Université de Montréal

Study of *cox1* trans-splicing in *Diplonema papillatum* mitochondria

par

Yifei Yan

Département de biochimie
Faculté de médecine

Mémoire présenté à la Faculté de médecine
en vue de l'obtention du grade de maîtrise
en biochimie

janvier 31, 2011

Université de Montréal
Faculté de médecine

Ce mémoire intitulé:

Study of *cox1* trans-splicing in *Diplonema papillatum* mitochondria

présenté par :

Yifei Yan

a été évalué par un jury composé des personnes suivantes :

Dr. Pascal Chartrand

Dre. Gertraud Burger

Dre. Léa Brakier-Gingras

TABLE OF CONTENTS

List of Figures and Tables

List of Abbreviations

A6: mitochondrial ATPase subunit 6
AMV: avian myeloblastosis virus
ATP: adenosine triphosphate
BLAST: basic alignment search tool
cDNA: complementary DNA
cox1: cytochrome c oxidase subunit 1
CTP: cytosine triphosphate
CYB: apocytochrome b
DNA: deoxyribonucleic acid
dNTP: deoxy-ribonucleotides
DTT: dithiothreitol
EBS: exon binding site
EDTA: ethylenediaminetetraacetic acid
eIF: eukaryotic translation initiation factor
EtBr: ethidium bromide
gRNA: guide RNA
GTP: guanosine triphosphate
IBS: intron binding site
IPTG: isopropyl β-D-1-thiogalactopyranoside
IRIC: l'Institute de recherche en immunologie et en cancérologie (IRIC)
LB: lysogeny broth
mRNA: messenger RNA
nad5: NADH dehydrogenase subunit 5
ND7: NADH dehydrogenase (mitochondrial complex I) subunit 7
NEB: New England Biolabs
PCR: polymerase chain reaction
PNK: polynucleotide kinase
Prp8: pre-mRNA processing factor 8 homolog (*S. cerevisiae*)
RNA: ribonucleic acid
rns: small subunit ribosomal RNA
rRNA: ribosomal RNA
RT: reverse transcription
SAM : S-adenosyl-methionine
SL-RNA: spliced leader RNA
snRNA: small nuclear RNA
snRNP: small nuclear ribonucleoprotein
sRNA: small RNA
TAP: tobacco acid phosphatase
tRNA: transfer RNA
TUTase: terminal U transferase
UTP: uridine triphosphate
UTR: untranslated region

For my parents, Lillian, and Amelia; in the memory of Dr. Jerry Price

ACKNOWLEDGEMENTS

I would like to thank Dr. Gertraud Burger for giving me the wonderful opportunity to study as a master student under her supervision. Her extraordinary vision and unparalleled patience helped me through out the study. I thank my parents for their altruistic support even during their difficult times. I thank Shona Teijeiro, Lise Forget, and the rest of the lab for their day to day assistance on my project. I thank Dr. Franz Lang for his guidance and motivation during the preparation of the manuscript for our publication in *Trends in Genetics*. I thank Dr. Pascal Chartrand, Dr. Normand Brisson and their students for generously lending me the lab equipments and helping me on experimental techniques; and I thank Dr. Luc DesGroseillers, the parrain, for useful advices on my study and my project.

RÉSUMÉ

Diplonema papillatum est un organisme unicellulaire qui vit dans l’océan. Son génome mitochondrial possède une caractéristique spéciale: tous les gènes sont brisés en de multiples fragments qui s’appellent modules. Chaque module est codé par un chromosome différent. L’expression d’un gène exige des épissages-en-trans qui assemblent un ARN messager complet à partir de tous les modules du gène. Nous avons précédemment montré que le gène *cox1* est encodé dans neuf modules avec six Us non encodés entre le module 4 et le module 5 de l’ARN messager mature [1]. Nous n’avons identifié aucune séquence consensus connue de site d’épissage près des modules. Nous spéculons qu’un ARN guide (gRNA) a dirigé l’épissage-en-trans du gène *cox1* par un mécanisme qui est semblable à l’édition d’ARN par l’insertion/la suppression des Us chez les kinétoplastides, le groupe sœur des diplonémides. Nous avons trouvé que les six Us sont ajoutés au bout 3’ de l’ARN d’une façon semblable à ceux ajoutés par le TUTase lors de l’édition de l’insertion des Us chez les kinétoplastides. Nous avons construit des profils de gRNA de l’épissage-en-trans avec les expressions régulières basé sur notre connaissance des gRNAs dans l’édition d’ARN chez les kinétoplastides. Selon la complémentarité partielle entre le gRNA et les deux modules adjacents, nous avons généré des amorces pour RT-PCR visant à détecter des séquences qui sont assorties à un des profils de gRNA. Une expérience pilote in vitro n’a pas permis de reconstituer l’épissage-en-trans des modules 3, 4, et 5, suggérant que nous devons améliorer nos techniques.

Mots clés: *Diplonema papillatum*, mitochondrie, *cox1*, épissage-en-trans, ARN guide, kinétoplastides, RT-PCR

ABSTRACT

Diplonema papillatum is a single cellular organism that lives in the ocean. Its mitochondrial genome possesses a special feature: all genes are fragmented in multiple pieces that are called modules and each module is encoded by a different chromosome. Expression of a gene requires trans-splicing that successfully assemble a full-length mRNA from all modules of the gene. It was previously shown that the *cox1* gene is encoded in nine modules that are all located on different chromosomes; moreover, a stretch of six non-encoded Us exist between Module 4 and 5 in the mature mRNA [1]. No consensus sequence of known splicing sites was identified near the modules. We speculate that trans-splicing of the *cox1* gene is directed by guide RNAs (gRNAs) via a mechanism that is similar to U-insertion/deletion editing in kinetoplastids, the sister group of diplonemids. We have detected populations of small RNA molecules that could come from mitochondrial. We found that the six Us were added to the 3' end of Module 4 in a similar way to the Us added by the TUTase in kinetoplastid U-insertional editing. Sequence profiles of possible trans-splicing gRNAs were constructed in regular expressions based on our knowledge of known gRNAs in kinetoplastid RNA editing. According to the complementarity between the gRNA and the two adjacent modules, primers were designed for RT-PCR that aims to detect gRNA sequences. Among the results, we identified sequences that match or partially match the gRNA profiles. A pilot in vitro assay did not reconstitute trans-splicing of module 3, 4 and 5, suggesting that further technical improvements are needed.

Key words: *Diplonema papillatum*, mitochondrion, *cox1*, trans-splicing, guide RNA, kinetoplastids, RT-PCR

1. INTRODUCTION

Splicing originally refers to the process that removes non-coding sequences, introns, which interrupt the coding sequences from the primary transcripts so that the coding sequences can correctly join and form a contiguous messenger RNA ready for translation. In order to better study trans-splicing, a variation of the conventional intron-splicing, we will briefly recapture our knowledge about conventional splicing.

1.1 Intron types and distribution in nature

Introns are sequences present in the primary transcript that are subsequently removed and thus absent from the mature RNA molecule. Introns are categorized into tRNA or archaeal introns, group I introns, group II introns, and spliceosomal introns.

1.1.1 tRNA and/or archaeal introns

Despite their name, the phylogenetic distribution of tRNA and/or archaeal introns is ubiquitous, in Eukarya, Archaea and Bacteria. In bacteria, tRNA introns are self-splicing [2]; while in archea and eukaryotes, tRNA splicing requires a cascade of enzymes including endonucleases that specifically recognize the splice sites [3]. Such recognition depends on the conserved structural features of the tRNA such as the anticodon loop and the bulge-helix-bulge motif (BHB). Eukaryotic nucleus-encoded tRNA introns are always located between the first two nucleotides 3' to the anticodon.. The cleavage site is determined by the distance from the highly conserved structural features of the mature molecule, the 3' end of the anticodon stem [4]. Archaebacterial tRNA introns are found in additional locations of the tRNA: in the anticodon stem, the anticodon loop [5], or the extra arm [6]. The recognition of intron/exon boundaries requires special sequences and tertiary structures at the boundary, where a bulge-helix-bulge structure is formed by a stem of four base-pairs flanked by two 3-nt loops [7-9].

1.1.2 Group I and group II introns

Both group I and II introns can be found in mitochondria. They differ by their secondary and tertiary structures as well as mechanisms of splicing. Group I and group II introns are often incorrectly referred to as autocatalytic [10], due to the fact that several members can carry out splicing *in vitro*. In fact, many of them do not possess such property and additional protein factors are required for splicing [11-13]; moreover, none of the mitochondrial introns in plants are self-splicing (see discussion in [14, 15]).

1.1.2.1. Group I introns

Group I introns are found in the nucleus and organelles of diverse eukaryotes such as fungi, plants and numerous protist groups. They are absent from mitochondria of animals, except for several early diverging metazoa such as corals, sponges, sea anemones, and placozoan (for references, please see [16]). Group I introns are rarely found in Eubacteria and bacterial phages and are completely absent in Archea [17-19]. The secondary structure of group I introns usually consists of approximately 10 paired regions (P1-10) that fold the molecule into a conserved secondary structure [20] (Fig. 1). Although in some cases particular paired regions may be absent in a group I intron, the sequences surrounding P1, P3, P4, P6 and P7 constitute the "intron core" of the secondary and tertiary structure. Group I introns are removed from the precursor RNA sequence by a two-step mechanism. In the first step, an external guanine nucleotide attacks the 5' splice site and causes a cut in the phosphate backbone, then it is covalently bound to the newly cut site on the intron. The second step involves the attack of the 3' splice site by the 3' end OH of the cleaved exon from step 1. This results in the ligation of the exons and the release of the intron sequence [17].

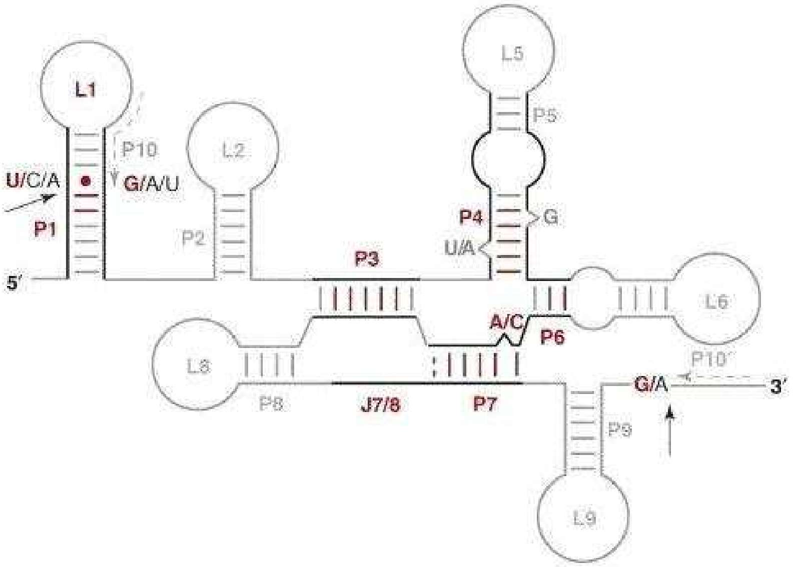


Figure 1. Secondary structure of a typical mitochondrial group I intron
Black lines indicate the conserved intron core. Pairings (P), loops (L) and junctions (J) are numbered according to conventions that emphasize secondary structure. Universally conserved structural elements are shown in red, while variable portions are shown in gray. The minimum number of conventional base pairs (A-U, G-C and G.U) in helices is shown in red. One base pairing in P7 (broken) is absent in several instances. It should be noted that J3/4 and J3/7 vary in size and have, therefore, not been listed as part of an invariant core structure. P10 and P10′ are the two halves of a helical interaction. This figure is adapted from [14].

1.1.2.2 Group II introns

Group II introns preferentially occur in the organelles of land plants, fungi, and protists [18]. They were rarely identified in Archaea and Bacteria (for reviews see [14, 15]) and they are absent from the mitochondrial DNA of metazoa. The secondary structure of a group II intron usually consists of six helical domains, extending from a central hub, and denoted as domains I to VI [21]. Though, the most distinctive feature of group II introns is domain V, which contains an AGC triad in the paired region crucial for the catalytic activity of the ribozyme [22], and a GNRA-tetraloop motif (N denotes any nucleotide and R stands for A or G). This GNRA motif is highly conserved and can be reliably used as a landmark sequence for group II introns [14]. The molecular process of group II intron splicing involves two successive transesterification reactions and it uses the 2'-OH of an internal and bulged adenosine residue to initiate the attack on the 5' splice site [18]. As the consequence, the group II intron that undergoes splicing is excised as a lariat structure [23]. Both steps of the reaction are reversible; however, the reverse reaction of the second step is much slower than the forward one under physiological conditions, rendering the reverse-splicing an inefficient process. The reverse-splicing process can be manipulated by altering the conditions to be as efficient as the forward one [24], and interestingly, the reverse reaction works efficiently with DNA targets as well [25].

