub-genome, 241 in the CLBne and 745 in the CLBmod. Among the identified ncRNA candidates in the CLBe and CLBne sub-genomes, the majority is composed of *T. brucei*-like snoRNA candidates (72 and 81 sequences, respectively), RNase P candidates (66 and 52 sequences) and tRNA candidates (61 and 60 sequences). Additionally, there are miRNA candidates (24 and 28 sequences), rRNA candidates (5 and 7 sequences), uRNA candidates (5 sequences each), and other (9 and 8 sequences, including Phe leader RNAs and SRP RNAs). To assess the quality of this approach, I then decided to retrieve all previously annotated ncRNA candidates and check if the proposed annotation attributed by INFERNAL was correct.

A.3 Investigating Rfam ncRNA Candidates that were Already Annotated

BOWTIE2 was used to perform a stringent genome mapping using the ncRNA candidates FASTA files previously retrieved from INFERNAL. As a result, the chosen strategy for genome mapping reported (as requested) a single genome alignment to each of the sequences. Using HTSEQ union mode, I have retrieved 136 annotated ncRNA candidates from the CLBe sub-genome (Table 7).

HTSeq Mode	**Gene feature**	**CLBe**	**CLBne**	**CLBmod**	**Total**
	Aligned	242	241	745	1228
Union	Annotated	136 (56.2%)	135 (56.0%)	430 (57.7%)	701
	Non annotated	99 (40.9%)	101 (41.9%)	312 (41.9%)	512
	Ambiguous	7 (02.9%)	5 (02.1%)	3 (00.4%)	15
Intersect strict	Annotated	113 (46.7%)	98 (40.7%)	109 (14.6%)	320
	Non annotated	129 (53.3%)	143 (59.3%)	636 (85.4%)	908
	Ambiguous	0 (00.0%)	0 (00.0%)	0 (00.0%)	0

Table 7: Sequences mapped to the reference genome and further annotated according to the GFF file. The table presents the number of sequences successfully mapped to the reference genome using BOWTIE2 and subsequent annotation status of the corresponding genome region according to HTSEQ analyses using either the union mode or the intersect-strict mode.

For 121 sequences the RFAM analogue identified with INFERNAL corroborated the previous genomic annotation. Another 15 ncRNA candidates retrieved with INFERNAL are annotated as either hypothetical proteins (4 snoRNAs, 5 miRNAs and 3 RNaseP had hits with 8 different hypothetical proteins) or specifically annotated as a methyltransferase (hit with RNaseP) and a pseudogene for dynein heavy chain (hit with a human long-non coding RNA). From the CLBne sub-genome I have retrieved 135 annotated ncRNA candidates. For 126 sequences the INFERNAL match corroborated the previous genomic annotation. Another 9 ncRNA candidates retrieved with INFERNAL are annotated as either hypothetical proteins (2 snoRNAs, 2 miRNAs and 3 RNaseP had hits with 7 hypothetical proteins) or specifically annotated as an endopeptidase (hit with RNaseP) and a helicase (hit with a miRNA). Finally, from the CLBmod sub-genome, all annotated ncRNAs were correctly predicted (Table 8).

ncRNA Class	**CLBe**	**CLBne**	**CLBmod**	**Total**
T. brucei like snoRNAs	52	56	269	377
5S rRNAs	5	6	135	146
tRNAs	59	58	5	122
Other rRNAs	0	1	18	19
U spliceosomal RNAs	5	5	3	13
Incorrectly predicted ncRNAs	15	9	0	24

Table 8: Distribution of the annotated sequences along different ncRNA classes. The distribution is given based on HTSeq results (union mode) and presented for each sub-genome individually as well as for the entire genome.

In total, only 24 from the 701 previously annotated ncRNA candidates seem to be false positives, representing only 3% of the total and indicating a very high accuracy of the prediction method. Therefore, this *in silico* protocol seems to be a valid strategy for the identification of novel ncRNAs in the *T. cruzi* CL Brener complete genome. Despite this high accuracy, I retrieved near 40% of all the previously annotated ncRNAs using

INFERNAL. In a further section I will present the number of known ncRNAs in trypanosomatids and below (Table 9) I compare the number of each annotated class as found in the TriTrypDB website and the portion retrieved with INFERNAL. The reason for both the apparently high accuracy and the low retrieval is most likely the stringency I have applied for filtering the results and the evolutionary distance between *T. cruzi* and most of the species that comprise the set of RNAs in RFAM.

	snoRNA	**tRNA**	**rRNA**	**snRNA**	**Total**
INFERNAL retrieved	377	122	164	13	676
Previously annotated	1442	115	219	19	1795

Table 9: The number of ncRNA candidates retrieved by INFERNAL compared to the number of ncRNAs previously annotated in the parasite.

A.4 Investigating Rfam ncRNA Candidates that were not Annotated

When considering all 512 candidate ncRNAs identified by INFERNAL and subsequently characterized as non-annotated by HTSEQ, less than 50 RFAM families are present, representing mainly 9 RNA functional classes (Table 10). The majority of these ncRNA candidates are from the *T. brucei* snoRNA, RNase P or miRNA classes, the last two represent important new findings not previously reported in trypanosomatids.

RNA Family	**CLBe**	**CLBne**	**CLBmod**	**Total**
T. brucei snoRNAs	9	18	239	266
RNase P	61	48	25	134
miRNAs	19	25	1	45
tRNAs	2	2	15	19
tmRNAs	0	0	13	13
Hammerhead	0	0	8	8
Phe leader RNAs	2	4	0	6
rRNA	0	0	5	5
SRP RNAs	2	1	0	3
Other	4	3	6	13

Table 10: The number of ncRNA candidates retrieved by INFERNAL and with no previous annotation. Only the top functional classes are presented and results are given per sub-genome and for the entire genome.

A.5 Searching for NONCODE ncRNAs in the *T. cruzi* CL Brener Genome

The NONCODE database (v.4) currently contains 366,977 ncRNA sequences from 10 different model organisms (from yeast to human): 148,172 human sequences, 74,963 mouse sequences, 60,253 rat sequences, 27,126 cow sequences, 24,008 *Caenorhabditis elegans* sequences, 12,836 zebrafish sequences, 12,122 chicken sequences, 3,853 *Arabdopsis thaliana* sequences, 3,193 fruit fly sequences, and 451 yeast sequences. From the total number of entries, over 200,000 represent long ncRNAs, thus enriching the analyses on this class of ncRNAs.

While assembling a comprehensive dataset from the species-specific FASTA files, I noticed that the subset of rat sequences is biased towards the extreme lengths, with a minimum length of 1 and a maximum length of nearly 1.5 million bases. To illustrate the discrepancies, while this dataset has a mean length of approximately 18,000, the human subset has a mean length of only near 650. Excluding rats, the length of deposited ncRNAs (which exclude tRNAs and rRNAs) ranges approximately from 10 to 100,000, with a mean value of 770 and a median of 450, which are much more plausible values. I therefore decided to exclude the subset of rat ncRNAs from the total dataset, being left with 306,724 sequences.

I have performed BLAST searches in the publicly available NONCODE data (excluding rat sequences) against the CL Brener genome using an e-value cutoff of 0.01. As a result, I have retrieved nearly 12,000 matches, which were further filtered by a cutoff of at least 80% coverage and 80% identity between query and subject, using an in-house PERL-written script. After filtering, there were nearly 500 hits representing less than 200 unique ncRNA candidates in the *T. cruzi* genome.

Interestingly, the candidates were not evenly distributed along the genome, but rather they were concentrated in 6 chromosomes of each the CLBe and the CLBne sub-genomes (chromosomes 17, 22, 41, 37, 8 and 39 in order of abundance), one extra chromosome of the CLBne sub-genome (9) and 21 contigs of the CLBmod sub-genome. As this physical concentration of candidates in specific chromosomes could indicate a functional relationship between candidates, I have listed the set of candidates in each chromosome and noticed that only chromosome 41 presented a high copy number of two specific ncRNAs, thus rejecting this hypothesis. Interestingly on chromosome 41 there are 39 ncRNA candidates which are multiple copies of only 4 NONCODE entries. Regarding species, over two thirds of the candidates identified in this step show sequence similarity to *Caenorhabditis elegans, Arabdopsis thaliana* and *Mus musculus* ncRNAs (five other species are also present), thus indicating a high level of evolutionary conservation for these RNAs. In fact, the most abundant elements in this subset of candidates are mammal RNAs transcribed as antisense to the ubiquitin gene (data not shown) and at this stage it is not possible to decide if these are analogous ncRNAs in *T. cruzi* or if these were identified only based on the high sequence identity of ubiquitin itself as a conserved protein. The second most abundant set of elements are from the 28S rRNA family and thus also highly evolutionarily conserved.

A.6 Searching for miRBase ncRNAs in the *T. cruzi* CL Brener Genome

I have performed a BLAST search in the CL Brener complete genome to find matches with the MIRBASE database of miRNAs. The MIRBASE database used (v.21) was composed of 28,645 hairpin sequences, from 223 species (the three most represented being mammals) and over 14,000 subclasses of miRNAs, ranging in size from 39 to 2,354 nucleotides (mean of 103 and median of 91). I have applied e-value restrictions to match the cutoff used in INFERNAL searches (0.01) and retrieved 3,484 hits, 99% (3,461) of these in the CLBmod sub-genome. I further filtered these results to contain only matches with at least 30% coverage of the hairpin sequence by the *T. cruzi* genomic region and at least 80% nucleotide identity and only 125 unique matches met this criteria. Interestingly all but one of the 125 predicted miRNAs are located in the CLBmod sub-genome, spread along 40 different CLBmod contigs. Although functional annotation is not yet available, it is worth mentioning that mir-7360 seems to be most abundant miRNA in the parasite genome.

A.7 Searching for Previously Identified *T. brucei* snoRNAs in the *T. cruzi* CL Brener Genome

In 2005, Liang and collaborators have used SNOSCAN (based on the algorithm of Lowe & Eddy, 1999) to identify snoRNA candidates in the genome of *T. brucei*. The study has revealed 91 snoRNA candidates from both the C/D (57) and H/ACA (34) classes, ranging from 56 to 125 nucleotides in length (mean of 82 and median of 78). I have used BLAST to scan this data and search for possible *T. cruzi* homologues. As a result, using an e-value threshold of 0.01, I have recovered 1,203 matches which were further filtered regarding their percentage of identity to the *T. cruzi* genome sequence and the percentage of the snoRNA which was covered by the alignment. When recovering matches with at least 80% of snoRNA sequence coverage and 80% sequence identity, I have found over 50 high-quality hits with 5 out of the 91 entries from the Liang *et al*. database. From these, there were 27 hits for the same C/D box snoRNA in different genomic positions and 25 hits with 4 H/ACA box snoRNAs.