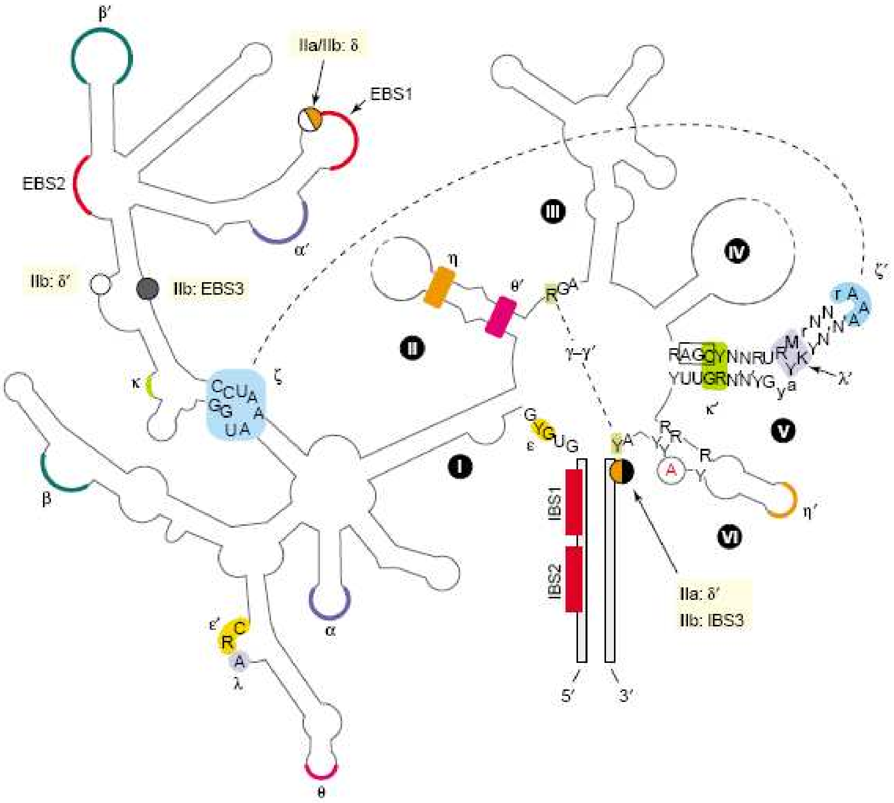


Figure 2. Typical secondary structure of a group II intron

A Group II intron core structure shows six domains (DI–DVI) that extend from a central hub (flanking exons in grey). Uppercase letters in domain V reflect >90%, lowercase 80–90% conservation; Y = pyrimidine; R = purine; M = A or C; K = G or U; N–N′ = canonical basepair. Domain V is shown as 9 bp + 5 bp separated by a dinucleotide bulge. The boxed nucleotidtes (AGC) are among those essential for catalysis as determined by detailed site-directed mutagenesis and self-splicing in vitro. The bulged A in domain VI is shown in red. Tertiary interactions are shown by color-coded Greek symbols, two of which are illustrated by dashed lines (namely γ–γ′ and ζ–ζ′). EBS = exon binding site (red), IBS = intron binding site (red). Flanking exons are shown in grey and the bulged "A" in DVI is shown in red.

1.1.3 Spliceosomal introns

Spliceosomal introns are common in nuclear genes of eukaryotes, but absent from mitochondrial and chloroplast genes, as well as eubacterial and archaeal genes. Exon-intron junctions of eukaryotic spliceosomal introns are highly conserved, with most of them having GU at the 5' boundary of the intron and AG at its 3' boundary. A minor class of introns in yeast, vertebrates and invertebrates has different splice site sequences, that is, AUAUCUU at the 5' end and CAC at the 3' end [26, 27]. The spliceosome is a ribonucleoprotein complex that contains five uridine-rich small nuclear RNAs, or snRNAs, designated as U1, U2, U4, U5 and U6. The snRNAs are highly conserved and range between 60 to 300 nt in length. Splice sites are recognized by small nuclear ribonucleoprotein particles (snRNPs) in the spliceosome, which are complexes formed by about 150 protein factors with 6 snRNAs in humans [28]. It was shown that U1 and U2 recognize the 5' and 3' splice junctions, respectively, via an ATP- and temperature-dependent base-pairing mechanism [29, 30]. Together with U4-U6 and U5 snRNAs, the snRNPs align the active sites and form the catalytic core that carries out splicing (for a review see [31]). In vitro studies showed that intron removal by the spliceosome follows a two-step reaction mechanism similar to that of group II introns [32, 33]. Further, sequence and structure similarities exist between the U6 snRNA and the domain V (DV) of group II introns [34, 35]. Figure 3 depicts the similarity between group II intron and the U6 RNA. Due to the parallelism between the two systems, it was suggested that U2 and U6 alone could form the catalytic center of the spliceosome (for review, see [36]). However, the spliceosome likely differs from the group II intron in the composition of the active site, because a protein called Prp8 was consistently found within the active site of spliceosome [37], while no protein was found to be associated with the active site of DV of the group II intron. Hence, refined structural data and improved techniques that overcome the difficulties imposed by the spliceosome assembly process are still needed to test whether its catalytic activity is truly a protein-free process.

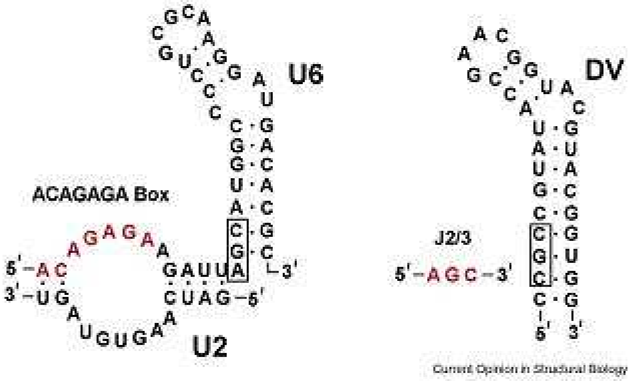


Figure 3. Similarity between U6 snRNA and DV of group II introns
The U6 RNA is on the left and the domain V of the *Oceanobacillus iheyensis* group II intron is on the right. The catalytic triad (boxed) is located at the base of the stem loop structures, with the two-nucleotide bulge spaced five base pairs away. The ACAGGA box (red) may form base triple interactions with the triad of the U6 snRNA, analogous to that observed between J2/3 and domain V [34].

1.2 Intron trans-splicing

Trans-splicing originally refers to the collaborative removal process involving two separate primary transcripts that each holds a moiety of an intron; consequently, the two independent transcripts are joined to generate a new continuous RNA molecule. A few variations of trans-splicing were discovered during the past two decades and they fall into the following categories: spliceosome-dependent, discontinuous group II, and recently, group I intron trans-splicing.

1.2.1 Spliceosome-dependent trans-splicing

This process was first demonstrated in the nucleus of *Trypanosoma brucei*, where a 39-nt RNA processed from a 140 nt-long "Spliced Leader RNA" (SL-RNA) is added to the 5' end of every pre-mRNA molecule in order to generate a translatable mRNA. Hence the process was referred to as "spliced-leader type" trans-splicing [38, 39]. The corresponding spliceosome has a somewhat different snRNP composition from that of conventional cis-splicing spliceosomes [40]. SL-trans-splicing was also discovered in the nucleus of worms [41], *Euglena* [42], and *Diplonema* [43]. More recent studies showed that non-SL-type trans-splicing of nucleus-encoded genes in human depends on spliceosomes [44, 45].

1.2.2 Discontinuous group II intron trans-splicing

Trans-splicing that involves discontinuous group II introns [46] was first observed in chloroplasts of land plants [47, 48]. In vitro group II intron trans-splicing was first demonstrated by Konarska's group, showing that intermolecular RNA-RNA hybridization increases the efficiency of trans-splicing significantly [49]; later on, protein factors were found to be required for this process in vivo [50].

1.2.3 Discontinuous group I intron trans-splicing

Very recently, cases of naturally occurring group I trans-splicing were discovered, all of them in mitochondria. The first reported instance is in the placozoan animal *Trichoplax adhaerens* [16] (Appendix 6) and the others in the lycophyte moss *Isoetes engelmannii* [51] and the green alga *Helicosporidium* [52]. In vitro trans-splicing of discontinuous group I introns was previously demonstrated possible for mammalian cells and the molecular requirements are thought to be similar to those of cis-splicing [53].

1.3 Trans-splicing in *Diplonema* mitochondria

None of the above models explains the trans-splicing phenomenon that we observed in *Diplonema* mitochondria (Kiethega et al., submitted). Based on knowledge gained from our preliminary studies, we postulate that *Diplonema* mitochondria represent a new type of trans-splicing, as we detailed below.

1.3.1 What is *Diplonema papillatum*?

D. papillatum is a single-celled bi-flagellated eukaryote that lives in the ocean (Fig. 4). Diplonemids, together with kinetoplastids and euglenids form the phylum of Euglenozoa, which are characterized by discoidal mitochondrial cristae, as well as a distinctive architecture of the flagellar and feeding apparatus [54, 55]. Based on molecular phylogeny, diplonemids and kinetoplastids diverged after euglenids [56, 57]. Figure 5 depicts a schematic phylogenetic tree of Euglenozoa.

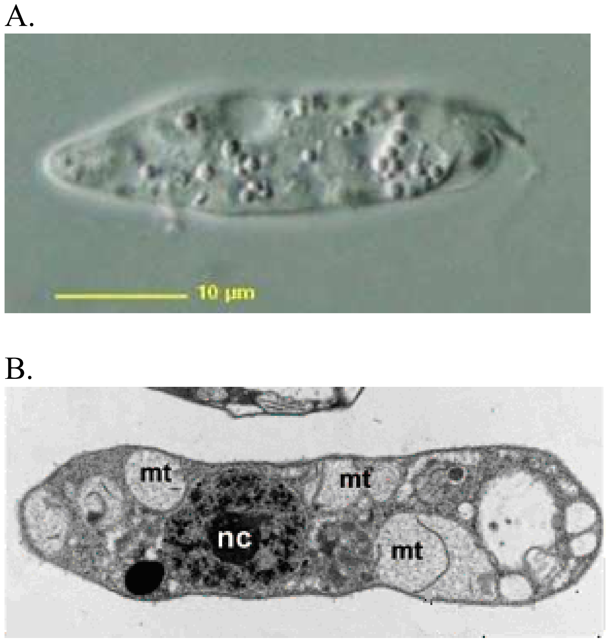


Figure 4. Cell morphology of *D. papillatum*
(A) *D. papillatum* cell under light microscope. (B) Ultrastructure of *D. papillatum* under electron microscopy. The image shows a longitudinally sectioned *Diplonema* cell. Prominent flat cristae are visible in the organellar lumen. nc, nucleus; mt, mitochondrion. Figure is taken from [1].

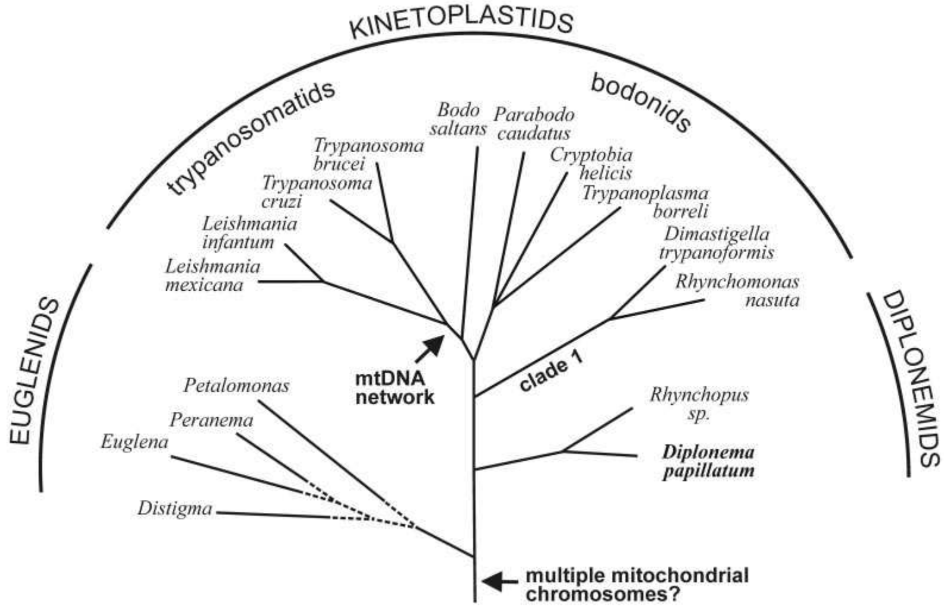


Figure 5. Schematic representation of the phylogenetic position of *D. papillatum*
Schematic phylogenetic tree of Euglenozoa is taken from [1]. The topology is taken from studies by Simpson and coworkers [56, 58] , where bootstrap support is significant for both the common ancestry of Euglenozoa and the diplonemids and kinetoplastids. The branching order within euglenids is tentative. Arrows indicate, as proposed in this report, the time points in evolutionary history when multiple mitochondrial chromosomes and mtDNA networks arose in Euglenozoa. Note that *Parabodo caudatus* was previously designated *Bodo caudatus*.

1.3.2 Mitochondrial genome

The mitochondrion is the "power plant" of a cell in terms of generating most of the ATP molecules required for cellular function. The genes encoded by the mitochondrial genome are involved in a relatively constant set of biological processes that are localized to the mitochondria. The functions of these genes include: respiration, oxidative phosphorylation, and translation. Less frequently, mitochondrial genes are involved in protein import/maturation, transcription, and RNA processing. The median of mitochondrial genome sizes is 30 kbp when animals are included; excluding animals, the median is 50 kbp. The largest mitochondrial genome is found among plants, with a size of 2400 kbp in the family of Cucurbitaceae [59].

1.3.3 Unusual genome organization and gene expression in *D. papillatum* mitochondria.

Akin to trypanosomes and euglenids, *Diplonema* contains one mitochondrion per cell [60]. The total size of the *D. papillatum* mitochondrial genome is probably around 600 kbp. It contains two types of circular chromosomes designated as class A and class B of sizes around 6.2 kbp and 7.2 kbp, respectively. Chromosomes of the same class are quasi-identical in sequence in the "constant region" that spans 95% of the whole length of the chromosome. The residual 5% consists of the "cassette" which holds a coding region and adjacent non-coding unique sequence. It is the cassette that distinguishes individual chromosomes from one another (Fig. 6).

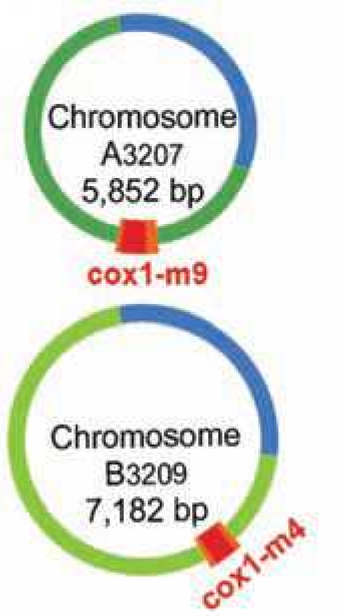


Figure 6. Two mitochondrial chromosomes of *D. papillatum*
The blue and green arcs together represent the constant regions. Blue region sequences are identical for all chromosomes and green region sequences are identical among chromosomes of the same class. The cassette consists of a gene module (red) and the flanking regions (orange) [61].

All mitochondrial genes coding for rRNAs, cytochrome oxidase subunits, NADH dehydrogenase subunits, etc., are found to be fragmented into several pieces and each cassette of a chromosome holds one piece termed a "module" [62]. While contiguous and complete DNA regions coding for these genes were not detected either in nuclear or in mitochondrial DNA, the corresponding contiguous mature transcripts were found. This implies that the assembly of modules occurs at the RNA level, i.e., involves trans-splicing [1, 61].