A.8 Searching for Previously Identified Trypanosomatid ncRNAs in the *T. cruzi* CL Brener Genome

The TriTrypDB is an integrated genomic and functional genomic database for pathogens of the Trypanosomatidae family. From this database I have downloaded the subset of all annotated ncRNAs, composed of 7,060 sequences from 20 different organisms of three genera: *Trypanosoma* (4,086 sequences from 6 species), *Leishmania* (2,670 sequences from 6 species) and *Crithidia* (304 sequences from only 1 sequence). The table below (Table 11) shows a summary of the annotated trypncRNAs, which include snoRNAs, tRNAs, rRNAs, spliced leader RNAs, snRNAs, SRP RNAs and *other*. Among all represented organisms, the species contributing the higher number of sequences is the *T. cruzi* CL Brener strain itself (with 1,795 sequences), in which most of the annotated sequences are snoRNAs.

	snoRNA	tRNA	Unspecified	rRNA	SL RNA	snRNA	SRP RNA	Other	Total
All species	3400	1388	1385	706	91	66	11	13	7060
CL Brener	1442	115	-	219	-	19	-	-	1795

Table 11: Previously annotated ncRNAs in trypanosomatid genomes. Only 6 main classes were previously identified in these organisms and 4 in the *T. cruzi* CL Brener strain.

After eliminating the already annotated ncRNAs from *T. cruzi* CL Brener, 5,265 RNAs remained in the database that was then used for BLAST searches. I have used an e-value cutoff of 0.01, which returned over 125,000 hits, mainly due to very short hits and redundant reporting of the same ncRNA being similar to many different genomic regions. When a 80% coverage and 80% identity filter was applied, only a little over 20,000 hits remained, representing approximately 1,000 unique ncRNAs in the *T. cruzi* CL Brener genome. Among the 1,036 candidates, 765 are located in the CLBmod sub-genome and the remaining 271 are nearly evenly distributed between the CLBe (in 17 chromosomes) and CLBe (18 chromosomes) sub-genomes.

A.9 Identifying tRNAs in the *T. cruzi* CL Brener Genome

Using INFERNAL and the TRNA.cm CM file provided by tRNAscan I have identified 124 putative tRNA genes in the complete genome of CL Brener. From the sub-genomes, CLBe presented 55 candidates, CLBne 56 and the CLBmod 13. Considering these 124 predictions, all but one of the 115 annotated tRNAs were predicted by this approach. The only one missing is from the CLBmod sub-genome, contig number 7392. The median and the mean CM scores and e-values from these predictions are of 64.5, 63.6, 1.5^{-10} and 1.5^{-09} respectively, which can be considered a very good result. One additional tRNA was predicted at CLBne chromosome 37, with a high CM score of 68.5 and a low e-value of 1.4^{-13}. I have also observed 9 new tRNA candidates in the CLBmod sub-genome, with no previous annotation and these should be further investigated. Interestingly, tRNAs range in size from 61 to 90 nucleotides with a mean of 74 and a median of 72 nucleotides, which is in agreement with their expected size of 70-80 nucleotides.

A.10 Summary of ncRNA Candidates Identified in the *T. cruzi* CL Brener Genome

A total of 2,673 unique ncRNA candidates were identified based on known database entries (Table 12). Since there is a degree of redundancy among these databases (e.g. both miRBase and the NONCODE databases contain miRNAs), I first excluded multiple ncRNA candidates which were mapped to the exact same genomic position with equivalent annotation, being left with 2,357 ncRNA candidates. In addition to the candidates mapped to the exact same genomic position, there were superposed entries which could be reduced to only one comprehensive entry.

Search tool	Based on	Database	Hits before filter	Hits after filter	Unique hits
Infernal	Structure	Rfam	2373	1228	1228
Infernal	Structure	tRNAScan	124	124	124
BLAST	Sequence	miRBase	3484	125	56
BLAST	Sequence	NONCODE*	11833	494	177
BLAST	Sequence	TrypRNAs	125743	20418	1036
BLAST	Sequence	*T. brucei* snoRNAs	1203	52	52
Total			144761	22441	2673
* excluding rat ncRNAs					

Table 12: The number of ncRNA candidates retrieved by each search tool in different databases. Numbers are presented before and after filtering. The number of non-redundant (unique) candidates is also shown.

I have then designed and applied an algorithm implemented in a simple in-house built PERL script to identify and exclude candidates which are fully contained within another. As a result, 1,777 comprehensive candidates were kept and further manually curated to eliminate reminiscent redundant candidates. The 1,595 further remaining candidates are spread throughout the sub-genomes (Table 13) being 253 from CLBe, 259 from CLBne and 1,083 from CLBmod. Interestingly, among these ncRNA candidates, the most represented classes are *T. brucei*-like snoRNAs, RNase P, SL RNAs, among others.

ncRNA class	**CLBe**	**CLBne**	**CLBmod**	**Total**
T. brucei-like snoRNAs	72	81	527	680
rRNAs	7	10	181	198
RNase P	65	52	23	140
tRNAs	56	58	18	132
T. brucei-like SL RNAs	0	0	111	111
miRNAs	21	25	47	93
NONCODE transcripts	15	13	34	62
C. fasciculata-like snoRNAs	0	2	35	37
Miscellaneous RNA	4	5	19	28
T. vivax-like TcY486 snoRNA	3	2	22	27
L. mexicana-like LmxM.00 ncRNA	0	0	22	22
T. congolensis-like 067 ncRNA	0	0	17	17
U spliceosomal RNAs (U2, U3, U6)	5	5	3	13
tmRNAs	0	0	13	13
Hammerhead	0	0	8	8
Phe leader RNA	2	4	0	6
Other snRNAs and snoRNAs	2	1	3	6
SRP	1	1	0	2

Table 13: The total number of curated and non-redundant ncRNA candidates. Numbers are presented per sub-genome and for the entire genome.

B. SPECIAL FEATURES OF THE NCRNA MOLECULES IN THE GENOME OF T. CRUZI *CL BRENER CLONE*

B.1 Description of Some ncRNA Classes Identified in the *T. Cruzi* Cl Brener Genome

The most abundant ncRNA class identified in this study is the snoRNAs with similarity to *T. brucei*. We have observed nearly 700 of those RNAs, mostly in the unassembled CLBmod genome. The high abundance of this class may reflect the previous annotation of such ncRNAs in the *T. brucei* genome, thus facilitating the identification of homologues in *T. cruzi*. This class of small RNAs is usually associated with modifications to process rRNA molecules their a mature state. The snoRNAs guide protein complexes that can perform either methylation (C/D box snoRNAs) or pseudouridylation (H/ACA box snoRNAs) of rRNAs. In the dataset there are only 78 C/D box-like snoRNAs and thus over 600 H/ACA box snoRNAs.

As presented in the table above, from all databases used in this study I was able to identify 132 tRNA candidates (Table 13). The only annotated tRNA not identified in the tRNAscan database was retrieved from the set of previously annotated trypanosomatid RNAs, thus successfully completing the retrieval of all annotated tRNAs. Additionally, I have annotated one missing tRNA on CLBne chromosome 37, with matching location to its analogue on the CLBe sub-genome. Finally, 16 new tRNAs were reported, being 13 from the CLBmod sub-genome, one from the chromosome 33 of the CLBne sub-genome and other 2 from the same chromosome of the CLBe sub-genome. Interestingly, most of these putative new tRNAs are selenocysteine-carrying (tRNA-Sec). Regarding strandness, given a chromosome or contig all tRNAs are in the same strand and regarding genomic location tRNAs are spread across 15 of the 41 chromosomes of each sub-genome. Some chromosomes (notoriously chromosomes 22 and 34 of each sub-genome) seem to be more enriched in tRNAs and these are usually located near each other (although no statistical test was applied to verify the significance of this result). A total of 93 miRNAs were identified, as presented above (Table 13). These range in size from 18 to 189 nucleotides, with a mean of 81, a median of 84 and a mode of 22, which is a typical size for processed miRNAs.

Regarding the 62 ncRNA candidates identified with similarity to NONCODE entries, there are 10 different types of ncRNAs. As mentioned, the two candidates with highest number of copies in the genome (30 and 21 copies of NONMMUT020592 and

NONMMUT074488, respectively) are both co-localized with ubiquitin genes and yet to be confirmed as ncRNAs. Based on BLAST searches I have identified 5 other NONCODE hits as putative rRNAs, comprising 8 *T. cruzi* ncRNA candidates. The last 3 NONCODE hits could not be assigned to any specific ncRNA class through BLAST searches.

While investigating the genome of *T. cruzi* CL Brener I have found 119 lncRNAs, being 24 annotated as rRNAs. The remaining 95 candidates are spread throughout all sub-genomes, being 22 from the CLBe, 20 from the CLBne and 59 from the CLBmod. Interestingly, 30 snoTBR17-like, 30 NONMMUT020592-like and 19 NONMMUT074488-like ncRNA candidates comprise the great majority of these.

Other less abundant classes include tmRNAs, hammerheads, Phe leader RNAs and SRPs. The tmRNAs are bacterial ncRNAs with both tRNA and mRNA functions that exist as part of a complex which is capable of solving stalled protein translation by recycling the ribosome and adding a proteolysis-inducing mark to the protein and mediating the degradation of the faulty mRNA in a process named trans-translation [reviewed by Keiler, 2008]. The hammerhead is the smallest catalytic RNA motif, comprising one of the best characterized ribozymes, found in most life forms that cleaves RNA molecules in a site-specific manner at phosphodiester bonds [Uhlenbeck, 1987; Symons 1992]. Phe leader RNAs are structurally conserved RNAs that contain short ORFs rich in codons for phenylalanine. In bacteria, these peptide leaders control expression of the downstream gene according to the levels of phenylalanine analogously to the tryptophan attenuation mechanism [Bertrand *et al*., 1976]. When phenylalanine is abundant, the translation of the short ORF generates a structural element that blocks downstream gene expression. On the other hand, when phenylalanine levels are low, ribosome stalling leads to an alternative structure formation enabling downstream gene expression. Last, SRPs are ncRNAs that bind to signal peptides emerging from ribosomes and target the associated protein to the endoplasmic reticulum [Nagai *et al*., 2003].

B.2 Genomic Localization of the ncRNA Candidates

I then decided to investigate the genomic localization of all 1,595 candidates in the annotated CL Brener genome. A PERL script was written to categorize ncRNA candidates in four main classes: (i) between two genes, (ii) at the ends of chromosomes/contigs, (iii) not accompanied by any annotated gene in this contig and (iv) related to a previously annotated gene. The first category contains 376 ncRNA candidates predicted to be located between two annotated genes in a CLBe or CLBne chromosome or CLBmod contig. The second category is composed by 82 ncRNA candidates which are located at the end of a chromosome or contig, before the first or after the last annotated gene. The third category comprises 247 ncRNA candidates predicted to be located in CLBmod contigs in which no other gene is currently annotated. The last category comprises 882 ncRNAs predicted to be located in a genomic position with at least partial positional overlap to previously annotated genes. This class was further subdivided in several subclasses: (A) ncRNA candidates which exactly or nearly exactly (≤ 5 nucleotides) co-localize with an annotated gene (500 occurrences), (B) ncRNAs which exactly or nearly exactly (≤ 20 nucleotides) co-localize with the first or last gene in a CLBmod contig (171 occurrences), (C) ncRNAs which are entirely contained within an annotated gene (150 occurrences), (D) ncRNAs which partially co-localize but are not fully contained in annotated genes (54 occurrences) and (E) ncRNAs with partial overlap to two annotated genes (7 occurrences). In addition, 8 ncRNA candidates were classified as a mixture of different categories and to avoid more complex analyses were excluded from further results. The figure below (Figure 5) is a graphic representation of each category described above and also a summary of number of ncRNA candidates in each category.

B.3 Assessing ncRNA Candidates Co-Localized with Previously Annotated Genes

I have assessed all 500 predicted ncRNAs with exact positional coincidence to previously annotated genes. As result, I observed that virtually all (but 7) of these match previous annotation, 336 are *T. brucei*-like snoRNAs (7 of which are currently annotated as hypothetical proteins), 114 are tRNAs, 22 are rRNAs, 17 are predicted as *T. congolensis*-like 067 ncRNAs (all annotated as snoRNA TB11C2C1) and 11 are uRNAs. Interestingly, from the 336 *T. brucei*-like snoRNAs, 45% are either TB11Cs4H3, TB6Cs1H1 or TB10Cs1H1, which are H/ACA-like snoRNAs predicted to guide the pseudouridylation of large subunit rRNA 3 at positions Ψ566, Ψ380 or Ψ659, respectively [Liang *et al*., 2005].

From the list of ncRNA candidates located within previously annotated genes (i.e. the ncRNA candidate is fully contained in a previously annotated gene), 68 are snoRNAs (6 annotated as hypothetical proteins), 30 are rRNAs, 18 are NONMMUT020592 (all annotated as polyubiquitin pseudogenes), 10 are TvY486-like ncRNAs, (all annotated as snoRNA TB11C2C2), 8 are miRNAs (6 coincide with annotated hypothetical proteins, 1 with a rRNA and 1 with a helicase), 8 are NONMMUT074488 (all annotated as polyubiquitin pseudogenes) and the remaining 8 are from different classes (7 annotated as rRNAs).

Predictions for all 171 ncRNAs candidates co-localized with either the first or last annotated gene of CLBmod contigs match previous annotations. From these 132 (nearly 80%) are rRNAs, 24 are *T. brucei*-like snoRNAs, 12 are TvY486-like snoRNAs and 3 are U2 RNAs. There are 54 ncRNA candidates with partial overlap to annotated genes and an overhang at either the 3' or 5' end. From these, 24 are *T. brucei*-like snoRNAs (1 annotated as a hypothetical protein), 17 are rRNAs, 8 are RNase P (6 annotated as hypothetical proteins), 2 are tRNAs, 2 are *T. congolensis*-like 067 ncRNAs (annotated as U6 uRNAs) and 1 is predicted as a tRNA and annotated as a mucin-associated protein. All 7 occurrences of predicted ncRNAs which span across two different annotated genes are located at chromosome 16 of either CLBe or CLBne genomes.

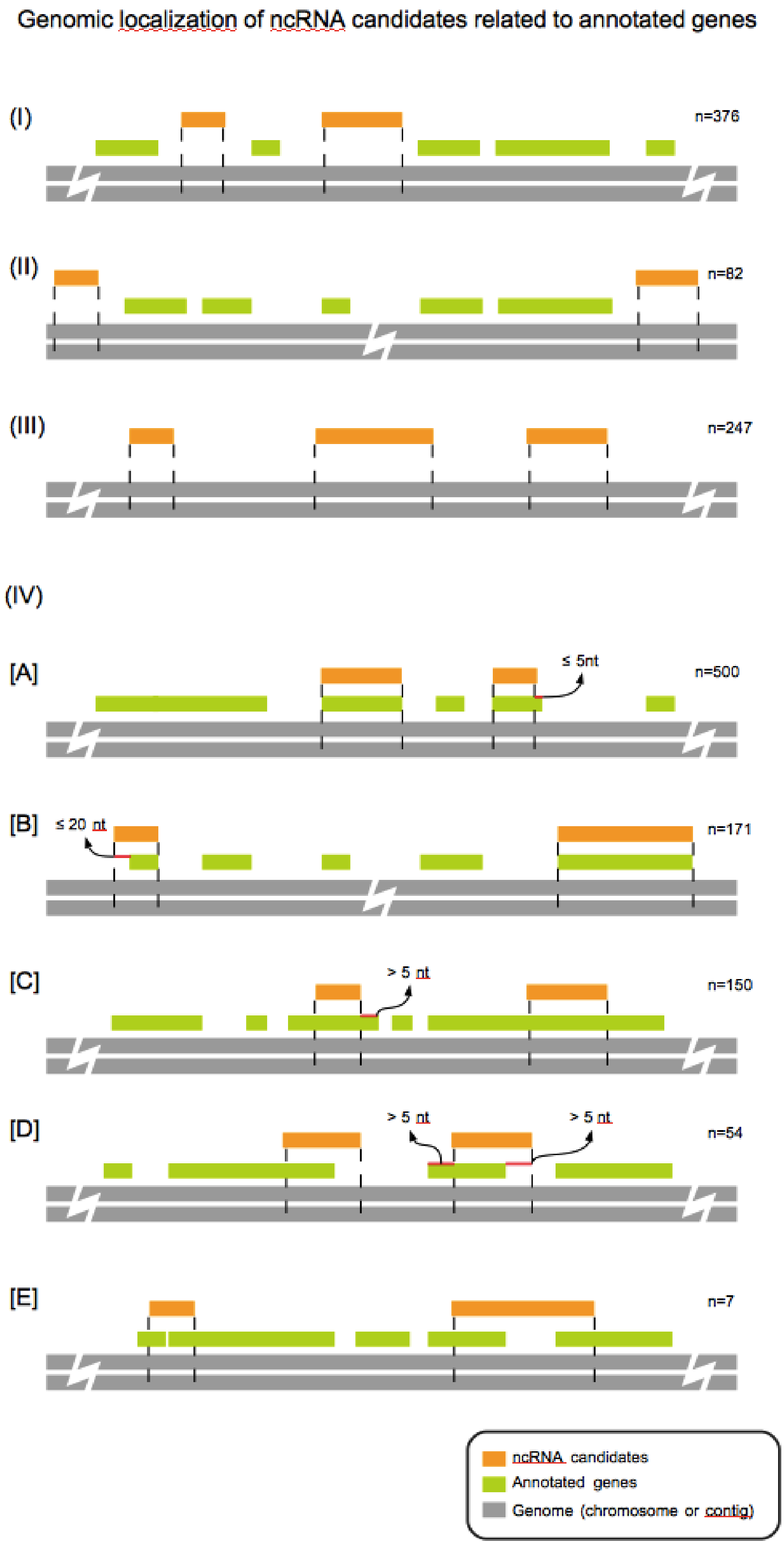


Figure 5: Classification of the 1,595 ncRNA candidates regarding their relative genomic localization to previously annotated genes in the *T. cruzi* CL Brener genome. Numbers indicate the amount of candidates in each class and the length cutoffs used in specific classes to include or exclude matches.

After assessing these results I noticed that in all these regions there is an annotated snoRNA partially superposed with a downstream annotated hypothetical protein, thus generating a double annotation for the ncRNA candidate. I have predicted 5 TvY486-like ncRNAs and 2 *C. fasciculata*-like snoRNA to be located in these regions of superposition.

In summary, I have identified over one third of the previously annotated snoRNAs (484 from ~1424), all tRNAs (115), nearly all rRNAs (over 200 from 219), and nearly all the uRNAs (15 from 19). On total, even with a very strict protocol, I have re-identified nearly half the previously annotated ncRNAs from the CL Brener genome (Table 14).

	snoRNA	**tRNA**	**rRNA**	**snRNA**	**Total**
CL Brener	1442	115	219	19	1795
Predicted	484 (33.6%)	115 (100%)	201 (91.8%)	15 (78.9%)	815 (45.4%)

Table 14: The number of previously annotated CL Brener ncRNA candidates (top) compared to the amount of ncRNA candidates retrieved in this work with annotation according to previous records (bottom).

B.4 Assessing New ncRNA Candidates

After confirming the correct prediction of nearly half the previously annotated ncRNA candidates, I have assessed the non-annotated ncRNA candidates. For the 376 ncRNA candidates located between two annotated genes, over 50% (195) are *T. brucei*-like snoRNAs, nearly 30% (109) are RNAse P and approximately 10% (39) are miRNAs. From the 247 non-annotated ncRNA candidates located in CLBmod contigs with no annotated genes, 111 are SL RNAs, 44 are miRNAs, 22 are RNAse P, 15 are tRNAs, 13 are tmRNAs, 13 are NONMMUT074488, 12 are NONMMUT020592 and 6 are from different classes. Last, from the 82 non-annotated ncRNA candidates located before the first or after the last annotated gene in CLBmod contigs, nearly 90% are snoRNAs, mainly TB11Cs4H1 (34) and TB11Cs4H2 (26).

Regarding less abundant ncRNA genes in all three location classes, there are 13 tmRNAs, 9 putatively new tRNAs, 8 tRNA-Sec, 8 hammerhead RNAs, 6 Phe leader RNAs and 2 SRP. Taken together, these results represent two different and complementary conclusions: first, the ncRNA candidates identified herein which are related to annotated genes are in general in agreement with previous annotation and second, the set of newly observed ncRNAs mainly comprises RNAse P (131), miRNAs (83) and SL RNAs (111).

B.5 Relative Localization of the ncRNA Candidates

Regarding the localization of ncRNA candidates relative to one another, I have designed a simple PERL script to identify candidates with less than 200 nucleotides length of separation. As a result, 157 candidates were shown to be partially superposed to their nearest neighbour and many of these are from the same class. These such cases could be further studied and putatively collapsed to an unique element in the event of a genome re-annotation. Interestingly, over 349 additional candidates are less than 200 nucleotides apart and 133 are less than 50 nucleotides apart from their nearest neighbours. Most notably, these ncRNA candidates include snoRNAs and tRNAs. In face of this tendency, I have performed an assessment of each of these classes independently. As a result, over half the tRNAs (59) and 156 of the snoRNAs are in tandem (< 200 nucleotides) with at least one other candidate of the same class. To illustrate this observation, on chromosome 16 of the CLBe sub-genome I have identified a cluster of 3 snoRNAs which are repeated 4 times in a short length of the genome (Figure 6).