1.3.4 *D. papillatum* mitochondrial *cox1* trans-splicing differs from all known trans-splicing processes

Marande and Burger [61] showed that mitochondrial cytochrome *c* oxidase subunit 1 (*cox1*) gene is fragmented into nine pieces and its mRNA is assembled from nine pieces of separately transcribed primary RNAs (Fig. 7). This trans-splicing process does not resemble any of the previously reported cases after detailed bioinformatic analyses were conducted (Kiethega et al., submitted). Firstly, in the regions adjacent to the coding modules, none of the highly conserved spliceosomal intron motifs mentioned earlier was detected. Secondly, these regions do not contain tRNA intron sequences.

Thirdly, none of the conserved group I or group II intron sequence motifs was detected in these regions. Lastly, the fact that module junction sites differ from all known insertion points of organellar introns further diminishes the possibility of the presence of discontinuous group I and II introns. The assembly of RNA modules into mRNA proceeds apparently by a new mechanism.

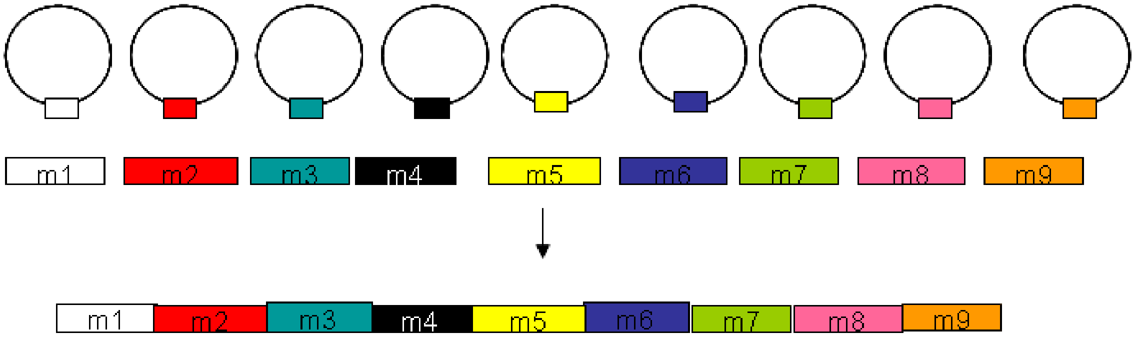


Figure 7. Gene structure of *cox1* in *D. papillatum* mitochondria
The *cox1* gene is encoded in nine individual pieces on different chromosomes. Circles represent different chromosomes and colored rectangles are the gene modules of *cox1*. The nine modules are assembled at the RNA level into a translatable mRNA.

1.4 Trans-splicing could be assisted by helper RNAs

1.4.1 Does anti-sense mRNA assist trans-splicing?

A mechanism was recently identified in the ciliate *Oxytricha trifallax* where larger RNA molecules are used as templates to direct the assembly process of genomic DNA sequences [63]. Such long RNA template could also be present in *D. papillatum* but to direct the assembly of RNA modules. Alternatively, *Diplonema* trans-splicing may be directed by small guide RNA (gRNA) in a way similar to mitochondrial RNA editing in kinetoplastids. The process and machinery involved in ciliate DNA splicing are not known; in contrast, kinetoplastid RNA editing has been intensively studied during the past 20 years.

1.4.2 Guide RNA-directed editing in the mitochondria of kinetoplastids

Kinetoplastids include several parasitic genera, such as *Leishmania* and *Trypanosoma*. The mitochondrial genome of trypanosomes consists of circular DNA molecules, termed maxicircles (14 to 40 kbp) and minicircles (1.0 to 1.4 kbp). There is a single type of maxicircle, present in dozens of copies, which encodes typical mitochondrial genes, and hundreds of different types of minicircles, present in thousands of copies, which encode gRNA [58]. Genes encoded on maxicircles were often found to lack initiation or termination codons; in some cases, they include frameshifts. RNA editing corrects these shortcomings by inserting or deleting precise numbers of uridine (U) residues at distinct positions. Such precision is achieved through the help of the gRNAs (for a review, see [64]).

1.4.3 The discovery of gRNA in trypanosome RNA editing

The Simpson's group radio-labeled small RNA molecules isolated from the trypanosome mitochondria and used them as probes in Southern blot against membrane bound mitochondrial DNA. They detected regions on the mitochondrial minicircles to be gRNA-encoding. Further confirmation was obtained by cloning and sequencing these RNAs and testing them in vitro [65]. Since then, a plethora of gRNA molecules were identified in kinetoplastid mitochondria assisting the editing events. These gRNAs were reported to contain a 5' anchor region that is usually around 5-12 nt in length and pairs perfectly with the pre-mRNA downstream of the edited site to hold the pre-mRNA in position (Fig. 8). RNA duplex stabilization involves not only Watson-Crick but also G-U base pairing. The central part of the gRNA is the “guiding region” and it appears in two forms: in insertion editing, the guiding region contains additional “guiding nucleotides” that act as a template and they will pair with the newly inserted Us; in deletion editing, the guiding region contains no nucleotide that pairs with the Us in the pre-mRNA and the un-matched Us will be deleted. The guiding region could be as long as 25-35 nt and it may direct editing at multiple neighboring sites on the pre-mRNA [66]. Non-guiding nucleotides may also exist in the guiding region (curved lines in the “guiding region” depicted in Figure 8). They do not bind to pre-mRNA and they can form a bulge when gRNA initially hybridizes to mRNA. Base-pairing may occur within the bulge itself and secondary structures can form; however, they do not affect editing [67]. Upon the completion of editing, the guiding nucleotides must pair with the edited sequence and the non-guiding nucleotides would pair with the nucleotides next the editing site. The anchoring region at the 3’ of the gRNA is also called a “tether”, which pairs in many cases with about 11 nt of purine-rich sequence in the pre-mRNA further to its 5’ end (5 to 24 nt away from the editing site) [66]. In the 3’ tether of the gRNA, mutations that increases base-pairing stability with 5’ moiety of the mRNA decreases the formation of edited products [68]. At the very 3’ end of the gRNA, there is usually an oligo-U tail of 5-24 nt that binds to the purine-rich region in pre-mRNA further upstream, providing additional stabilization.

In summary, the following features are common for gRNAs to mediate the recognition of the editing site: 55-75 nt in size, perfect 5’ base-pairings between gRNA and pre-mRNA, single stranded guiding nucleotides and single stranded region of pre-mRNA, 3’ end tethering of the pre-mRNA (Fig. 8). The gRNA sequences found in kinetoplastids so far are all co-linear with the edited mRNA around the editing sites, i.e., there are no bulges left in the gRNA-mRNA duplex once editing is completed.

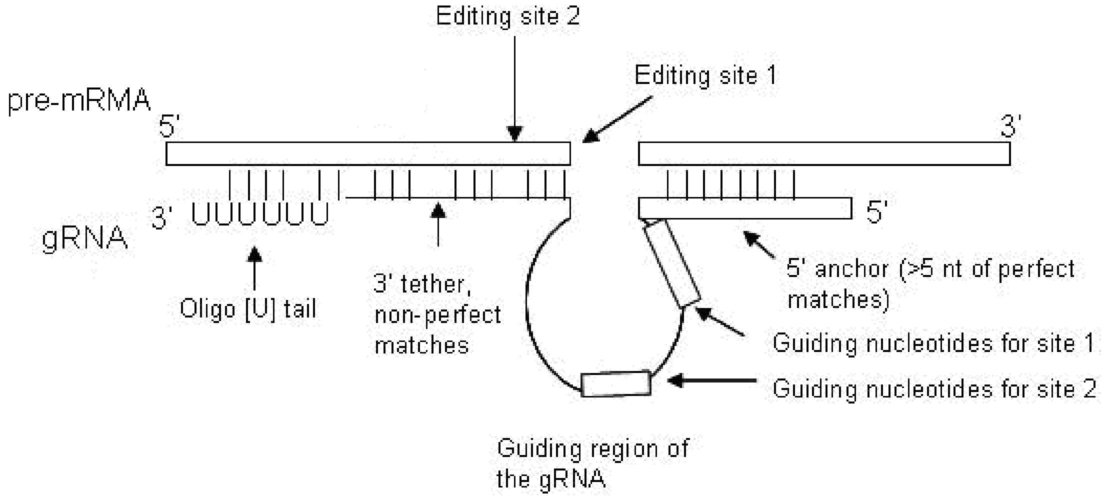

Editing site 2
Editing site 1
pre-mRMA
5'
3'
3' UUUUUU
5'
gRNA
Oligo [U] tail
3' tether, non-perfect matches
5' anchor (>5 nt of perfect matches)
Guiding nucleotides for site 1
Guiding nucleotides for site 2
Guiding region of the gRNA

Figure 8. Common features of the pre-mRNA-gRNA duplex in kinetoplastids
The top strand represents the mRNA with the 5' end on the left. The bottom strand represents the gRNA with the 5' end on the right. The square-shaped sequence represents sequences that are coding or complementary to coding sequence. The line represents sequences that are non-coding. The gRNA in this figure encodes two U-insertion sites. The editing of the two sites proceeds in the 3' to 5' direction along the pre-mRNA.

1.4.4 The trypanosome RNA editing machinery

Experimental evidence shows that U-insertion/deletion editing in trypanosomes mitochondria involves endonucleolytic cleavage and ligation. A multi-protein complex of about 20S, called editosome, carries out RNA editing in a stepwise fashion [64]. Specific recognition of editing sites on the target pre-mRNA via the help of gRNA molecules is the first step of editing (Fig. 9A). The gRNA molecule contains nucleotides complementary to the sequence both upstream and downstream of the editing site. The guiding nucleotides are between the complementary nucleotides on both ends and they mismatch the mRNA at the nucleotides to be edited. A cleavage will be made by an endonuclease of the editosome on the pre-mRNA at the 3' of the first base that is not paired to the gRNA's 5' anchor [69] (Fig. 9B). This creates a nick on the pre-mRNA, exposing the 3' OH group of the cleaved strand. In U-insertion editing, the next step is that TUTase adds a stretch of Us to the exposed 3' end of the pre-mRNA, then excess number of Us will be trimmed by a U-exonuclease to the length specified by the guiding nucleotides in the guide RNA; in the case of U-deletion editing, only a U-exonuclease will be involved in this step to cut the excessive Us [70-72] (Fig. 9C). The U-removal process leaves a phosphate group at the 3' end of the upstream moiety of the pre-RNA, which renders the 3' end non-ligatable. Once the template sequence matches the target sequence, a phosphatase removes the phosphate group [73] and the 3' of the edited site is religated to the 5' by an RNA-ligase [74] (Fig. 9D). After the editing of a given site is completed, the gRNA is displaced by a helicase, and the editosome proceeds to the next editing site in the 5' direction along the pre-mRNA. In summary, under the direction of gRNA molecules, the trypanosomal RNA editosome executes the following enzymatic activities: RNA endonuclease, TUTase, phosphatase, RNA ligase and helicase.

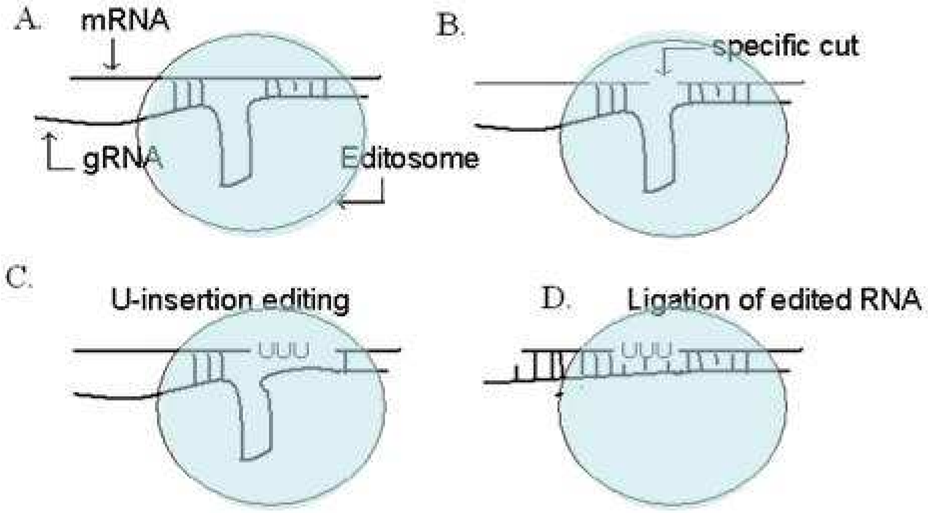


Figure 9. A schematic representation of U-insertion editing in kinetoplastids
The mRNA is on top and the guide RNA is at the bottom. The ellipsoid shape represents the editosome complex. An endonucleolytic cut is made in the mRNA at the 5' end of the nucleotide that pairs with the last nucleotide of the 5' anchor region of the gRNA (see Fig. 8 for the 5' anchor region of gRNA). Multiple Us are subsequently added to the 3' end of the cut and trimmed to the exact length that is specified by the complementary nucleotides in the gRNA. The nick is then ligated and upon completion of editing, the mRNA perfectly pairs with the gRNA at the editing site.

1.4.5 Experimental demonstration of gRNA function in kinetoplastid mitochondria

As soon as RNA editing was discovered in kinetoplastids, several *in vitro* systems were set up attempting to reproduce this process.

1.4.5.1 In vitro systems at the organelle level

Isolated mitochondrial vesicles from *T. brucei* were shown to incorporate ^{32}P-UTP into endogenous RNAs post-transcriptionally [75]. A similar observation was made using a mitochondrial extract of *L. tarentolae* [76]. These experiments confirmed the outcome of U-insertion editing; however, they only monitored the general incorporation of radio labeled nucleotides, hence are not sufficient to study the U-insertion editing mechanism.

1.4.5.2 In vitro systems based on a mitochondrial extract

The mitochondrial extract based system has been refined to test specific hypotheses on requirements of RNA editing. A study of gRNA-dependent editing was reported for both U-insertion and U-deletion in vitro [77]. In a *T. brucei* system, synthetic gRNA was added and the U-insertions and U-deletions occurred at the expected editing sites of *A6* mRNA [78-80]. Despite the low efficiency of the *T. brucei* in vitro system, the signals of the labeled intermediates are strong enough to allow detection. A similar study was conducted with a *L. tarentolae* system with exogenously supplied synthetic gRNA, where blockage of the 3'-OH of the U-tail of the gRNA, which is required for the hypothesized transesterification reaction that supplies the Us, did not have any effect on editing. This result showed that transesterification was not involved in the transfer of Us [81]. Using the mitochondrial extract-based in vitro system, gRNA-independent editing was also studied. For the mitochondrial genes *CYB* and *ND7* in *L. tarentolae*, primer extension assay revealed that the 3' UTR of the mRNA can fold back and form a short RNA duplex with sequences upstream and downstream of the editing site. Hence the mRNA acts as a gRNA for its own editing [82]. Further studies with this in vitro system showed that an AU-rich sequence just upstream of the edited site was responsible for the activation of U-insertion editing in the absence of gRNA [83].

1.4.5.3 In vitro system based on purified protein/complexes

Once the mitochondrial extract-based in vitro system was successfully tested for the editing process, proteins and complexes from the extract could be isolated and tested for their roles. For example, in *T. brucei*, two ~57 kDa and one ~50 kDa bands were isolated as part of the minimal editing complex. The 57 kDa complex was later purified and identified as an RNA ligase [84] and the 50 kDa remains to be characterized. In a less efficient system, *L. tarentolae*, Aphasizhev's group affinity-purified proteins involved in RNA-binding from the mitochondrial extract to better understand their roles using biochemical assays [85].

To sum up, the establishment of in vitro systems allowed the determination of the roles that individual components play in the editing process. Further, different features of the gRNAs and mRNAs can be tested for their necessity in editing.

Our previous studies in *D. papillatum* demonstrated the presence of a stretch of six non-encoded Us at the junction of Modules 4 and 5 [61]. This suggests that a trypanosome type U-insertion editing machinery exist in *Diplonema* mitochondria. In fact the functions available in the hypothetical editosome would suffice the requirements of trans-splicing in *Diplonema* mitochondria, given a properly designed gRNA that directs the ligation of the transcripts of two adjacent modules.

2. WORKING HYPOTHESES FOR THE PROJECT

From the above, we can formulate two working hypotheses regarding trans-splicing in *Diplonema* mitochondria:

1. Small gRNAs exist and help assemble every two neighboring module transcripts. Such gRNAs should comprise sequences complementary to the junctions of two neighboring modules (Fig. 10A)

2. Alternatively, a complete antisense mRNA (referred in the following as “long gRNA”) exists and acts as a template to direct the simultaneous assemblage of all module transcripts of a given gene (Fig. 10B).

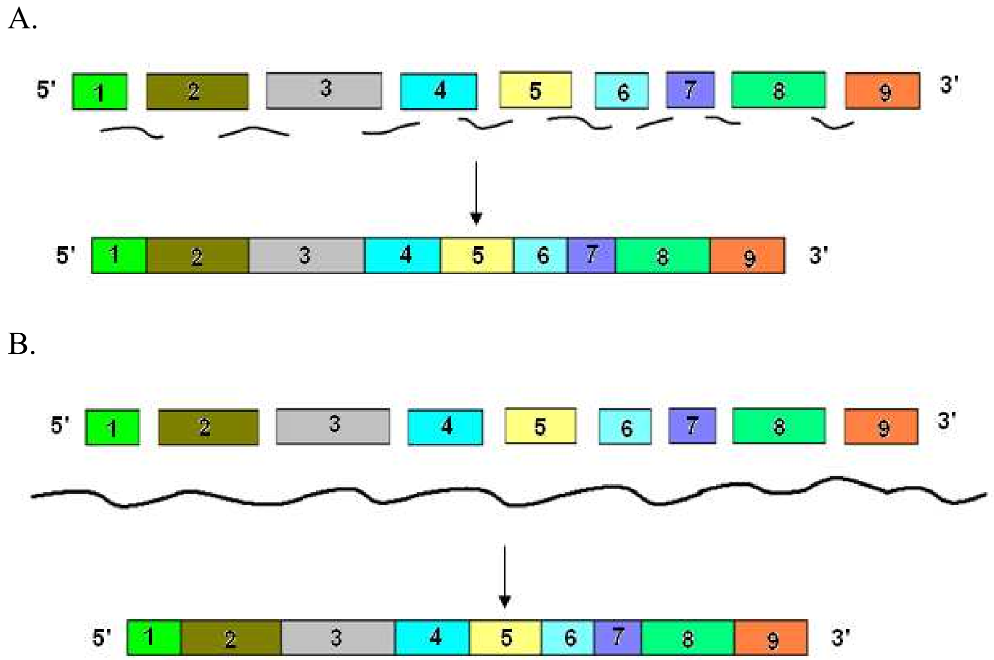


Figure 10. Working hypotheses on trans-splicing of *cox1*
(A) Multiple small gRNAs direct the joining of each junction. (B) A large RNA complementary to pre-mRNA acts as a template and directs the joining of multiple junctions.

3. MATERIALS AND METHODS

3.1 DNase I treatment of total RNA

RNase-free DNase I from Roche® was used to digest the residual DNA present in the total RNA extracted from *D. papillatum* according to the manufacturer's instructions. Briefly, 10 µL of 1 µg/µL of total RNA, 5 µL of 10X reaction buffer provided by the supplier, and 0.5 µL of 10 units/µL of DNase I were combined in a total reaction volume of 50 µL. The reaction mixture was incubated at 37 °C for 15 minutes. The enzyme was heat-deactivated at 75 °C for 10 minutes. The reaction mixture was phenol-chloroform extracted and ethanol-precipitated.

3.2 Tobacco Acid Phosphatase (TAP) treatment of RNA

TAP was purchased from Epicentre Biotechnologies® and treatment conditions followed instructions of the commercial supplier. Briefly, the reaction mixture contained 20 µg of total RNA from the previous step, 1X reaction buffer, and 2.5 units of the enzyme in a total volume of 20 µL. Incubation was carried out at 37 °C for one hour. The RNA was phenol-chloroform extracted and ethanol-precipitated, then resuspended in 20 µL of RNase-free water.

3.3 Polynucleotide Kinase treatment of RNA

T4 Polynucleotide kinase (PNK) was used following the instructions of the commercial supplier, New England Biolabs (NEB). Briefly, the reaction mixture contained 20 µg of RNA, 1X reaction buffer, 1 mM ATP, and 10 units of T4 PNK enzyme in a final volume of 50 µL. The reaction was incubated at 37 °C for one hour. The enzyme was inactivated at 65 °C for 20 minutes.

3.4 RNA circularization

Circularization of RNA was achieved by RNA ligation at low concentration using T4 RNA ligase from Roche®. Briefly, the ligation mixture contained 20 ng/µL of RNA, 1X reaction buffer, 0.1 mM ATP, 10 ng/µL of BSA, and 0.15 unit/µL of the ligase. The total reaction volume depended on the amount needed and we kept the total volume under 20 µL. The reaction mix was incubated at 16 °C overnight. RNA was phenol-chloroform extracted and ethanol precipitated. RNA was resuspended at 200 ng/µL in RNase-free water.

3.5 Reverse Transcription

We used avian myeloblastosis virus (AMV) reverse transcriptase from Roche® for this reaction. The reaction mix contained 50 ng/µL of RNA, 0.5 µM of the primer, 1X of the first strand reaction buffer, 100 µM DTT, 1mM of each dNTP, 1 unit/µL of the enzyme. The reaction was usually carried out in a final volume of 20 µL. The RNA and the primers were pre-incubated at 72 °C for 2 minutes, then other ingredients were added and the total reaction mix was incubated at 42 °C for 1 hour.

For convergent RT-PCR across the modules, the following primers were used in the RT:

M1/M2: dp142
M2/M3: dp88
M3/M4: dp146
M4/M5: dp129
M5/M6: dp150
M7/M8: dp84
M8/M9: dp154

For divergent RT-PCR, the following primers were used in the RT:

M1/M2: dp144
M2/M3: dp138
M3/M4: dp148
M5/M6: dp152
M7/M8: dp141
M8/M9: dp155

3.6 PCR

Expand High Fidelity PCR system from Roche® was used in the PCR following instructions from the commercial supplier. Briefly, the reaction mixture contained 2 µL of the cDNA synthesized from the reverse transcription reaction, 3 µL of each of the 10 mM primers, 1X reaction buffer, 0.2 mM of each dNTP, and 0.5 µL (3.5 U/µL) of the Taq enzyme in a final volume of 50 µL. The annealing temperature was calculated as 2X (A or T) + 4X (C or G) -5 °C. When primers of different melting temperatures were used in the same batch of reactions, a gradient PCR program was set up on the PTC-200 Peltier Thermal Cycler from MJ-Research PCR machine. The thermal cycles were: 96 °C for 2 minutes, 88 °C for 1 minute, 96 °C for 20 seconds, annealing temperature for 20 seconds, 72 °C for 1 minute, go to step 3 (96 °C for 20 seconds) 24 times, 72 °C for 5 minutes, and 4 °C forever.

For convergent PCR across the modules, the following primers were used:

M1/M2: dp142, dp143

M2/M3: dp88, dp80

M3/M4: dp146, dp147

M4/M5: dp129, dp109

M5/M6: dp150, dp151

M7/M8: dp84, dp85

M8/M9: dp154, dp41

For divergent PCR, the following primers were used:

M1/M2: dp144, dp145

M2/M3: dp138, dp139

M3/M4: dp148, dp149

M5/M6: dp152, dp153

M7/M8: dp141, dp140

M8/M9: dp155, dp156

3.7 DNA end-repair and purification

Two units of T7 DNA polymerase (NEB) and 2 units of Klenow enzyme (NEB) were added to the PCR mix as soon as the PCR cycles finish so that blunt ends were produced for cloning. The mix was incubated at 12 °C for 30 minutes. The products were visualized on a 0.8% agarose gel for their sizes. A low melting 0.8% agarose gel was used for gel purification of the PCR products of certain size ranges. DNA was either electro-eluted from the gel or extracted with QiaBeads™ gel-purification kit from Qiagen®. The reaction was stopped by heating at 65 °C for 2 minutes.

3.8 Cloning PCR fragments into vectors

T4 PNK form Roche® was used to phosphorylate the gel-purified PCR products. The reaction mix contained final concentration of 10 ng/µL of DNA, 1 mM of ATP, 1X reaction buffer, and 1 units/µL of the enzyme. The reaction mix was incubated at 37 °C for 30 minutes.

Using T4 DNA ligase, the phosphorylated product was ligated at 14 °C overnight into a blunt-end linearized and de-phosphorylated vector, pBFL6cat, that carries blue-white selection markers (B.F. Lang, unpublished). The reaction mix contains10 ng/µL of DNA (including the insert and the vector), 1 mM of ATP, 1X ligase buffer, and 1 unit/µL of T4 DNA ligase. The amount of insert DNA to be added was calculated to satisfy the molar ratio between the insert and the vector of 2:1.

Transformation of 1/10 of the ligation mix into *E. coli* cells (DH5α strain) was carried out according to protocols outlined in *Molecular Cloning* by Maniatis [86]. Bacteria were plated onto LB-agar plate (contained 5 µg/mL chlroramphenicol, 5 µg/mL tetracycline, 10 uM IPTG, and 40 µg/mL X-GAL) for blue-white selection. For the inserts whose sizes were known to be small, both blue and white colonies were picked because we would not miss clones that contained in-frame small inserts that failed to disrupt the *lacZ* gene function. Each picked colony was placed into a well in the 96-well culture block from SARSTEDT®. After overnight growth in LB (contained the same ingredients as the LB-agar plate except for the X-Gal and IPTG), plasmid DNA was extracted using the Qiagen 96-well mini-prep kit and was resuspended in 40 µL volume in 96-well Corning® plates at -20 °C.

3.9 Sequencing reactions

BigDye® Terminator v3.1 Cycle Sequencing Kit from Applied Biosystems was used for the sequencing reaction. From the DNA plate, 1 µL of each plasmid (approximately 20 ng/µL) was taken and added into a 96-well PCR reaction plate. Each well contained 0.7 µL of the BigDye, 0.6 µL of the 5X reaction buffer, 0.4 µL of the M13+ primer that is upstream of the cloning site and 7.3 µL of water. Sequencing reaction cycles were set up on the PTC-200 Thermal Cycler according to the user's manual. After the reaction, DNA was precipitated and washed in 70% ethanol. Then the pellets were dried in the dark, and the plate was wrapped in aluminum foil and sent to the sequencing platform at l'Institut de recherche en immunologie et en cancérologie (IRIC).

3.10 Sequence analysis

Sequencing results were downloaded from the website of the sequencing facility. After unzipping the files, we used in-house Perl scripts to retrieve sequences and remove the vector sequences from them. The resulting sequences were combined in a multi-sequence FASTA file. Using local BLAST we queried the primer sequences against the sequencing results, and we identified the positions of the primers in each sequence. Using the MotSearch program, we looked for the hypothesized gRNA sequences that matched the profiles in Appendix 1. *D. papillatum* mitochondrial DNA sequences of around 250 kbp and mitochondrial cDNA sequences of about 20 kbp were also queried against the sequencing results. Results were organized in Excel files where the configuration of the sequences were categorized and counted.

3.11 *D. papillatum* cell culture and preparation of sub-cellular fractions

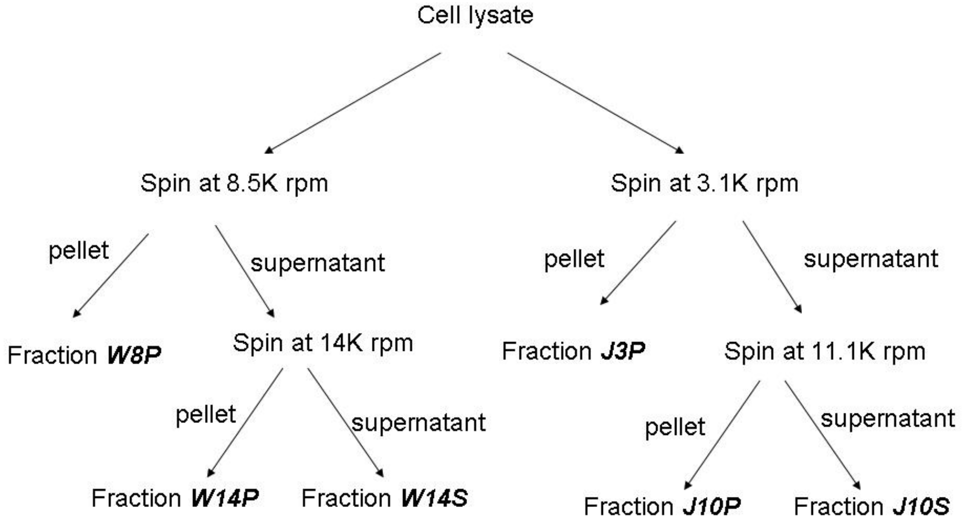


Figure 11. A flow chart of the fractionation process of the cell lysate
The left branch of fractionation followed the centrifugation speeds used by William Marande who previously worked on the project in the lab; hence, the fractions were designated with the prefix "W". The fractionation scheme on the right was adapted from a whole-cell-extract preparation protocol on mammalian cells used in Dr. Jerry Pelletier's lab, hence the prefix "J". The suffixes "P" and "S" represent the pellet and the supernatant, respectively.