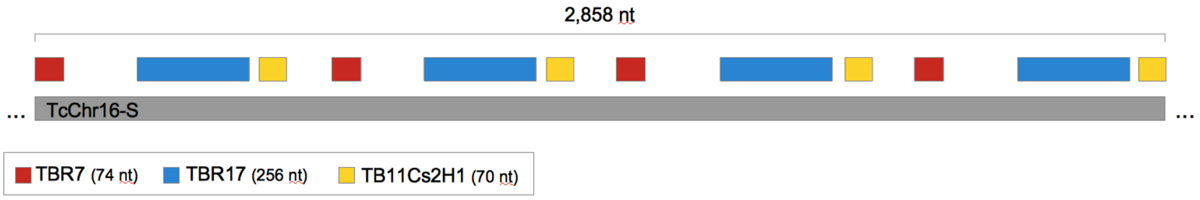


Figure 6: An example of snoRNA clusters in the CL Brener genome. This example depicts a cluster of three different snoRNAs repeated four times in a space of under 3,000 nucleotides of the CLBe chromosome 16.

C. VALIDATING ncRNA CANDIDATES USING RNA-SEQ DATA

C.1 Retrieval of Previously Generated RNA-Seq Data

When searching for entries related to the term “Trypanosoma cruzi” in the SRA database, I have found 36 matches, being 10 of those related to RNA data of different types. The remaining 26 entries found are related to whole genome sequencing projects. From the 10 entries for RNA sequences, eight are from the previously mentioned work from Sabatini and colleagues regarding the role of base J. Within this project there are four sets of reads from wild type parasites (each of a different life-cycle stage) and four sets from parasites depleted from base J synthesizing enzymes. I have retrieved the first four sets of reads, yielding a total of approximately 850 million bases distributed in more than 22 million 38-nucleotides reads from the four life-cycle stages of the wild type parasite altogether. Reads are distributed along four life-cycle stages as follows: ~3 million for amastigote parasites, ~4 million reads for epimastigote parasites, ~6 million reads for trypomastigote parasites and ~9 million reads for metacyclic parasites.

C.2 Quality Assessment

FASTQC was used to ensure the high quality of reads and to guide the decision of wether or not to trim the sequences or discard possible adaptor-containing reads. As an overall result, sequence quality was high (considering the restricted 38 nucleotides length of reads, and the technology used to generate the sequences) and no significant contamination seemed to be present. The per-base quality value observed for reads of all sets was always of 26 or higher, enough to place this data within the highest quality region. The per-sequence quality scores peak were at 37 for all sets, representing a good quality. Based on this results I have therefore chosen not to trim the already short reads and keep the full dataset as downloaded from the SRA, without manual Bioinformatics intervention.

C.3 Counting RNA-Seq Reads that Map to ncRNA Candidates

The alignment-free SAILFISH method was used to calculate the expression counts of each ncRNA candidate based on the individual expression of the k-mers derived from each ncRNA sequence. I have used k-mers of 20 nucleotides (or 20-mers), as the shortest sequence identified in this work is only 18 nucleotides. The generated index was composed of 44,732 20-mers originated from the 1,595 sequences. The k-mers originated from reads of all four life-cycle stages were independently mapped to this index sequences and results are summarized in the table below (Table 15).

Stage	Number of 20-mers generated for the RNA-Seq samples	Percentage of 20-mers mapped to the ncRNA database	Anchored reads	Uniquely anchored reads
Ama	49.7×10^6	1.90%	350	171
Epi	73.9×10^6	1.20%	421	173
Meta	172.5×10^6	0.47%	402	177
Trypo	121.3×10^6	0.95%	432	184

Table 15: Statistics of the results produced by SAILFISH for counts of reads mapped to the ncRNA dataset.

I have found over 300 expressed ncRNA candidates in the available RNA-Seq data. It is worth noting that I have not applied any expression cutoff and therefore any level of expression was considered as a positive count. The table below (Table 16) therefore presents counts of how many candidates from each class were observed to be expressed, regardless of expression levels. The last column however contains the number of ncRNAs in each class which are differentially expressed in different life-cycle stages, according to SAILFISH results on the RNA-Seq data. SAILFISH counts were analyzed using EDGER with a minimum of 5 reads per million as cutoff for gene expression and a false discovery rate of 0.05. As a result, 10% of the expressed ncRNAs were considered as differentially expressed between the four life stages according to the described parameters.

ncRNA class	**Ama**	**Epi**	**Meta**	**Trypo**	**Total**	**Differentially Expressed**
RNase P	128	126	131	127	132	13
miRNA	34	31	37	38	39	3
rRNA	10	21	16	23	39	8
SL RNA	34	34	34	34	34	0
T. brucei-like snoRNA	13	6	15	11	17	0
snoRNAs	4	4	4	4	4	0
NONCODE	10	11	8	14	14	0
Miscellaneous RNAs	7	7	7	8	8	3
Phe leader	6	5	5	6	6	2
tRNA	0	3	2	0	5	0
U spliceosomal RNA	3	3	1	3	3	1
Total	249	251	260	268	301	30

Table 16: The number of transcripts expressed (at any level) in each category, per life-cycle stage. Expression was measured on the RNA-Seq data from Sabatini and collaborators.

V. DISCUSSION

In the course of this work I have scanned the genome of *T. cruzi* in search of ncRNAs that may play important roles in the parasite biology. We have selected both structure-based and sequence-based methods to identify candidates which were further curated and filtered. Mainly, we have applied a minimum identity cutoff, a minimum e-value cutoff and a minimum coverage cutoff. The first two filters were set in either INFERNAL (for structure-based searches) or BLAST (for sequence-based searches) and the latter was implemented as separate PERL-written scripts. These scripts were different for INFERNAL-generated and BLAST-generated results. Since the INFERNAL run did not provide specific hits but general RNA families, the PERL scripts first assessed the length of each database entry to generate metrics for each family and further considered only those matches within the expected range. BLAST on the other hand has allowed me to retrieve each alignment and thus calculate the query coverage and further apply a filter using PERL scripts. The coverage filter in combination with the e-value cutoff should improve the accuracy of these predictions by excluding false positives such as matches that are likely to be identified just by chance and also matches that are too short in comparison to the full length of the database entry.

Only 3% of the INFERNAL matches to the Rfam database with previous annotation do not match the genome annotation, indicating a very low false discovery rate for this method in this context. In addition, most of these putative false positives are annotated as hypothetical proteins (which make up to approximately 50% of the annotated proteins in the entire CL Brener genome) and therefore revision is needed to properly annotate such entries. This high accuracy is compensated by a low retrieval rate of previously annotated ncRNAs. Using INFERNAL and further filters as described, only around 40% of all the previously annotated ncRNAs were retrieved. Nevertheless, this was an interesting survey which allowed me to identify a large amount of *T. brucei*-like snoRNAs, RNase P ncRNAs and miRNAs, among other candidates. INFERNAL was also used to characterize tRNAs using the database contained within tRNAscan. To expand the catalogue of ncRNAs in *T. cruzi*, I have employed sequence-based searches using BLAST to scan for candidates in different databases, namely: NONCODE, MIRBASE, snoRNAs from Liang *et al.*, 2005 and the TriTrypDB website. As the described databases can be redundant, I have designed scripts to reduce multiple predictions of the same ncRNA candidate at the same genomic position to one unique prediction. As a result both the structure-based and the sequence-based genome scans, we were able to identify 1,595 ncRNAs. As mentioned, these ncRNAs candidates are distributed along all three sub-genomes, being 253 from CLBe, 259 from CLBne and 1,083 from CLBmod. The candidates from both the CLBe and the CLBne sub-genomes are scattered across 36 of the 41 chromosomes each. Interestingly, most of the candidates that are present in the CLBe sub-genome have a counterpart in CLBne and vice-versa. From the 1,595 ncRNAs, over 700 represent new ncRNAs, with the most abundant classes being snoRNAs, SL RNAs and miRNAs, respectively.

The study of small RNAs in trypanosomatids has been concentrated in the class of snoRNAs and this is probably the best known class of ncRNAs in these organisms. In 2005, Liang and collaborators [Liang *et al*., 2005] have published a very detailed survey on snoRNAs in *T. brucei* and showed that such genes are organized in clusters and repeated several times in the chromosome. Accordingly, we have found *T. cruzi* snoRNAs to be disposed in tandem and usually a small cluster is repeated in the chromosome. It is worth noting that a high percentage of the predicted snoRNAs are located in short CLBmod contigs and thus it is not possible to assess the repetitive nature of such candidates until the genome is fully assembled. In the assembled portion of the CL Brener genome the snoRNA clusters seem to be syntenic between the CLBe and the CLBne sub-genomes and the distances between two consecutive elements in a cluster tend to be highly conserved, even though their exact genomic location may be shifted when comparing two sub-genomes.

Regarding function, as presented before snoRNAs are involved in the mechanism of rRNA modification by guiding enzymes that perform RNA methylation or pseudourylation. In the trypanosomatid *C. fasciculata*, a surprising number of nearly 100 methylation sites were identified in rRNAs [Gray, 1979]. These modifications are known to increase ribosomal stability, a feature that in plants is believed to be important through drastic temperature changes, during which the translational machinery must remain active [Liang *et al*., 2005]. The authors suggest that trypanosomes would also benefit from ribosomal stabilization when migrating from the insect to the mammalian host and thus experiencing a temperature rise of ~10°C. In a different perspective, *T. cruzi* is exposed to various different types of stresses during its life-cycle, one of the most severe probably resulting from the inhospitable environment in the insect vector intestine. During its cycle, *T. cruzi* frequently experiences changes in pH and osmolarity and considerable oxidative stress [Kollien & Schaub, 2000]. Interestingly, epimastigote forms of the parasite have been shown to be highly resistant to ionizing radiation [as shown by Takeda *et al*., 1986 and addressed by Regis-da-Silva *et al*., 2006]. Regarding its natural resistance to oxidative stresses, it has been speculated by our group if this could be the underlying cause of the parasite resistance to ionizing radiation [e.g. Vieira *et al*., 2014]. In this sense, in addition to the authors hypothesis, I also suggest that the stabilization of ribosomes may play an important role in the resistance of *T. cruzi* to oxidative stress and ionizing radiation. Additionally, snoRNAs have been also shown to exert their function on tRNAs, snRNAs and other small RNAs [reviewed by Bachellerie *et al*., 2002]. Interestingly, the identification of snoRNAs acting upon tRNAs in Archaea has revealed an ancient origin for such molecules. A specific H/ACA-like molecule is responsible for directing pseudouridylation on the SL RNA of trypanosomatids, although its function in the biogenesis of this molecule is not clear [reviewed in Liang *et al*., 2005]. Interestingly, in this survey I have identified the non-annotated snoRNA responsible for SL-RNA pseudourydilation (SL-associated RNA or SLA1) in the CL Brener genome. Characterization was first achieved by comparison with the cluster in which SLA1 should be located, according to the map of snoRNAs presented by Liang *et al.* in 2005 and the description made by Dunbar *et al*. in 2000, placing the snoRNAs TBR7 and TBR17 in the same cluster. Notably, this is the cluster I have chosen to exemplify the tandem repeats in Figure 16 and it also contains TBR7 and TBR17. I have then performed BLAST searches which confirmed the prediction of the SLA1 gene according to previous reports [Sturm *et al.*, 2003].