Two liters of *D. papillatum* culture was grown to a density of 1.6X10^6 cells/mL after approximately a week in 0.33% Instant Ocean, 0.1% bacto-tryptone, and 1% horse serum. Cells were harvested at 6,000 rpm with a GSA rotor. Cell pellets were resuspended in 1.5 volume of STE buffer (250 mM sucrose, 20 mM Tris of PH 7.9, and 2 mM EDTA) and let it sit for 10 minutes for proper mixing. The cell resuspension was then passed though a 22 gauge needle 10 times, and 60% of sucrose was added immediately to the lysate in 6:50 volume ratio (e.g. 2.4 mL of 60% sucrose will be added to 20 mL of lysate). The lysate was then divided in two (Fig. 11): one was for the preparation of a fraction called W8P following a protocol previously used in the lab [1], and the other one was for the preparation of the fraction called J10P using a modified centrifugation scheme. To make W8P , the cell lysate was divided into 1 mL aliquots in Eppendorf tubes and centrifuged at 8,500 rpm (7,650 *g*) in an Eppendorf Microfuge 5417C at 4 °C for 10 minutes to clear the lysate. The pellet portion was labeled as W8P, which would typically contain the *Diplonema* nucleus, and the supernatant portion was labeled as W8S. The W8S was futher centrifuged at 14,000 rpm (20,800 *g*) to precipitate the mitochondria in the Microfuge 5417C for 15 minutes. The pellet was labeled as W14P, which contains the mitochondria, and some of the microsomes and membranes, while the supernatant was labeled as W14S, which contains the rest of the membranes, ribosomal RNAs and proteins. For the other lysate, aliquots of 1 mL in eppendorf tubes were centrifuged in the microfuge at 4 °C for 10 minutes at 3,100 rpm (1,000 *g*) for 10 minutes. The first centrifugation is at a lower speed than the first fractionation scheme because the diplonemid mitochondria are known to be large. A slower speed will avoid precipitating mitochondria with nuclei during the first centrifugation. The supernatant was labeled J3S while the pellet was labled J3P. Further centrifugation was performed on J3S in the microfuge at 11,100 rpm (13,000 *g*) at 4 °C for 20 minutes, and the supernatant was labeled J10S, which contains membranes, proteins and ribosomal RNA, while the pellet portion was labeled as J10P, which should contain mitochondria.

The mitochondria-containing fractions W8P and J10P were further lysed with a final concentration of 0.3% Triton X-100. The Triton-lysed fraction was cleared by centrifugation at 10,600 rpm (12,000 *g*), then the supernatant portion was used for the in vitro assay. The cell extract preparation conditions were adapted from the mitochondrial DNA purification protocol from previous work in the lab [1].

3.12 In vitro transcription and radio-labeling of substrates

Module 3, Module 4, Module 4 with six Ts, and Module 5 were amplified by PCR using primers that contained the T7 RNA polymerase promoter sequence at the 5' end. Primers used were dp157-163 (Appendix B). The PCR products were run on an agarose gel and the corresponding bands were gel-purified and confirmed by sequencing. In vitro transcription was carried out following the instructions in the T7 RNA polymerase kit from NEB. The reaction contained 1X RNA polymerase buffer, 5 µL of 10 mg/mL BSA, 1.25 µL murine RNaseIn® (NEB), 2.5 µL of 10 mM ATP, CTP, and GTP, 1 µL of 10 mM UTP, 3 µL of α-^{32}P-UTP (10 µCi/µL), 10 µL of 20 ng/µL DNA templates, and 1 µL T7 RNA polymerase in a total volume of 50 µL. The reaction was incubated at 37 °C for 2 hours, phenol-chloroform extracted and passed through a G-50 column. RNA was then washed and precipitated in 70% ethanol, resuspended in 10 µL of RNase-free water, and kept at -80 °C for future use.

3.13 In vitro trans-splicing assay

An in vitro trans-splicing experiment was performed by incubating a *D. papillatum* crude mitochondria-enriched extract, unlabeled total RNA, together with internally labeled *cox1* RNA Module 3, 4 or 5. Two different crude extracts were used: W8P and J10P. Each in vitro reaction contained 25 mM of HEPES (PH 7.5), 5 mM magnesium acetate, 50 mM potassium acetate, and 1 mM DTT, 1 mM ATP, 15 µL of either one of the extracts, 0.6 µL of the NEB murine RNase inhibitor, 3 µL of the radio-labeled RNA substrate, and 1 µL of the total RNA (200 ng/µL) in a total volume of 30 µL. The reaction was incubated at 37 °C for 4 hours. Then RNA was phenol-chloroform extracted, and run on a 6% PAGE in 1X TBE, with 7M urea, at 5 V/cm for one hour then at 10 V/cm for five hours and 20 minutes. The gel was exposed to a film at -80 °C for 48 hours.

3.14 Capping reaction

Capping enzyme was ordered from the Epicenter® Biotechnologies. Briefly, each capping reaction contained 1X reaction buffer, 0.1 mM S-adenosyl-methionine (SAM), 1 unit/µL of RNAse inhibitor, 1 µCi/µL of α-^{32}P-GTP from Perkin Elmer, 1 µg/µL of total RNA, and 0.4 units/µL of the ScriptCap™ enzyme. The capped RNA was phenol-chloroform extracted, ethanol-precipitated, and resuspended in 20 µL of RNase-free water.

3.15 Gel-purification of radio-labeled probes and Southern blot

A 12% polyacrylamide gel with 6M urea was used to separate the total labeled RNA. Labeled bands were cut out and soaked in the elution buffer (0.1% SDS, 0.5 M ammonium acetate, and 10 mM magnesium acetate) at 4 °C overnight. The elution was centrifuged briefly to rid of the residual polyacrylamide.

Meanwhile, total DNA was run on a 0.7% agarose gel in multiple lanes. A picture of the gel was taken under UV with a ruler. The gel was denatured in denaturing solution (1.5 M NaCl and 0.5M NaOH) for 15 minutes at 4 °C and then rinsed in neutralization solution (1.5 M NaCl, 0.5 M Tris of pH 7.5, and 1 mM EDTA) twice for 15 minutes at 4 °C. The gel was allowed to equilibrate with 10X SSPE (1.8 M NaCl, 0.1 M phosphate buffer of pH 7.7, and 10 mM EDTA) at 4 °C for 10 minutes. One hundred milliliter of phosphate buffer of pH 7.7 was prepared by combining 10.5 ml of 1 M NaH_2PO_4 and 89.5 ml of 1 M Na_2HPO_4. DNA in the gel was then transferred to a Hy-bond™ membrane overnight using 2 layers of Whatman™ filter paper and 3 cm of paper towels stacked underneath the membrane. The DNA sample lanes on the membrane were then cut so that each lane can be hybridized to a different probe. The membranes were pre-hybridized with degraded and denatured salmon sperm DNA in the hybridization buffer (1% SDS, 1.5X SSPE, 0.5% Denhardt's reagent, and 0.1 mg/ml degraded salmon sperm DNA).

Denatured probes were then added to a fresh hybridization buffer and hybridization was carried out at 65 °C overnight. The membranes were washed a few times using buffers with descending ionic strength. Humid membrane was then exposed to film for 48 hours.

4. EXPERIMENTAL RESULTS

4.1 Radio-labeling of mitochondrial RNA with capping enzyme

Guide RNAs of kinetoplastids involved in U-insertion/deletion editing are small RNA molecules encoded by mitochondrial DNA. We asked whether such a population of small RNAs exists in *D. papillatum* mitochondria as well. Employing an RNA capping method that was previously used in the study of trypanosomes [87], we intended to detect the existence of these small RNAs and locate their genes in the genome. The rationale of the approach is briefly described below.

In all eukaryotes, the 5' triphosphate end of primary transcripts in the nucleus is modified by the RNA 5'-triphosphatase (RTP), RNA guanylyl transferase (RGT), and RNA (guanine-7-) methyltransferase (RNMT) [88, 89]. The cap structure is an inverted and 7-methylated guanine nucleotide that is connected via a 5'-5' triphosphate bridge to the first guanine residue of the mRNA. It is often written as GpppRNA cap, or M7G cap. Other methylation positions are known while the inverted guanine is invariant among all cap structures. Capping of primary transcript, however, does not take place in the mitochondria. Using a commercially provided capping enzyme that was engineered to combine RTP and RNMT activity, we intended to specifically radio-label the 5' ends of all mitochondrial primary transcripts. Labeled RNAs were separated by electrophoresis and used as probes in Southern blot to identify their origin in the genome.

We have identified three major bands that were radio-labeled after running a polyacrylamide denaturing gel (Fig. 12). The very top band corresponds to the RNA molecules that did not enter the gel. We estimated sizes range of the labeled bands is between 50 to 300 nucleotides. Comparing to the EtBr stained unlabeled total RNA, we see that the RNA band at about 190 nt position could not be capped, while the top and bottom radio-labeled bands is not visible in EtBr staining.

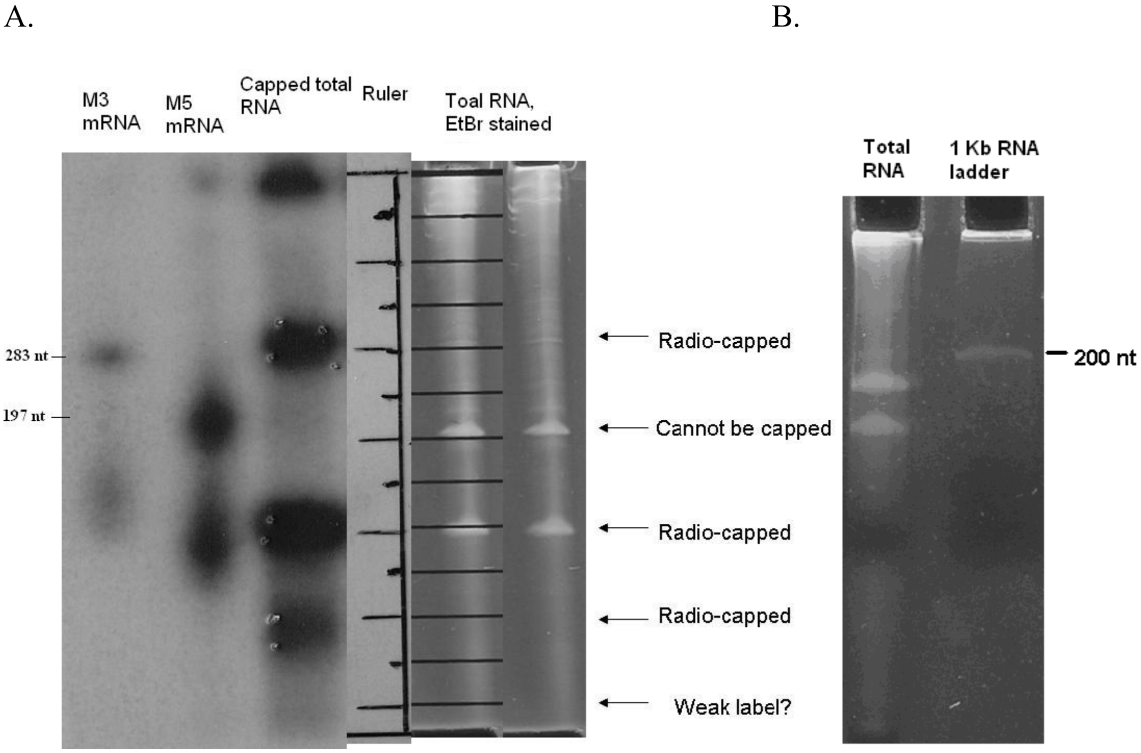


Figure 12. Capping of total RNA of *D. papillatum* in the presence of radio-labeled GTP
(A) Lane 1 and 2 contain ^{32}P-labeled M3 and M5 by in vitro transcription in the presence of α-^{32}P-GTP. Upper band corresponds to the in vitro transcription product and the lower smear probably corresponds to the partially degraded RNA. Lane 3 is the total RNA with ^{32}P incorporated in the cap. Unlabeled total RNA (right-most two lanes) was run on the same gel and the lanes were cut to be visualized by ethidium bromide staining and UV light. A picture was taken with a ruler on the side, so that we can compare the bands to the radio-labeled ones. (B) Total RNA and 1Kb RNA ladder were electrophoresed side by side and visualized by ethidium bromide staining. The lowest band of the ladder corresponds to 200 nt.

Probably due to the low activity of the probes, Southern hybridization did not detect on which genome those RNA species are encoded.

4.2 Trans-splicing intermediates containing the Module 4 and 5 junction

To confirm that trans-splicing proceeds accurately at the junction of Module 4 and 5, and to better understand how the six Us appear between these two modules, we performed RT-PCR on circularized RNA molecules to capture RNA intermediates from the trans-splicing process. To circularize RNA molecules, we treated RNA with TAP enzyme and T4 PNK with 3'-phosphatase activity, followed by RNA auto-ligation at low concentration (Fig. 13).