RNase P ribonucleases are protein-RNA complexes essential for RNA polymerase III transcription, apparently by both exerting a direct role in transcription and by processing tRNAs and other small RNAs in coordination with transcription [Reiner *et al*., 2006]. The RNA polymerase III of *T. brucei* is also well characterized and has been shown to transcribe tRNAs, 5S RNAs and U-rich snRNAs [Palenchar & Bellofatto, 2006]. In addition, the *T. brucei* Rnase P enzyme is a nuclear and mitochondrial complex and its molecular constitution differs from other eukaryotes by lacking the RNA molecule [Taschner *et al*., 2012]. Putative homologues to this type of RNA-independent RNase P (named proteinaceous RNase P, PRORP) have been found in all available trypanosomatid genomes, including *T. cruzi* [Taschner *et al*., 2012]. However, previous Bioinformatic searches have failed to identify any RNase P RNA candidate in trypanosomatid genomes [Rosenblad *et al*., 2006]. This can therefore be the first time an RNA-based RNase P is reported in trypanosomatids and this finding can subside further reassessments of the enzyme evolutionary pathway. RNA-based RNase P complexes seem to be more ancient and widespread and it is believed these have evolved to an RNA-independent form in more complex organisms. In any case, a more detailed survey searching for possible RNA-based RNase P protein homologues in the parasite genome and assessing their expression would be necessary to conclusively state that this enzyme is present in *T. cruzi*. In this sense, it would be first necessary to find a close species with an RNA-based Rnase P and further scan the genome of *T. cruzi* to search for a possible homologue. Further, it would be interesting to assess the tridimensional structure of the protein complex and its interaction with the RNA molecule. Additionally, since only a putative homologue of RNA polymerase III was reported in this parasite and no expression data is available at this stage, even if proven to exist, the function of the RNA-based RNase P complex in this parasite would need to be further defined. Lastly, SL RNAs were detailed discussed in the previous chapter and in brief these ncRNAs are involved in the processing of *T. cruzi* polycistronic transcripts. In summary, all three of the most abundant ncRNA classes identified in the CL Brener genome are related to post-transcriptional processing of other RNAs, both protein-coding or small ncRNAs (mainly tRNAs and rRNAs).

The small RNA content on the *T. cruzi* transcriptome has been previously studied, with the conclusion that these were processed RNAs mainly derived from tRNAs, rRNAs, snoRNAs and snRNAs in a non-random pattern and no homology to known miRNAs was found for the remaining transcripts [Franzén *et al*., 2011b]. The tRNA-derived small RNAs have been first described by Garcia-Silva *et al*. as actively produced RNAs which are recruited to cytoplasmic granules, especially in stress situations [Garcia-Silva *et al*., 2010]. The reporting of miRNA-like sequences in the parasite genome is surprising, since *T. cruzi* lacks the classical miRNA-processing machinery [DaRocha *et al.*, 2004; El sayed *et al*., 2005]. Regarding the findings presented here I have two different hypothesis for the processing of precursor RNAs in mature forms of miRNAs in the parasite. The first possibility is that *T. cruzi* would employ a non-canonical RNAi machinery, which is supported by the finding of an Aragonaute-like protein, a previously identified protein with a Piwi domain but no PAZ domain [Silva *et al*., 2010]. The second possibility is that *T. cruzi* would export precursor miRNA molecules to be processed by the host RNAi machinery and this would represent a very interesting observation if proved to be plausible. Coincidently, Mourier and collaborators when working on the discovery of structured RNAs in *Plasmodium falciparum* have raised the same hypotheses [Mourier *et al*., 2008].

Interestingly, among the less abundant ncRNA classes within the newly identified candidates, there are tmRNAs and Phe leader RNAs which are supposed to only exist in the prokaryotic kingdom. Both of these classes, if proven to be functional in an eukaryote as *T. cruzi* would have an important effect in the understanding of the protist evolution. In this sense, adapted mechanisms would have to exist in the parasitic cell in order to accommodate such molecules which are adapted to the comparatively simple bacterial gene expression system. Additionally, there were new tRNAs, especially tRNA-Sec, hammerheads RNA components and SRPs, all of these with known function in eukaryotes and involved in basic biological processes, mainly regarding gene expression. In summary, the majority of the ncRNAs identified here play direct or indirect roles in the mechanisms of transcription and translation, indicating a broad involvement of these RNAs in the complex and unusual gene expression machinery of *T. cruzi*. This observation is parallel to what has been reported in prokaryotic organisms, where ncRNAs act mainly on the regulation of gene expression [reviewed by Sorek and Cossart, 2010]. Nevertheless, the involvement of ncRNAs with gene expression has been reported in other organisms as well [e.g. Eddy *et al*., 2001; Mattick & Makunin, 2006;].

There are 119 lncRNAs in the 1,595 ncRNA candidates dataset, the majority of which composed by 30 snoTBR17-like ncRNAs, 30 NONMMUT020592-like and 19 NONMMUT074488-like ncRNAs. A sequence search in RNACentral has revealed that the latter two are mouse lncRNAs homologous to human lncRNAs which are transcribed in the antisense of ubiquitin genes. This observation likely explains why all of these ncRNAs are located in regions previously annotated as containing polyubiquitin protein-coding genes in the genome of *T. cruzi*. Further investigation is necessary but these may be false positives of ncRNA candidates which are only retrieved by BLAST given their sequence similarity to murine ubiquitin genes. On the other hand, snoTBR17 are small nucleolar RNAs of usually ~120 nucleotides that seem to present longer members in *T. cruzi*. Taken together these results point to a general absence of long non-coding RNAs in the genome of T. cruzi, which is therefore mainly populated by small ncRNAs. In contrast, a recent survey for the genome-wide identification of lncRNAs in *Plasmodium falciparum* has reported 164 lncRNAs with high confidence, among which 69 were functionally annotated, pointing to a significative contribution of such RNAs to the parasite biology [Liao *et al*., 2014].

VI. CONCLUSIONS

Taken together the results presented in this chapter point to a *T. cruzi* CL Brener genome highly populated with ncRNAs. In fact, the number of ncRNA candidates retrieved in this survey alone (1,595) is nearly the same as the number of functionally annotated proteins (e.g. proteins annotated as something else than *hypothetical protein*) which is of 1,743. This would suggest an organism which is as ruled by the effector protein molecules as it is by the regulatory ncRNAs. There is no report in the literature on large-scale surveys of ncRNAs in trypanosomatids, except for specific classes. Therefore, this could be one of the most comprehensive compendium of such RNAs in a species of the genera *Trypanosoma*. To my understanding, the main contributions of the present work are the corroboration of nearly half the previously identified ncRNAs using different methodologies and/or filters and the identification of nearly 700 new ncRNA candidates. The general overview of ncRNAs in *T. cruzi* as presented here indicates these are mainly involved with gene expression regulation processes, a scenery that resembles what has been reported in prokaryotes. I now intend to further validate the expression of these candidates, mainly using next-generation sequencing technologies to generate high quality and deep RNA sequence data.

REFERENCES

Agabian N 1990. Trans splicing of nuclear pre-mRNAs. Cell 61:1157-1160.

Allen MA, Hillier LW, Waterston RH, Blumenthal T 2011. A global analysis of C. elegans trans-splicing. Genome Research 21:255-264.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ 1990. Basic local alignment search tool. Journal of Molecular Biology 215(3):403-410.

Anders S, Huber W 2010. Differential expression analysis for sequence count data. Genome Biology 11:R106.

Anders S, Pyl PT, Huber W 2014. HTSeq-A Python framework to work with high-throughput sequencing data. bioRxiv.

Aslett M, Aurrecoechea C, Berriman M, Brestelli J, Brunk BP, Carrington M, et al. 2010. TriTrypDB: a functional genomic resource for the Trypanosomatidae. Nucleic Acids Research 38:D457-462.

Bachellerie JP, Cavaillé J, Hüttenhofer A 2002. The expanding snoRNA world. Biochimie 84(8):775-790.

Bachvaroff TR, Place AR 2008. From stop to start: tandem gene arrangement, copy number and trans-splicing sites in the dinoflagellate Amphidinium carterae. PloS One 3:e2929.

Backofen R, Bernhart SH, Flamm C, Fried C, Fritzsch G, Hackermüller J, Hertel J, Hofacker IL, Missal K, Mosig A, Prohaska SJ, Rose D, Stadler PF, Tanzer A, Washietl S, Will S 2007. RNAs everywhere: genome- wide annotation of structured RNAs. Journal of Experimental Zoology Part B: Molecular and Developmental Evolution, 308(1), 1-25.

Basch PF 1981. Cultivation of Schistosoma mansoni in vitro. I. Establishment of cultures from cercariae and development until pairing. The Journal of Parasitology 67:179-185.

Beranova-Giorgianni S 2003. Proteome analysis by two-dimensional gel electrophoresis and mass spectrometry: strengths and limitations. Trends in Analytical Chemistry 22:9.

Berger S 2014. Infectious Diseases of Mexico. Global Infectious Disease and Epidemiology Network (GIDEON) Informatics ISBN:9781498801379.

Bern C & Montgomery SP 2009. An estimate of the burden of Chagas disease in the United States. Clinical Infectious Disease 49:e52-54.

Berriman M, Ghedin E, Hertz-Fowler C, Blandin G, Renauld H, Bartholomeu DC, et al. 2005. The genome of the African trypanosome Trypanosoma brucei. Science 309(5733):416-422.

Bertrand K, Squires C, Yanofsky C 1976. Transcription termination in vivo in the leader region of the tryptophan operon of Escherichia coli. Journal of Molecular Biology 103(2):319-337.

Blumenthal T, Gleason KS 2003. Caenorhabditis elegans operons: form and function. Nature Reviews - Genetics 4:112-120.

Boguski MS, Lowe TM, Tolstoshev CM 1993. dbEST--database for "expressed sequence tags". Nature Genetics 4:332-333.

Boothroyd JC, Cross GA 1982. Transcripts coding for variant surface glycoproteins of Trypanosoma brucei have a short, identical exon at their 5′ end. Gene 20:281-289.