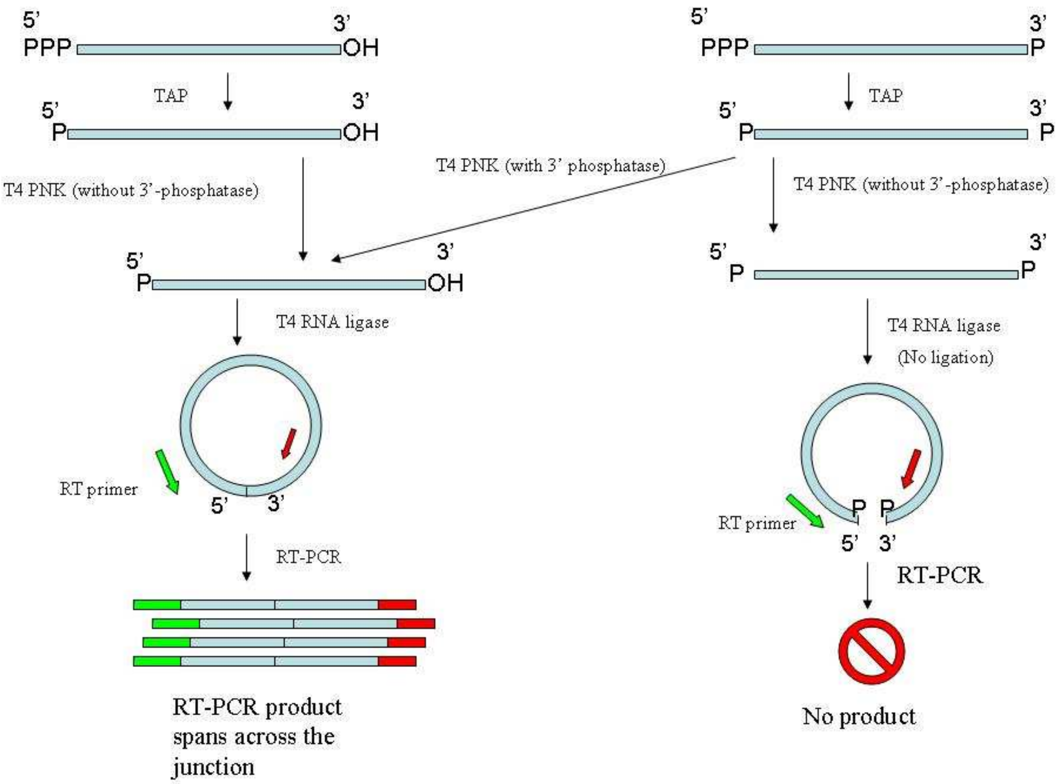


Figure 13. Difference between different PNK enzymes for subsequent circularization of RNA Mitochondrial transcripts in *D. papillatum* may have at their 5' end a single phosphate (if processed) or a tri-phosphate (in primary transcript). At their 3' end, they could have an -OH group or a phosphate group (in the newly added U-tail in trypanosome U-insertional editing). Some transcripts are phosphorylated at 3' ends. When an RNA molecule with 3'-OH was treated with TAP and T4 PNK that lacks the 3' phosphatase activity, the 5' tri-phosphate was converted to a mono-phosphate and it was readily ligated by T4 RNA ligase (pathway on the left). For an RNA molecule that was phosphorylated at the 3' end, the wild-type PNK (with 3'-phosphatase activity) was used for successful auto-ligation; otherwise, the RT-PCR could not proceed (pathway on the right).

The synthesis of cDNA was primed in Module 5 in the reverse direction from its 3' end, and then three PCR reactions were carried out with different primer pairs. All primer pairs consisted of a primer that was located near the 3' end of Module 5 in the forward direction, and its counterpart was located near the 5’ end of Module 3, 4 or 5 respectively, in the reverse direction. All PCR products were pooled together before cloning into vectors. This was my first batch of sequencing (i62). The second batch (i65) of sequences was prepared from a different PCR: using the same cDNA, the primer in Module 3 was omitted in the PCR. We only intended to detect intermediates that contained Module 4 and Module 5 sequences so that we could better understand how trans-splicing took place at Modules 4 and 5 junction accompanying the insertion of six Us. Sequencing results of the first RT-PCR from 48 clones is summarized in panel A of Figure 14. Panel B contained sequencing results for the second batch of 96 clones of RT-PCR products targeting intermediates containing the M4/M5 junction. This batch was labeled “i65”. In both cases, no Module 5 with six Us attached to its 5’ end was detected and no Module 4 was found to be directly connected to Module 5. On the other hand, Module 4 was found to have six Us attached to its 3’ end with or without the adjacent modules. Module 4 with 5’ and/or 3’ non-coding regions were detected to be abundant in both cases, while “cleanly” processed Module 4 occurred only rarely.

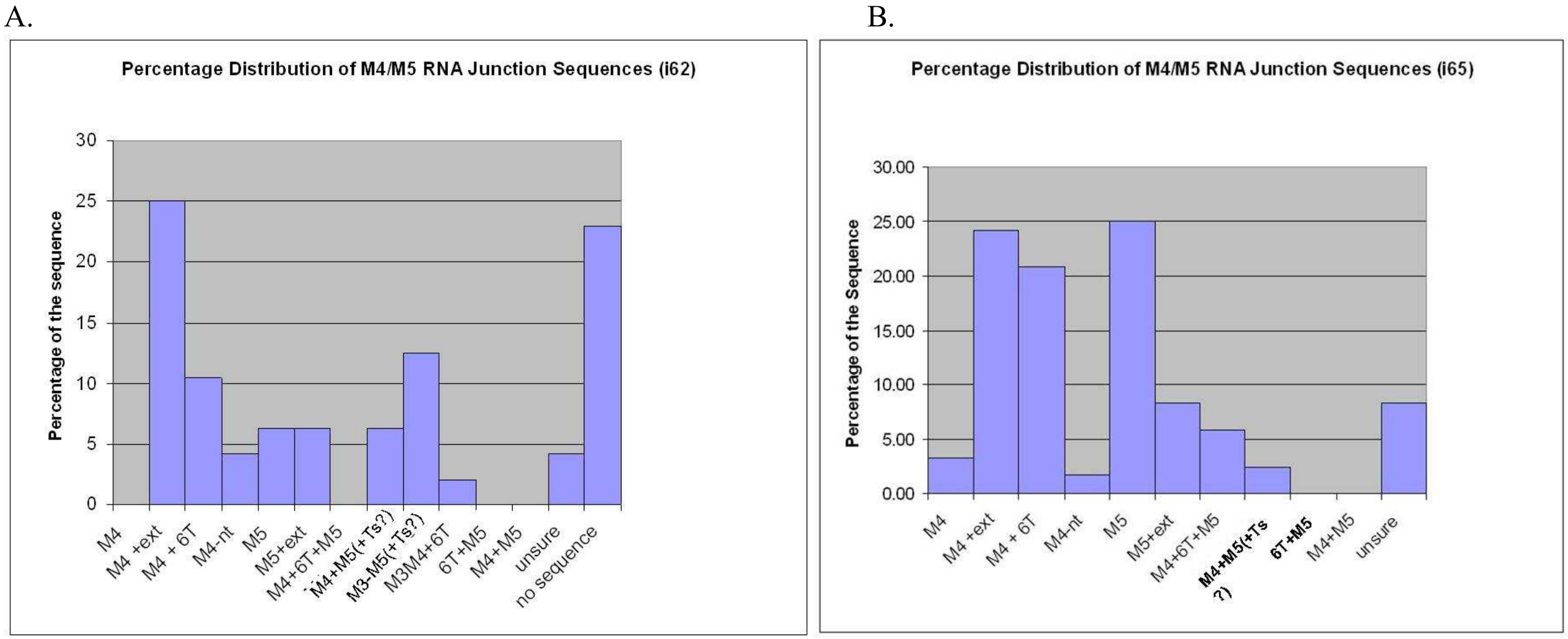


Figure 14. Sequencing RNA species that contain M4/5 junctions
M4, Module 4 coding sequence; M4 + ext, Module 4 with 5' and/or 3' extension; M4 + 6T, Module 4 with six T(U for RNA)s attached to its 3' end; M4-nt, Module 4 with some nucleotides missing; M5, Module 5 coding sequence only; M5 + ext, Module 5 with 5' and/or 3' extensions from the primary transcript; M4 + 6T + M5, Module 4 and Module 5 joined together with six T in the between; M4 + M5 (+Ts?), Module 4 and Module 5 are joined, but whether there are six Us in between cannot be detected by RT-PCR because the primers chosen do not cover this region; M3-M5(+Ts?), Module 3 is joined with Module 5, but whether there is a Module 4 with six Ts in between cannot be detected in this RT-PCR due to the primers chosen; M3M4 + 6T, Module 3 and Module 4 are joined and there are six Ts attached at the 3' end of the Module 4; 6T + M5, there are six Ts attached to the 5' end of Module 5; M4 + M5, Module 4 and Module 5 are directly linked without anything between them; unsure, the sequences detected are not legible; no sequence, sequence is either empty or the read quality is too low to be analyzed. For each type of sequence detected, its frequency among the same batch of sequences is counted, as indicated by the y-axis.

4.3 Long anti-sense RNA does not exist

If a long complementary RNA exists to direct the module assembly of the *cox1* gene, then we should be able to identify RT-PCR products that span across multiple modules of the gene. In order to detect such long anti-sense molecule present at very low concentration, we performed a RT-PCR followed by a nested PCR reaction to enrich the product. Since shorter PCR products are favored in the reaction, we only aimed to amplify part of the hypothetical anti-sense RNA that spans across Module 5 and 9 (Fig. 15). We used for the positive control the same set of primers except for the RT primer which targets the sense RNA.

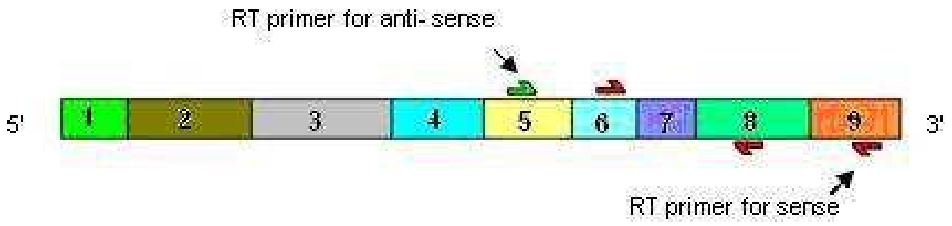


Figure 15. Experimental design of RT-PCR to detect long anti-sense RNA
A primer (green) in the anti-sense direction was used for reverse transcription. Following the reverse transcription, two rounds of PCRs were performed using nested primers shown here. The first PCR used the outer pair (in Modules 5 and 9) and the second round used inner pair (in Modules 6 and 8). As a positive control, a primer in the sense direction (in Module 9, red) was used for cDNA synthesis. Then the same two rounds of PCR were performed and we expected to amplify a region spans about 400 nt (between Module 6 and 8) on the mature mRNA. From left to right, the primers are: dp82, dp67, dp35, and dp38.

We focused on products that span Modules 5 through 9. As shown is Figure 16, the nested RT-PCR did not yield a product indicative for an anti-sense RNA molecule that spans across more than four modules (lane 1). The positive control (lane 2) showed that under the same experimental conditions, the mRNA was successfully detected for its region between Module 5 and 9.

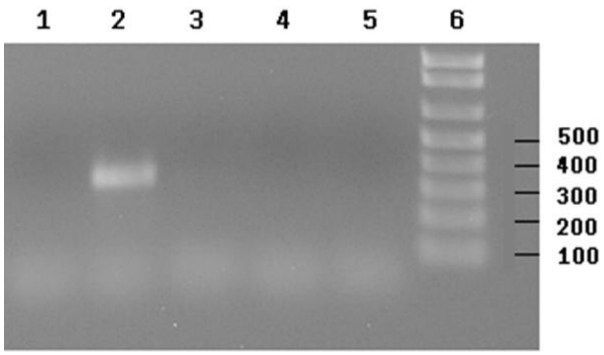


Figure 16. RT-PCR results of anti-sense RNA detection.
Lane 1: Detection of anti-sense of Modules 5 through 9. For reverse transcription of the antisense, a primer annealing in Module 5 was used. Primer pairs used in nested RT-PCRs flank Modules 5 and 9, then 6 and 8. Lane 2 (positive control): Reverse transcription of the sense mRNA between Modules 5 and 9, followed by nested PCR using the same primers as in Lane 1. Lanes 3-5: Negative controls of RT and both PCR reactions.

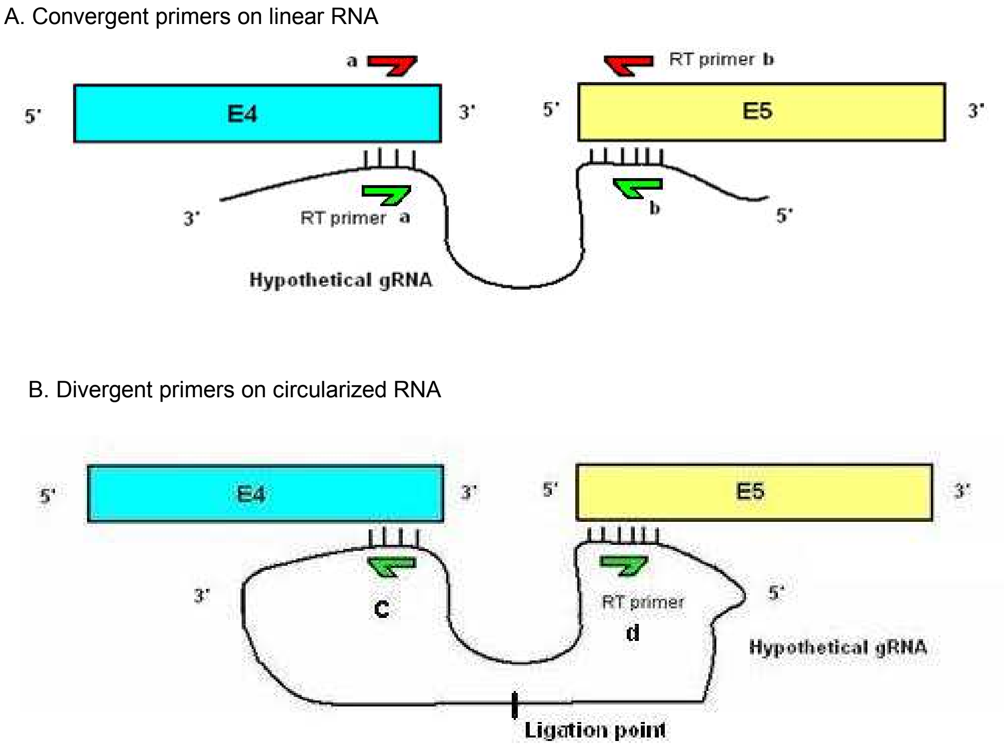
Hence, no long anti-sense RNA likely exists to direct the post-transcriptional assembly of the whole *cox1* transcript.

4.4 RT-PCR and Amplicon Sequencing identified sequences that match gRNA profiles

If small gRNAs exist for module joining, we should be able to amplify the guiding nucleotides in the gRNA using primers complementary to the nucleotide sequences at the module borders. Three sets of RT-PCR were designed.

In the first set of experiments, we performed a reverse transcription (RT) reaction on the gRNA using a primer that was identical in sequence to the nucleotides at the 3' end of Module 4 (green primer *a* in panel A of Fig. 17). In the subsequent PCR, we used the RT primer and a primer (green primer *b* in panel A of Fig. 17) that was identical to the reverse complementary sequence of the nucleotides at the 5' end of Module 5. In the expected product we should detect the sequence of primer *a*, the guiding nucleotides of the gRNA, and the sequence of primer *b*. To be sure that the products do not come from ligation of the two primers, the primers were designed not to cover the last nucleotide of Module 4 and the first nucleotide of Module 5. If the RT-PCR product came from real amplification of RNA, we would detect the two uncovered nucleotides in the sequence of the product; otherwise, if the product came from the ligation of primers, those two nucleotides would be absent. Similar primers were designed for the junctions of M1/M2, M2/M3, M3/M4, M5/M6, M7/M8, M8/M9. The same primers were used in the amplification of the mRNA as a positive control (primer pair in red, panel A of Fig. 17).

In the second set of RT-PCR experiments, instead of linear RNA, we used circularized RNAs as templates and diverging primers in the PCR to obtain sequences beyond the binding sites of the primers (panel B in Fig. 17). For each module junction, the primers were exactly complementary to the sequences of those used in the RT-PCR on linear RNA. For example, we used an RT primer (primer *d*, in green) that was identical in sequence to the sense nucleotides at the 5' end of Module 5 to amplify the sequences in M4/M5 gRNA. The subsequent PCR used primer *c*, reverse complementary in sequence to the 3' end nucleotides of Module 4, and primer *d* for amplification. Primer c and *d* were the reverse complementary of the convergent RT-PCR primers *a* and *b*, respectively. The advantage of this method is that small RNAs can be reverse-transcribed into larger DNA molecules so that we can detect, purify, clone and sequence them more easily. Similar RT-PCRs were carried out upon junctions of M1/M2, M2/M3, M3/M4, M5/M6, and M7/M8.


A. Convergent primers on linear RNA

B. Divergent primers on circularized RNA

Figure 17. Experimental design of RT-PCR to detect partial sequences of gRNAs
(A) The green primer *a* on the left-hand-side of the junction was used for reverse transcription of potential gRNA molecules that were anti-sense to the mRNA. A convergent primer pair (*a* and *b*, in green) were used in the RT-PCR (primer *a* being the RT primer) to amplify the guiding nucleotides in the gRNA. For the positive control, primer *b* (red) was used in the RT targeting the mRNA, and the same primer pair (*a* and *b* in red) were used in the PCR. (B) When the total RNA was circularized, reverse transcription was primed by primer *d* targeting anti-sense sequences. PCR was carried out using the primer pair in divergent directions (primers *c* and *d*) to amplify the non-guiding sequence of the gRNA.

If our RT primers (primer *a*) can bind efficiently to the gRNA and initiate cDNA synthesis, then we might be able to amplify longer 5' sequences of the gRNA by changing the specific reverse primer (primer *b*) to a non-specific primer in the PCR. We expected to detect the guiding nucleotides and the binding site of primer *b* in the products, so that we could be more confident that the RT-PCR using two specific primers indeed targeted the gRNA; moreover, we intended to sequence further toward the 5' end of the gRNA. Hence, in the third set of RT-PCR experiments, we used the SMART primer from Clontech® as a non-specific primer. The SMART primer, together with the specific primer, could amplify all cDNAs that have the SMART IIA oligonucleotides tagged at the 3' ends (template switching, as depicted in Fig. 18). As a consequence, all newly synthesized cDNAs would contain the same short sequence at their 3' ends. We chose a Module 4 and 5 junction primer (dp129) for this experiment because the presence of the guiding nucleotides of six inserted Us are easier to identify among the sequences.

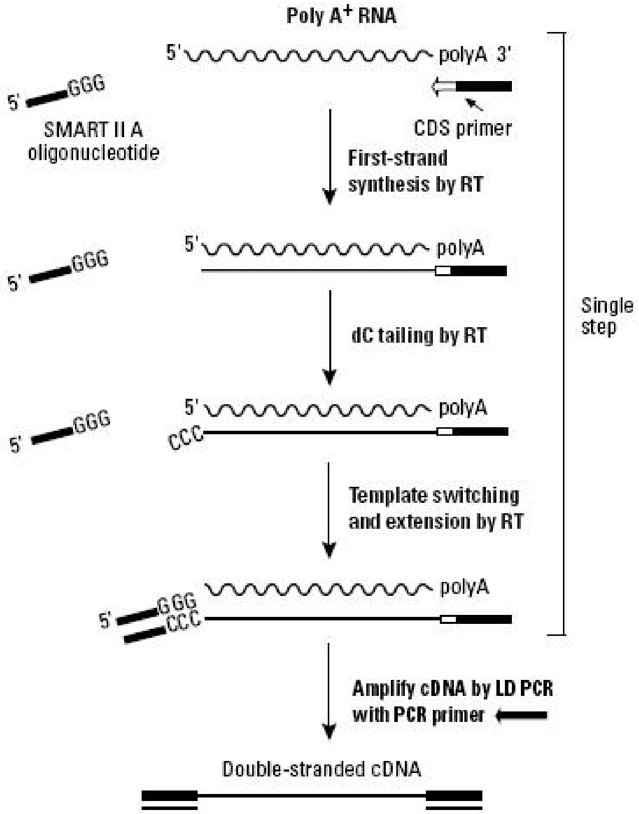


Figure 18. Reverse transcription of RNA using normal RT primer (CDS primer) and Clontech® SMART primer (SMART II A oligonucleotide)
The first strand of all cDNA molecules were tagged at their 3' ends with sequences complementary to the SMART primers by template switching directed the SMART IIA oligonucleotide. Amplification was then carried out by PCR using the SMART primer and the RT primer (Figure is taken from SMART™ PCR cDNA Synthesis Kit Manual).

RT-PCR products were separated by agarose gel, then cloned into a vector and sequenced. For the first set of RT-PCR, possibly because the RT-PCR products from convergent primers were very short (about 50-70 bp), we could not tell very well if they were integrated into the plasmid when we tried to visualize the purified plasmids on an agarose gel. As a result, in many cases we sequenced empty vectors ('Bad reads' in Table 1). Nevertheless, we identified sequences that were perfectly matching the gRNA profiles in five out of eight module junctions (anti-sense to the junction, in Table 1), using

MotSearch program. For example, for the M4/ M5 junction, we have identified 13 times the exact sequences of the module junction: primer *a* + CTTTTTTC + primer *b* (refer to Figure 17A for primer design) or primer *d* + GAAAAAAG + primer *c*. This indicates that an RNA molecule might exist and it could base-pair with the joining ends of M4 and M5 RNA; in addition, it contained a stretch of six As (complementary to six Us) between the pairing regions with M4 and M5. The clones that contained sequences perfectly matching the gRNA profiles are listed in Appendix 3. We have also observed artifacts such as ligated primers in our sequences. Other possible sequences were detected between the primers; however, they do not conform to the profiles of hypothetical gRNAs. Similar analysis was performed for all other module junctions. Besides the perfectly matching anti-sense sequences, we did not detect any potential gRNAs that would mediate trans-splicing.

Table 1. Sequencing results of RT-PCR set 1: hypothetical gRNAs targeted by specific primers converging towards the module junctions

Module junctions (Good reads/Total reads)	Sequence Identity	Frequency
M1/M2 (28/36)	**Anti-sense to the junction** [a]	0
	small rRNA (*rns*) [b]	24
	Unknown	4
	Bad reads [c]	8
M2/M3 (29/36)	**Anti-sense to the junction**	6
	rns	3
	rns + unknown	2
	M2	2
	M1 + M2	3
	Unknown	13
	Bad reads	7
M4/M5 (33/96)	**Anti-sense to the junction**	13
	Unknowns	9
	Ligated primers	5
	Very short unknowns	5
	Bad reads	64
M5/M6 (7/24)	**Anti-sense to the junction**	5
	M3 + 5' extension	1
	Unknown	1
	Bad reads	17
M7/M8 (23/36)	**Anti-sense to the junction**	0
	Unknowns	21
	One primer + M7 mRNA	1
	One primer + unknown	1
	Bad reads	13
M8/M9 (57/60)	**Anti-sense to the junction**	1
	Unknown	2
	Unknown + constant chromosome sequences	17
	known gene modules or chromosome sequences	38
	Bad reads	2

[a] No significant match with junction nucleotide sequences were found due to the mistakes in primer dp143.

[b] Gene encoding the cytoplasmic small subunit ribosomal RNA

[c] Bad reads are readings that contain either no sequence, or the quality of the sequence is too low to be read.

In the second set of RT-PCR experiments performed on circularized RNA, we did not identify any gRNA using MotSearch. Annotation of the sequences by searching the cDNA and the genomic database revealed that every RT-PCR reaction has non-specifically amplified some mRNA (Table 2) as well as unknown sequences. The unknowns were classified by size and they may correspond to non-guiding nucleotide sequences in gRNAs. A list of the unknown-sequence-containing clones was compiled for further analysis of gRNA candidates (Appendix 4). The PCR reaction for the M7/M8 junction requires to be repeated due to a mix-up of the primers: dp84 was mistakenly used instead of dp140. This explains why partial downstream module sequences were detected while other PCRs never detected downstream modules that do not contain any one of the primers.

Table 2. Sequencing result of RT-PCR set 2: circularized gRNA targeted by specific primers diverging from the module junctions

Module junctions (Good reads/Total reads)	Sequence Identity	Frequency
M1/M2 (27/48)*	**Partial M1**	20
	Unknown	6
	Primer ligation	1
	Bad reads	11
M2/M3 (35/48)	**M2 with 5' extension**	21
	Unknown	11
	Primer ligation	2
	Bad reads	13
M3/M4 (23/24)	**M3 with 5’ extension**	11
	Unknown	12
	Bad reads	11
M5/M6 (15/24)	**M5 with 5' extension**	8
	Unknown	7
	Bad reads	9
M7/M8 (33/48)	**M8 + partial M9**	6
	Unknowns	27
	Bad reads	15

* Primer dp144 contains a C at position 6 as an error. This could explain why no partial sequences from module 2 were detected.

In the third set of experiments, we used a specific primer (dp129) in combination with the non-specific primer, SMART primer, in the cross-module RT-PCR (refer to Fig. 17A, dp129 as primer *a* and SMART primer as primer *b*). We performed the same analysis on the sequences as in other RT-PCR experiments. We did not detect any M4/M5 guiding nucleotide sequences according to the profile, and we did not detect the reverse primer (dp109) among the sequences either (Table 3). None of the sequences was identified in the cDNA library or the genomic sequences of *D. papillatum* or in the non-redundant nucleotide online database of the NCBI.

Table 3. Sequencing results of RT-PCR set 3: RT-PCR across M4/M5 using dp129 in combination with SMART primer

Sequence identity	Frequency
dp129 + unknown + SMART	26
dp129 + unknown	5
Short, unmatched sequences	6
Bad reads	17
Good reads / total reads	31/48

4.5 Trans-splicing was not observed in the preliminary in vitro assay

We needed a system to validate the functions of hypothetical gRNAs identified. Therefore, an in vitro system was tested for *Diplonema* to observe trans-splicing. The basic approach was to prepare a mitochondrial extract and incubate radio-labeled RNA modules with it. If trans-splicing took place, then we could detect an increase in size of the radio-labeled RNA molecule. Due to the morphology of mitochondria and their locations inside the cell, pure mitochondria were hard to isolate from *Diplonema*, and another member of our group attempted to establish a protocol. As a starting point, the in vitro system was based on a whole cell extract. The substrates were two adjacent *cox1* modules and they were generated by in vitro transcription in the presence of ^{32}P-labeled nucleotides so that they were internally radio-labeled. The products of the in vitro assay were resolved on a denaturing polyacrylamide gel.

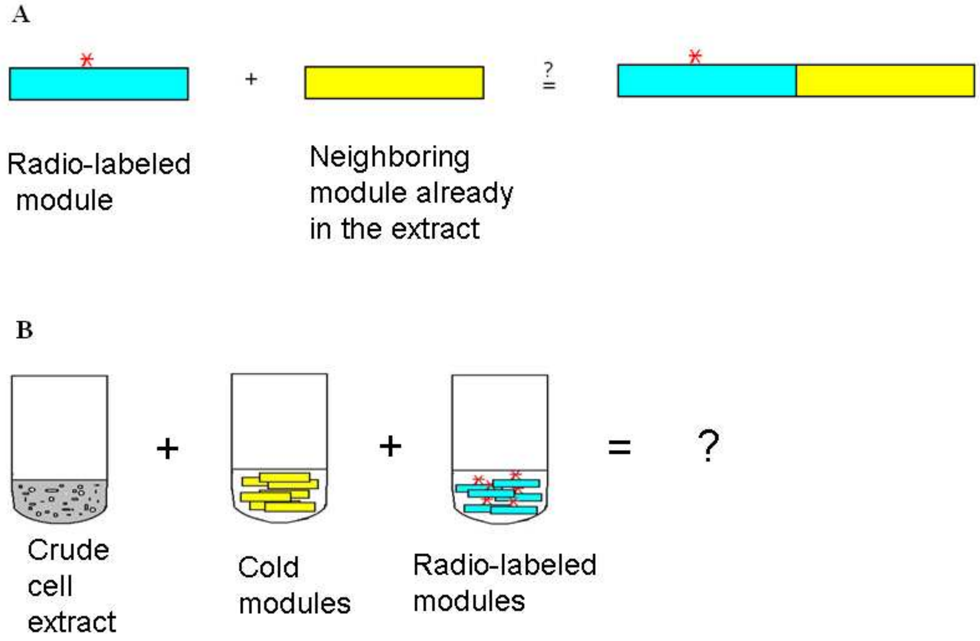


Figure 19. An illustration of the principle of in vitro assay for *cox1* trans-splicing
(A) ^{32}P-labeled modules were incubated with unlabeled RNA and we intended to detect a trans-spliced product. (B) Crude cell extracts prepared from *D. papillatum* were used in the incubation. In order to detect the formation of the products, we ran the reaction products on a polyacrylamide gel to see whether there was an increase in size of the radio-labeled RNA molecules.

Two types of crude cell extracts prepared were incubated with P-32 labeled cox1 RNA modules and unlabeled total RNA. Under the current experimental conditions, we did not observe trans-splicing.