Borst P, Sabatini R 2008. Base J: Discovery, Biosynthesis, and Possible Functions. Annu Review on Microbiology 62:235-51.

Brandão A, Urmenyi TP, Rondinelli E, Gonzalez A, de Miranda AB, Degrave W 1997. Identification of transcribed sequences (ESTs) in the Trypanosoma cruzi genome project. Memórias do Instituto Oswaldo Cruz 92(6):863-866.

Brehm K, Jensen K, Frosch M 2000. mRNA trans-splicing in the human parasitic cestode Echinococcus multilocularis. The Journal of Biological Chemistry 275:38311-38318.

Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP, Bateman A. 2012. Rfam 11.0: 10 years of RNA families. Nucleic Acids Research 41:D226-232.

Bustin SA, Benes V, Garson JA, Hellemans J, Huggett J, Kubista M, Mueller R, Nolan T, Pfaffl MW, Shipley GL, Vandesompele J, Wittwer CT 2009. The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. Clinical Chemistry 55:611-622.

Cass CL, Johnson JR, Califf LL, Xu T, Hernandez HJ, Stadecker MJ, Yates JR, 3rd, Williams DL 2007. Proteomic analysis of Schistosoma mansoni egg secretions. Molecular and Biochemical Parasitology 155:84-93.

Castro C, Craig SP, Castañeda M. Genome organization and ploidy number in Trypanosoma cruzi 1981. Molecular Biochemistry Parasitology 4(5-6):273-82.

Castro-Borges W, Simpson DM, Dowle A, Curwen RS, Thomas-Oates J, Beynon RJ, Wilson RA 2011. Abundance of tegument surface proteins in the human blood fluke Schistosoma mansoni determined by QconCAT proteomics. Journal of Proteomics 74:1519-1533.

Chagas C 1909. Nova tripanozomiaze humana: estudos sobre a morfolojia e o ciclo evolutivo do Schizotrypanum cruzi n. gen., n. sp., ajente etiolojico de nova entidade morbida do homem. Memórias do Instituto Oswaldo Cruz 1(2):159-218.

Chan PP, Lowe TM 2009. GtRNAdb: A database of transfer RNA genes detected in genomic sequence. Nucleic Acids Research 37:D93-D97.

Clark MB, Choudhary A, Smith MA, Taft RJ, Mattick JS 2013. The dark matter rises: the expanding world of regulatory RNAs. Essays in biochemistry 54(1):1-16.

Cory S, Spahr PF, Adams JM 1970. Untranslated nucleotide sequences in R17 bacteriophage RNA. In Cold Spring Harbor Symposia on Quantitative Biology 35:1-12.

Coura JR & Borges-Pereira J 2010. Chagas disease: 100 years after its discovery. A systemic review. Acta tropica 115(1):5-13.

Coura JR & Viñas PA 2010. Chagas disease :a new worldwide challenge. Nature 465(7301S):S6-7.

Crick F 1970. Central dogma of molecular biology. Nature 227(5258):561-563.

Crooks GE, Hon G, Chandonia JM, Brenner SE 2004. WebLogo: A Sequence Logo Generator. Genome Research 14:1188-1190.

Curwen RS, Ashton PD, Johnston DA, Wilson RA 2004. The Schistosoma mansoni soluble proteome: a comparison across four life-cycle stages. Molecular and Biochemical Parasitology 138:57-66.

Dan A, Pereira MH, Pesquero JL, Diotaiuti L, Beirão PS 1999. Action of the saliva of Triatoma infestans (Heteroptera: Reduviidae) on sodium channels. Journal of Medical Entomology 36(6):875-9.

DaRocha WD, Otsu K, Teixeira SM, Donelson JE 2004. Tests of cytoplasmic RNA interference (RNAi) and construction of a tetracycline-inducible T7 promoter system in Trypanosoma cruzi. Molecular and Biochemical Parasitology 133(2):175-86.

Davis RE 1996. Spliced leader RNA trans-splicing in metazoa. Parasitology Today 12:33-40.

Davis RE, Hardwick C, Tavernier P, Hodgson S, Singh H 1995. RNA trans-splicing in flatworms. Analysis of trans-spliced mRNAs and genes in the human parasite, Schistosoma mansoni. The Journal of Biological Chemistry 270:21813-21819.

De Souza W 2002. Basic Cell Biology of Trypanosoma cruzi. Current Pharmaceutical Design 8(4):269-285.

Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Research 22(9):1775-1789.

Dias JCP 2007. Southern Cone Initiative for the elimination of domestic populations of Triatoma infestans and the interruption of transfusion Chagas disease: historical aspects, present situation, and perspectives. Memórias do Instituto Oswaldo Cruz 102:11-18.

Dias JCP, Amato Neto V, Luna EJDA 2011. Alternative transmission mechanisms of Trypanosoma cruzi in Brazil and proposals for their prevention. Revista da Sociedade Brasileira de Medicina Tropical 44(3):375-379.

Dias JCP, Silveira AC, Schofield CJ 2002. The impact of Chagas disease control in Latin America: a review. Memórias do Instituto Oswaldo Cruz 97(5):603-612.

Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A 2012. Landscape of transcription in human cells. Nature 489:101-108

Dunbar DA, Chen AA, Wormsley S, Baserga SJ 2000. The genes for small nucleolar RNAs in Trypanosoma brucei are organized in clusters and are transcribed as a polycistronic RNA. Nucleic Acids Research 28(15):2855-2861.

Drummond SC, Pereira SRS, Silva LCDS, Antunes CMDF, Lambertucci JR 2010. Schistosomiasis control program in the state of Minas Gerais in Brazil. Memórias do Instituto Oswaldo Cruz 105(4):519-523.

Duszenko M, Ginger ML, Brennand A, Gualdrón-López M, Colombo MI, Coombs GH, Coppens I, Jayabalasingham B, Langsley G, deCastro SL, Menna-Barreto R, Mottram JC, Navarro M, Rigden DJ, Romano PS, Stoka V, Turk B, Michels PAM 2011. Autophagy in protists. Autophagy 7(2):127-158.

ENCODE Project Consortium 2004. The ENCODE (ENCyclopedia of DNA elements) project. Science 306(5696):636-640.

Eddy SR 2001. Non-coding RNA genes and the modern RNA world. Nature Reviews Genetics 2(12):919-929.

Ekanayake DK, Minning T, Weatherly B, Gunasekera K, Nilsson D, Tarleton R, Ochsenreiter T, Sabatini R 2011. Epigenetic Regulation of Transcription and Virulence in Trypanosoma cruzi by O-Linked Thymine Glucosylation of DNA. Molecular Cell Biology 31(8):1690-700.

El-Sayed NM, Myler PJ, Bartholomeu DC, Nilsson D, Aggarwal G, Tran AN, et al. 2005. The genome sequence of Trypanosoma cruzi, etiologic agent of Chagas disease. Science 309(5733):409-415.

Ewing B, Hillier L, Wendl MC, Green P 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research 8:175-185.

Faulkner GJ, Brown MK, Taylor DF, Steptoe AL, Wani S, Bethel G, Robertson AJ, Perkins AC, Bruce SJ, Lee CC, Ranade SS, Peckham HE, Manning JM, McKernan KJ, Grimmond SM 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nature Methods. 5:613-619.

Fenwick A, Webster JP 2006. Schistosomiasis: challenges for control, treatment and drug resistance. Current Opinion in Infectious Diseases 19:577-582.

Ferreira HO 1976. Ensaio terapêutico-clínico com o benzonidazol na doença de Chagas. Revista do Instituto de Medicina Tropical de São Paulo 18(5):357-364.

Franco GR, Adams MD, Soares MB, Simpson AJ, Venter JC, Pena SD 1995. Identification of new Schistosoma mansoni genes by the EST strategy using a directional cDNA library. Gene 152:141-147.

Franzén O, Ochaya S, Sherwood E, Lewis MD, Llewellyn MS, Miles MA, Andersson B 2011a. Shotgun sequencing analysis of Trypanosoma cruzi I Sylvio X10/1 and comparison with T. cruzi VI CL Brener. PLoS Neglected Tropical Diseases 5(3):e984.

Franzén O, Arner E, Ferella M, Nilsson D, Respuela P, Carnici P, Hayashizaki Y, Aslund L, Andersson B, Daub CO 2011b. The Short Non-Coding Transcriptome of the Protozoan Parasite Trypanosoma cruzi. PLoS Neglected Tropical Diseases 5(8): e1283.

Garcia-Silva MR, Frugier M, Tosar JP, Correa-Dominguez A, Ronalte-Alves L, Parodi-Talice A, Rovira C, Robello C, Goldenberg S, Cayota A 2010. A population of tRNA-derived small RNAs is actively produced in Trypanosoma cruzi and recruited to specific cytoplasmic granules. Molecular Biochemical Parasitology 171: 64-73.

Gilbert W 1986. Origin of life: The RNA world. Nature 319(6055).

Gray MW 1979. The ribosomal RNA of the trypanosomatid protozoan Chritidia fasciculata: Physical characteristics and methylated sequences. Canadian Journal of Biochemistry 57:914-926.

Gray DJ, McManus DP, Li Y, Williams GM, Bergquist R, Ross AG 2010. Schistosomiasis elimination: lessons from the past guide the future. The Lancet Infectious Diseases 10:733-736.

Greif G, Ponce de Leon M, Lamolle M, Rodriguez M, Piñeyro D, Tavares-Marques LM, Reyna-Bello A, Robello C, Alvarez-Valin F 2013. Transcriptome analysis of the bloodstream stage from the parasite Trypanosoma vivax. BMC Genomics. 14:149.

Griffiths-Jones S, Grocock RJ, Van Dongen S, Bateman A, Enright AJ 2006. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Research 34:D140-144.

Guimaraes RJPS, Freitas CC, Dutra LV, Oliveira G, Carvalho OS 2013. Multiple regression for the schistosomiasis positivity index estimates in the Minas Gerais state - Brazil at small communities and cities level. Parasitic Diseases - Schistosomiasis, Chapter 1 ISBN:9789535109426 InTech.

Gunzl A, Bruderer T, Laufer G, Schimanski B, Tu LC, Chung HM, Lee PT, Lee MG 2003. RNA polymerase I transcribes procyclin genes and variant surface glycoprotein gene expression sites in Trypanosoma brucei. Eukaryotic cell 2:542-551.

Hastings KE 2005. SL trans-splicing: easy come or easy go? Trends in Genetics 21:240-247.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Metha JP, Collins FS, Manolio TA 2009. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proceedings of the National Academy of Sciences of the United States of America 106(23):9362-9367.

Hofacker IL 2003. Vienna RNA secondary structure server. Nucleic Acids Research 31(13):3429-3431.

Huang X, Madan A 1999. CAP3 A DNA sequence assembly program. Genome Research 9:868-877.

Huang XY, Hirsh D 1989. A second trans-spliced RNA leader sequence in the nematode Caenorhabditis elegans. Proceedings of the National Academy of Sciences of the United States of America 86(22):8640-8644.

Ivens AC, Peacock CS, Worthey EA, Murphy L, Aggarwal G, Berriman M, et al. 2005. The genome of the kinetoplastid parasite, Leishmania major. Science 309(5733):436-442.

Kalinna BH, Brindley PJ 2007. Manipulating the manipulators: advances in parasitic helminth transgenesis and RNAi. Trends in Parasitology 23:197-204.

Kapranov P, Cheng J, Dike S, Nix DA, Duttagupta R, Willingham AT, et al. 2007. RNA maps reveal new RNA classes and a possible function for pervasive transcription. Science 316(5830):1484-1488.

Keiler KC 2008. Biology of trans-translation. Annual Reviews in Microbiology 62:133-51.

Keeling PJ, Burger G, Durnford DG, Lang BF, Lee RW, Pearlman RE, Roger AJ, Gray MW 2005. The tree of eukaryotes. Trends in Ecology and Evolution 20(12):670-676.

Kent WJ 2002. BLAT-The BLAST-Like Alignment Tool. Genome Research 12:656-664.

Knudsen GM, Medzihradszky KF, Lim KC, Hansell E, McKerrow JH 2005. Proteomic analysis of Schistosoma mansoni cercarial secretions. Molecular and Cellular Proteomics 4:1862-1875.

Kollien AH, Schaub GA 2000. Parasitology Today. The development of Trypanosoma cruzi in triatominae. 16(9):381-7.

Kowalczyk MS, Higgs DR, Gingeras TR 2012. Molecular biology: RNA discrimination. Nature 482(7385):310-311.

Krause M, Hirsh D 1987. A trans-spliced leader sequence on actin mRNA in C. elegans. Cell 49:753-761.

Langmead B, Salzberg S 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357-359.

Lari K, Young SJ 1991. Applications of stochastic context-free grammars using the inside-outside algorithm. Computer Speech & Language 5(3):237-257.

Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23(21):2947-2948.

Lasda EL, Blumenthal T 2011. Trans-splicing. Wiley Interdisciplinary Reviews: RNA 2(3):417-434

Lee MG, Van der Ploeg LH 1997. Transcription of protein-coding genes in trypanosomes by RNA polymerase I. Annual Review of Microbiology 51:463-489.

Leinonen R, Sugawara H, Shumway M 2011. The sequence read archive. Nucleic Acids Research 39:D19-21.

Lewis MD, Llewellyn MS, Gaunt MW, Yeo M, Carrasco HJ, Miles MA 2009. Flow cytometric analysis and microsatellite genotyping reveal extensive DNA content variation in Trypanosoma cruzi populations and expose contrasts between natural and experimental hybrids. International Journal for Parasitology 39(12):1305-1317.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25(16):2078-2079.

Liang X, Haritan A, Uliel S, Michaeli S 2003. Trans and cis Splicing in Trypanosomatids: Mechanism, Factors, and Regulation. Eukaryotic Cell 2(5):830-840.

Liang XH, Uliel S, Hury A, Barth S, Doniger T, Unger T, Unger R, Michaeli S 2005. 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 11(5):619-645.

Liao Q, Shen J, Liu J, Sun X, Zhao G, Chang Y, Xu L, Li X, Zhao Y, Zheng H, Zhao Y, Wu Z 2014. Genome-wide identification and functional annotation of Plasmodium falciparum long noncoding RNAs from RNA-seq data. Parasitology Research 113(4):1269-1281.

Lidie KB, van Dolah FM 2007. Spliced leader RNA-mediated trans-splicing in a dinoflagellate, Karenia brevis. The Journal of Eukaryotic Microbiology 54:427-435.

Liu C, Bai B, Skogerbø G, Cai L, Deng W, Zhang Y, Bu D, Zhao Y, Chen R 2005. NONCODE: an integrated knowledge database of non-coding RNAs. Nucleic Acids Research 33:D112-D115.

Livak KJ, Schmittgen TD 2001. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. Methods 25:402-408.

Logan-Klumpler FJ, De Silva N, Boehme U, Rogers MB, Velarde G, McQuillan JA, Carver T, Aslett M, Olsen C, Subramanian S, Phan I, Farris C, Mitra S, Ramasamy G, Wang H, Tivey A, Jackson A, Houston

R, Parkhill J, Holden M, Harb OS, Brunk BP, Myler PJ, Roos D, Carrington M, Smith DF, Hertz-Fowler C, Berriman M 2012. GeneDB--an annotation database for pathogens. Nucleic Acids Research 40:D98-108.

Lorenz R, Bernhart SH, Siederdissen CH, Tafer H, Flamm C, Stadler PF, Hofacker IL 2011. ViennaRNA Package 2.0. Algorithms for Molecular Biology 6:26.

Lowe TM, Eddy SE 1999. A computational screen for methylation guide snoRNAs in yeast. Science 283:1168-71.

Lowe TM, Eddy SR 1997. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. Nucleic Acids Research 25:955-964.

Macedo AM, Machado CR, Oliveira RP, Pena SDJ 2004. Trypanosoma cruzi: genetic structure of populations and relevance of genetic variability to the pathogenesis of chagas disease. Memórias do Instituto Oswaldo Cruz 99(1):1-12.

Main J, Gunasekera K, Roditi I 2012. RNA-Seq Analysis of the Transcriptome of Trypanosoma brucei. RNA Metabolism in Trypanosomes. Nucleic Acids and Molecular Biology. 2012;28:237-265

Mann VH, Morales ME, Kines KJ, Brindley PJ 2008. Transgenesis of schistosomes: approaches employing mobile genetic elements. Parasitology 135:141-153.

Martins-Melo FR, Alencar CH, Ramos Jr AN, Heukelbach J 2012. Epidemiology of mortality related to Chagas' disease in Brazil, 1999-2007. PLoS Neglected Tropical Diseases 6(2):e1508.

Martins-Melo FR, Ramos Jr AN, Alencar CH, Heukelbach J 2014. Prevalence of Chagas disease in Brazil: A systematic review and meta-analysis. Acta tropica 130:167-174.

Massad E 2008. The elimination of Chagas' disease from Brazil. Epidemiology & Infection 136(9):1153-1164.

Mathieson W, Wilson RA 2010. A comparative proteomic study of the undeveloped and developed Schistosoma mansoni egg and its contents: the miracidium, hatch fluid and secretions. International Journal for Parasitology 40:617-628.

Matsumoto J, Dewar K, Wasserscheid J, Wiley GB, Macmil SL, Roe BA, Zeller RW, Satou Y, Hastings KE 2010. High-throughput sequence analysis of Ciona intestinalis SL trans-spliced mRNAs: alternative expression modes and gene function correlates. Genome Research 20:636-645.

Mattick JS, Makunin IV 2006. Non-coding RNA. Human Molecular Genetics 115:R17-R29.

McCarthy FM, Gresham CR, Buza TJ, Chouvarine P, Pillai LR, Kumar R, Ozkan S, Wang H, Manda P, Arick T, Bridges SM, Burgess SC 2010. AgBase: supporting functional modeling in agricultural organisms. Nucleic Acids Research 39(Database issue):D497-506.

McCarthy FM, Wang N, Magee GB, Nanduri B, Lawrence ML, Camon EB, Barrell DG, Hill DP, Dolan ME, Williams WP, Luthe DS, Bridges SM, Burgess SC 2006. AgBase: a functional genomics resource for agriculture. BMC Genomics 7:229.

Morin R, Bainbridge M, Fejes A, Hirst M, Krzywinski M, Pugh T, McDonald H, Varhol R, Jones S, Marra M 2008. Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques 45:81-94.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5:621-628.

Mourão MM, Dinguirard N, Franco GR, Yoshino TP 2009a. Phenotypic screen of early-developing larvae of the blood fluke, schistosoma mansoni, using RNA interference. PLoS Neglected Tropical Diseases 3:e502.

Mourão MM, Dinguirard N, Franco GR, Yoshino TP 2009b. Role of the endogenous antioxidant system in the protection of Schistosoma mansoni primary sporocysts against exogenous oxidative stress. PLoS Neglected Tropical Diseases, 3:e550.

Mourier T, Carret C, Kyes S, Christodoulou Z, Gardner PP, Jeffares DC, Pinches R, Barrell B, Berriman M, Griffiths-Jones S, Ivens A, Newbold C, Pain A 2008. Genome-wide discovery and verification of novel structured RNAs in Plasmodium falciparum. Genome Research 18:2:281-92.

Nagai K, Oubridge C, Kuglstatter A, Menichelli E, Isel C, Jovine L 2003. Structure, function and evolution of the signal recognition particle. EMBO Journal 22(14):3479-85.

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science. 320:1344-1349.

Nawrocki EP, Eddy SR 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 29(22):2933-2935.

Nawrocki EP, Kolbe DL, Eddy SR 2009. Infernal 1.0: inference of RNA alignments. Bioinformatics 25(10):1335-1337.

Nilsen TW 1993. Trans-splicing of nematode premessenger RNA. Annual Review of Microbiology 47:413-440.

Nilsson D, Gunasekera K, Mani J, Osteras M, Farinelli L, Baerlocher L, Roditi I, Ochsenreiter T 2010. Spliced Leader Trapping Reveals Widespread Alternative Splicing Patterns in the Highly Dynamic Transcriptome of Trypanosoma brucei. PLoS Pathogens 6(8):e1001037.

Ostermayer AL, Passos ADC, Silveira AC, Ferreira AW, Macedo V, Prata AR 2011. The National

Survey of seroprevalence for evaluation of the control of Chagas disease in Brazil (2001-2008). Revista da Sociedade Brasileira de Medicina Tropical 44:108-121.

Palenchar JB, Bellofatto V 2006. Gene transcription in trypanosomes. Molecular & Biochemical Parasitology 146:135-141

Patro R, Mount SM, Kingsford C 2013. Sailfish: alignment-free isoform quantification from RNA-seq reads using lightweight algorithms. Nature Biotechnology 32:462-464.

Pearce EJ, Freitas TC 2008. Reverse genetics and the study of the immune response to schistosomes. Parasite Immunology 30:215-221.

Pessôa SB, Martin AV 1982. Parasitologia Médica - 11ª edição. Guanabara Koogan,

Pouchkina-Stantcheva NN, Tunnacliffe A 2005. Spliced Leader RNA-Mediated trans-Splicing in Phylum Rotifera. Molecular Biology and Evolution 22(6):1482-1489.

Prata A, Macedo V, Pires LL 1977. Tratamento da doença de Chagas com o benzonidazol. Sociedade Brasileira de Medicina Tropical 27.

Prata A, Macedo V, Porto G, Santos I, Cerisola JA, Silva N 1975. Tratamento da Doenca de Chagas pelo Nifurtimox (Bayer 2502). Revista da Sociedade Brasileira de Medicina Tropical 9(6).

Preußer C, Jaé N, Bindereif A 2012. mRNA splicing in trypanosomes. International Journal of Medical Microbiology 302(4-5):221-224.

Protasio AV, Tsai IJ, Babbage A, Nichol S, Hunt M, Aslett MA, De Silva N, Velarde GS, Anderson TJ, Clark RC, Davidson C, Dillon GP, Holroyd NE, LoVerde PT, Lloyd C, McQuillan J, Oliveira G, Otto TD, Parker-Manuel SJ, Quail MA, Wilson RA, Zerlotini A, Dunne DW, Berriman M 2012. A systematically improved high quality genome and transcriptome of the human blood fluke Schistosoma mansoni. PLoS Neglected Tropical Diseases 6:e1455.

Rajkovic A, Davis RE, Simonsen JN, Rottman FM 1990. A spliced leader is present on a subset of mRNAs from the human parasite Schistosoma mansoni. Proceedings of the National Academy of Sciences of the United States of America 87:8879-8883.

Rassi Jr A, Rassi A, Marin-Neto JA 2010. Chagas disease. The Lancet 375(9723):1388-1402.

Regis-da-Silva CG, Freitas JM, Passos-Silva DG, Furtado C, Augusto-Pinto L, Pereira MT, DaRocha WD, Franco GR, Macedo AM, Hoffmann JS, Cazaux C, Pena SD, Teixeira SM, Machado CR 2006. Characterization of the Trypanosoma cruzi Rad51 gene and its role in recombination events associated with the parasite resistance to ionizing radiation. Molecular and Biochemical Parasitology 149:191-200.

Reiner R, Ben-Asouli Y, Krilovetzky I, Jarrous N 2006. A role for the catalytic ribonucleoprotein

RNase P in RNA polymerase III transcription. Genes & Development, 20(12):1621-1635.

Requena-Méndez A, Albajar-Viñas P, Angheben A, Chiodini P, Gascón J, Muñoz J, Chagas Disease COHEMI Working Group 2014. Health Policies to Control Chagas Disease Transmission in European Countries. PLoS Neglected Tropical Disease 8(10):e3245.

Rinn JL, Chang HY 2012. Genome regulation by long noncoding RNAs. Annual Review of Biochemistry 81.

Robinson MD, McCarthy DJ, Smyth GK 2010. EdgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:1.

Rosenblad MA, López DM, Piccinelli P, Samuelsson T 2006. Inventory and analysis of the protein subunits of the ribonucleases P and MRP provides further evidence of homology between the yeast and human enzymes. Nucleic Acids Research 34(18):5145-5156.

Ross LH, Freedman JH, Rubin SC 1995. Structure and expression of novel spliced leader RNA genes in Caenorhabditis elegans. The Journal of Biological Chemistry 270(37):22066-22075.

RNAcentral Consortium. 2014. RNAcentral: an international database of ncRNA sequences. Nucleic Acids Research gku991.

Sather S, Agabian N 1985. A 5' spliced leader is added in trans to both alpha- and beta-tubulin transcripts in Trypanosoma brucei. Proceedings of the National Academy of Sciences of the United States of America 82(17):5695-5699.

Schattner P, Brooks AN, Lowe TM 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs Nucleic Acids Research 33(2): W686-689.

Schofield CJ 1994. Triatominae: biology & control. Eurocommunica Publications ISBN:1898763003

Senkovich O, Bhatia V, Garg N, Chattopadhyay D 2005. Lipophilic antifolate trimetrexate is a potent inhibitor of Trypanosoma cruzi: prospect for chemotherapy of Chagas' disease. Antimicrobial agents and chemotherapy 49(8):3234-3238.

SeqClean: https://sourceforge.net/projects/seqclean

Silva MRG, Tosar JP, Frugier M, Pantano S, Bonilla B, Esteban L, Serra E, Rovira C, Robello C, Cayota A 2010. Cloning, characterization and subcellular localization of a Trypanosoma cruzi argonaute protein defining a new subfamily distinctive of trypanosomatids. Gene 466(1-2):26-35.

Skelly PJ, Da'dara A, Harn DA 2003. Suppression of cathepsin B expression in Schistosoma mansoni by RNA interference. International Journal for Parasitology 33:363-369.

Smithers SR, Terry RJ 1965. The infection of laboratory hosts with cercariae of Schistosoma mansoni and the recovery of the adult worms. Parasitology 55:695-700.

Symons RH 1992. Small catalytic RNAs. Annual Review in Biochemistry 61: 641-671.

Sorek R, Cossart P 2010. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. Nature Reviews Genetics 11:9-16.

Stover NA, Kaye MS, Cavalcanti AR 2006. Spliced leader trans-splicing. Current Biology 16(1):R8-9.

Stover NA, Steele RE 2001. Trans-spliced leader addition to mRNAs in a cnidarian. Proceedings of the National Academy of Sciences of the United States of America 98:5693-5698.

Sturm NR, Vargas NS, Westenberger SJ, Zingales B, Campbell DA 2003. Evidence for multiple hybrid groups in Trypanosoma cruzi. International Journal for Parasitology 33(3):269-279.

Sturm NR, Campbell DA 2010. Alternative lifestyles: The population structure of Trypanosoma cruzi. Acta tropica 115(1):35-43.

Subileau M, Barnabé C, Douzery EJP, Diosque P, Tibayrenc M 2009. Trypanosoma cruzi: New insights on ecophylogeny and hybridization by multigene sequencing of three nuclear and one maxicircle genes. Experimental Parasitology 122(4):328-337.

Takeda GK, Campos R, Kieffer J, Moreira AA, Amato Neto V, Castilho VLP, Pinto PLS, Duarte MIS 1986. Effect of gamma rays on blood forms of Trypanosoma cruzi. Experimental study in mice. Revista Insternacional de Medicina Tropical Sao Paulo 28: 15-18.

Taschner A, Weber C, Buzet A, Hartmann RK, Hartig A, Rossmanith W 2012. Nuclear RNase P of Trypanosoma brucei: A Single Protein in Place of the Multicomponent RNA-Protein Complex. Cell Reports 2(1):19-25.

Taylor J & Bestetti RB 2009. Chagas' disease and its toll on the heart. European Heart Journal 30:2063-2072.

Tessier LH, Keller M, Chan RL, Fournier R, Weil JH, Imbault P 1991. Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in Euglena. EMBO Journal 10(9):2621-2625.

Uhlenbeck OC 1987. A small catalytic oligoribonucleotide. Nature 328:590-600.

Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP 2011. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell 147(7):1537-1550.

Univec Database http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html

Vandenberghe AE, Meedel TH, Hastings KE 2001. mRNA 5'-leader trans-splicing in the chordates. Genes and Development 15:294-303.

Verdun RE, Di Paolo N, Urmenyi TP, Rondinelli E, Frasch AC, Sanchez DO 1998. Gene discovery through expressed sequence Tag sequencing in Trypanosoma cruzi. Infection and Immunity, 66(11):5393-5398.

Vieira HGS, Grynberg P, Bitar M, Pires SF, Hilário HO, Macedo AM, Machado CR, Andrade HM, Franco GR 2014. Proteomic Analysis of Trypanosoma cruzi Response to Ionizing Radiation Stress. 9(5):e97526.

WHO 2013. Weekly epidemiological record Nº8 (81-88) from February 2013.

WHO 2014. Fact sheet N°115 from February 2014.

Wallace A, Filbin ME, Veo B, McFarland C, Stepinski J, Jankowska-Anyszka M, Darzynkiewicz E, Davis RE 2010. The nematode eukaryotic translation initiation factor 4E/G complex works with a trans-spliced leader stem-loop to enable efficient translation of trimethylguanosine-capped RNAs. Molecular and Cellular Biology 30:1958-1970.

Wang Z, Gerstein M, Snyder M 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics 10:57-63.

Weatherly DB, Boehlke C, Tarleton RL 2009. Chromosome level assembly of the hybrid Trypanosoma cruzi genome. BMC Genomics 10(1):255.

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V 2008. Database resources of the National Center for Biotechnology Information. Nucleic Acids Research 36:D13-21.

World Health Assembly 2010. Chagas disease: control and elimination. Resolution 63.20.

World Health Organization 2014. Chagas disease (American Tripanosomiasis). Fact Sheet 340.

Wu XJ, Sabat G, Brown JF, Zhang M, Taft A, Peterson N, Harms A, Yoshino TP 2009. Proteomic analysis of Schistosoma mansoni proteins released during in vitro miracidium-to-sporocyst transformation. Molecular and Biochemical Parasitology 164:32-44.

Yoshino TP, Dinguirard N, Mourão Mde M 2010. In vitro manipulation of gene expression in larval Schistosoma: a model for postgenomic approaches in Trematoda. Parasitology 137:463-483.

Yoshino TP, Laursen JR 1995. Production of Schistosoma mansoni daughter sporocysts from mother sporocysts maintained in synxenic culture with Biomphalaria glabrata embryonic (Bge) cells. The Journal of Parasitology 81:714-722.

Zerlotini A, Heiges M, Wang H, Moraes RL, Dominitini AJ, Ruiz JC, Kissinger JC, Oliveira G 2009. SchistoDB: a Schistosoma mansoni genome resource. Nucleic Acids Research 37:D579-582.

Zingales B, Andrade SG, Briones MRS, Campbell DA, Chiari E, Fernandes O, Guhl F, Lages-Silva

E, Macedo AM, Machado CR, Miles MA, Romanha AJ, Sturm NR, Tibayrenc M, Schijman AG 2009. A new consensus for Trypanosoma cruzi intraspecific nomenclature: second revision meeting recommends TcI to TcVI. Memórias do Instituto Oswaldo Cruz 104(7):1051-1054.

Zingales B, Miles MA, Campbell DA, Tibayrenc M, Macedo AM, Teixeira MM, Schijman AG, Llewellyn MS, Lages-Silva E, Machado CR, Andrade SG, Sturm NR 2012. The revised Trypanosoma cruzi subspecific nomenclature: rationale, epidemiological relevance and research applications. Infection, Genetics and Evolution 12(2):240-53.

Zingales B, Pereira MES, Oliveira RP, Almeida KA, Umezawa ES, Souto RP, Vargas N, Cano MI, Franco da Silveira J, Nehme NS, Morel CM, Brener Z, Macedo AM 1997. Trypanosoma cruzi genome project: biological characteristics and molecular typing of clone CL Brener. Acta tropica 68(2):159-173.