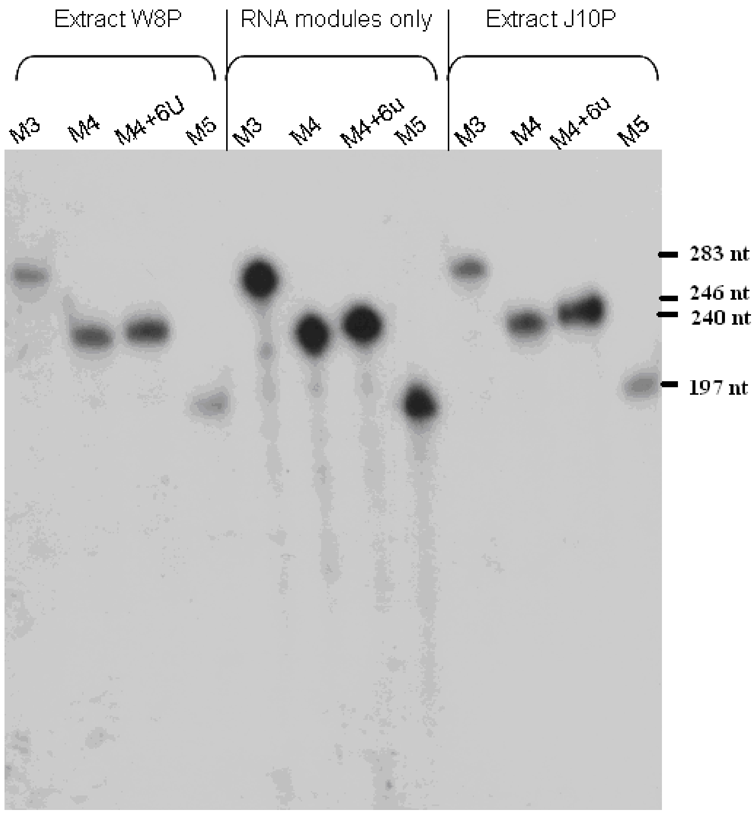


Figure 20. Trans-splicing in vitro using radio-labeled modules
M3: Module 3 coding RNA; M4: Module 4 coding RNA; M4+6U: Module4 coding RNA with six Us at 3' end; M5: Module 5 coding RNA. The first four lanes were incubated with *W8P*, the pellet fraction of a lower speed centrifugation that clears the extract; the last four lanes used *J10P*, which is a pellet fraction of the extract that underwent higher speed of centrifugation for clearing the extract. No trans-splicing was observed under both conditions. The middle four lanes were loaded with the labeled RNA only. No degradation occurs after 2.5 hours of incubation at 37 °C and 4 hours of electrophoresis on the 7M urea PAGE.

The pilot experiment indicates that the conditions used were suitable for the stability of RNA molecules. The following factors might contribute to the absence of trans-splicing in vitro. First, the substrates synthesized in vitro contain T7 promoter sequences at the 5' ends of the Module 3, 4 and 5. Module 4 and Module 5 sequences were identified among the natural RNA population from the mitochondria, therefore radio-labeling them for tracing the reaction is a reasonable design. However, the T7 promoter sequence does not naturally occur at the 5' ends of these modules. The altered 5' end sequences may not be recognized by the trans-splicing machinery, therefore no trans-splicing would happen at their 5' end.

Secondly, the preparation of the crude extract may have omitted certain components of the cell that were essential for the process to take place. In both preparation scheme, the mitochondria-containing fractions were lysed and used for in vitro assay. In the preparation of the *W8P* extract, the cell lysate was spun at 7650 *g* and the pellet was resuspended and kept as *W8P*. This fraction contained all heavy components except the nucleus of the cell. The *J10P* fraction was obtained by a high-speed centrifugation (13000 *g*) of the supernatant of a slow spin (1000 *g*). Hence, in both fractions, the lightest fraction of the cell was omitted. The lightest fraction contained, in addition to ribosomes, the RNAs and proteins that may be involved in trans-splicing. The hypothetical RNA molecules that participated in the trans-splicing could have been lost at that step. Although total RNA was added to the reaction mixture, the RNA was from a preparation of which the isolation protocol was not very efficient in retaining small RNAs. This could have led to the decreased efficiency of trans-splicing in the system. Moreover, in the preparation of *J10P*, the heaviest fraction was missing, suggesting that large complexes of proteins attached to the intracellular vesicles could have been lost. Trans-splicing is hence possibly carried out by multiple factors that were present in both in the light and heavy fractions of the cell lysate.

Thirdly, 0.3% of Triton was used during the preparation of the sub-cellular fractions and its presence might have dissociated protein complexes or ribonucleoprotein complexes that are crucial for trans-splicing.

A successful in vitro system would then rely on improved methods of enriching the mitochondrial fraction and synthesis of substrates of higher affinity to the trans-splicing machinery.

5. DISCUSSION

5.1 Populations of mitochondrial gRNAs may exist

In eukaryotes, the nucleus-encoded mRNAs are capped in order to be transported to the cytoplasm for translation; while mitochondrial mRNAs do not require capping for translation. Therefore, in the in vitro radioactive cap-labeling experiment that uses total RNA as substrate, we may be able to label mostly mitochondrial primary transcripts, but not nucleus-encoded mRNAs. Ribosomal RNAs in the cytoplasm may also be detected but can be distinguished by their size and concentration. We have identified RNA bands that are smaller than 283 nt from this experiment (Fig. 12). Comparing to Simpson's experiment where he detected trypanosome mitochondrial small RNAs of 60 to 100 nt in size using capping reactions and later identified them as gRNAs [65], we suspect that similar gRNAs may exist for trans-splicing in the mitochondria of *D. papillatum*. Simpson's experiment revealed only one capped RNA band, whereas we saw at least three bands. The difference could be due to the fact that Simpson used purified mitochondrial RNA. Whether these small RNAs are of mitochondrial origin can be tested by Southern hybridization to the genomic DNA, which is being followed by another member of the lab.

5.2 Modules 4 and 5 sequencing results suggest that trans-splicing and editing share common steps

From the analysis of sequencing results of the RNA intermediates that contain the junction sequence of Modules 4 and 5, we conclude that the stretch of six uridines must be attached to the 3' end of Module 4 before Modules 4 and 5 join. We previously stated that the six Us were not likely a mini-exon because its size is significantly (>10 times) smaller than all known coding regions identified in *D. papillatum*. Now there is a second reason: if it was a mini-exon, then it would have no preference for any of the two modules to attach to first and we should be able to find both M4-6Us and 6U-M5 in our sequences. Instead, the addition of Us to M4 is similar to that in kinetoplastid U-insertion editing, where a specific number of Us are added to the 3' end of a pre-mRNA molecule.

In addition, we noticed that Module 4 tailed by six Us cannot be detected in RNA circularization experiment unless it is dephosphorylated at its 3' end. In kinetoplastid

RNA-editing, a phosphatase is required to remove the phosphate group at the 3’ end of the U-tail inserted by the TUTase, possibly as a mechanism to protect the edited 3’ ends from being erroneously joined to other modules before editing is completed [73]. Our observation of the phosphorylated U-tail in *D. papillatum* suggests that the machinery required for trans-splicing shares a common function with that of kinetoplastid RNA U-insertion editing.

Other than trans-splicing, the modules seem to undergo processing, where their 5’ and 3’ non-coding regions in the primary transcripts are cut off. The precision of processing is uncertain as we have seen both modules containing exact coding regions and modules with a small part of the coding region missing (in about 30% of all detected Module 4 sequences alone). It could be either an artifact due to the instability of RNA or the incorrect cuts made by the processing machinery. However, the expression of genetic information is not hindered, since none of the trans-spliced modules is missing any part of the coding region. Figure 21 summarizes the proposed model of trans-splicing of M4 and M5 according to our results.

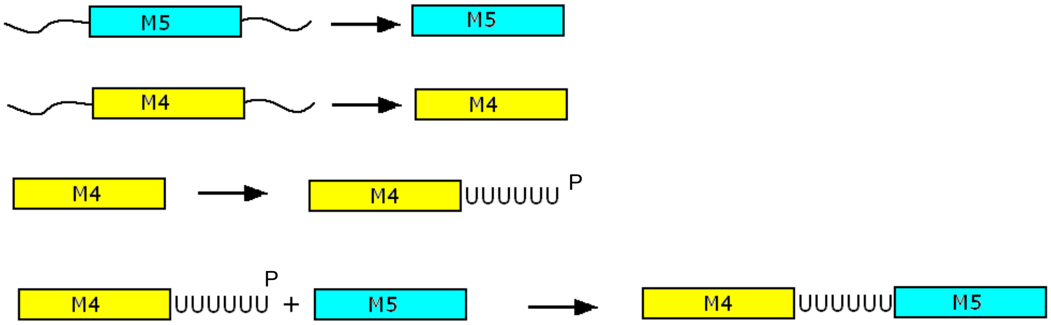


Figure 21. Proposed model of how the six Us are added between Module 4 and Module 5
First, Modules 4 and 5 are processed at both 5’ and 3’ ends so that only coding regions are left. Secondly, Module 4 is attached with six Us at the 3’ end, with a phosphate group at the 3’ end of the last U. Lastly, the phosphate group on U is removed and Module 5 is attached to the 3’ of the Us, completing the trans-splicing of M4 and M5.

The similar steps in the *D. papillatum* trans-splicing and the kinetoplastid U-insertional RNA editing suggest that similar factors may be involved. We suspect that in

D. papillatum trans-splicing, the recognition of the adjacent modules is probably carried out by the help of gRNAs.

5.3 Kinetic inference of Modules 4 and 5 trans-splicing

If we assume non-biased sampling of the sequenced clones, then the proportion of an RNA intermediate over all intermediates represents its relative abundance in the original steady-state RNA population in vivo. Hence, we can use the abundance of the intermediates to infer the kinetics of the proposed trans-splicing model for M4 and M5. Similar inferences were made in the study of kinetoplastid RNA editing where different amounts of partially edited transcripts were determined statistically after sequencing RT-PCR products [90]. According to our sequencing results of partially trans-spliced mRNAs, the fastest step is the addition of six Us to the 3' end of M4 and "pauses" that occur during 5' processing of M4 and the ligation of the U-tailed M4 with M5. The kinetics of the reactions is consistent with the hypothesis that both processing and ligation reactions require the formation of a duplex between mRNA and its corresponding gRNA for specific site recognition, since more time would be required for them to "find" each other among all kinds of RNA molecules. On the other hand, appending six Us to M4 happens after the duplex is already formed and the enzymatic reaction can be quick.

5.4 *D. papillatum* gRNAs may not conform to the "rules"

As several evidences indirectly support our hypothesis that trans-splicing in *D. papillatum* involves gRNA, we examined our candidates (see Section 5.4 and 5.5) for features similar to those of kinetoplastid gRNAs. For that, we constructed regular expressions, which are formulae that summarize patterns in strings of text, to represent the expected profiles of gRNAs of all module junctions (Appendix 1). We identified possible gRNA profiles for four of the eight module junctions (Table 1, "Anti-sense to the junction", and Appendix 3). This indicates that we might have amplified the guiding nucleotides of the gRNAs and the guiding nucleotides can pair with the junctions perfectly, without G:U pairing. The fact that we identified these guiding sequences repeatedly at most junctions (Appendix 3), thought at low frequency for each junction, suggests that the gRNAs may exist but at low a concentration. Due to the primer design, we could not detect sequences beyond the guiding regions in this experiment. RT-PCR using a specific primer and a SMART primer had the potential to amplify the gRNA more towards the 5’ end, beyond the guiding nucleotides. However, we did not recover the guiding nucleotides and reverse primer sequence for M4/M5 junction (results in Table 3), possibly because of their low concentration. Alternatively, this may suggest that the 5’ anchor of the gRNA deviate from the perfect sequence complementarity to the flanking regions of the M4/M5 junction. In other words, the gRNA sequences may differ from the “rules” that we specified in the Motsearch program.

The unknown sequences (see Table 1, Table 3), on the other hand, may contain sequences that are capable to act as gRNAs. The fact that their sequences were unknown means that they are not amplification products of any known mRNA. In order to examine their potential to bring two trans-splicing modules close to each other, we relaxed the search criteria in the gRNA profiles and analyzed the unknown product sequences of the convergent RT-PCR across M4/M5. The following three criteria were used to identify gRNA candidates among the unknowns. First, its sequence contains binding sites of both Module 4 and Module 5. Second, stem-loop structures can form between the two binding sites and bring them close to each other. Third, the nucleotides that are not in the stem-loop structures form a continuous stretch of six A/Gs. We found six sequences that satisfy these criteria. Figure 22 summarizes the possible gRNA conformations that we have detected. We used the Mfold program [91] to calculate the secondary structures and the outputs are included in Appendix 4.

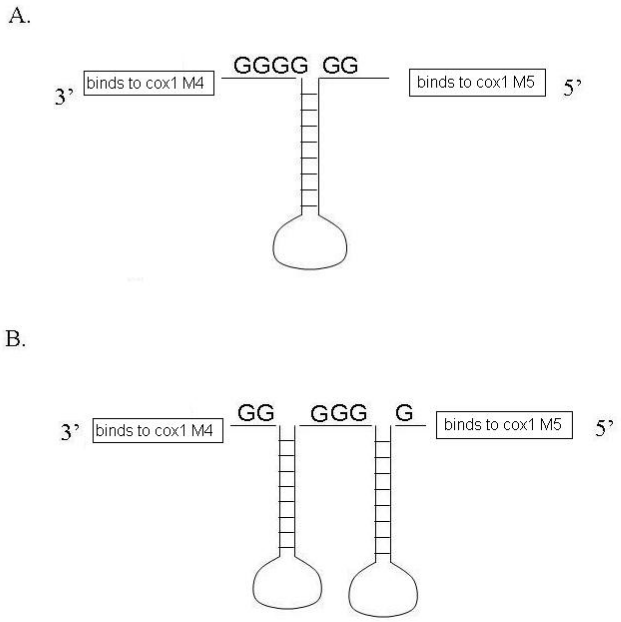


Figure 22. Potential gRNAs with secondary structures for six Us between M4/M5 in trans-splicing The guiding nucleotides for the six Us inserted between Module 4 and 5 may not conform to the known rules. Secondary structures in the sequence may help to bring distant guiding nucleotides together. Here we see different ways that a stretch of six of A or Gs can be formed between the two primers. (A) In clones dp6912 and dp6933, the sequences between the two primers can form a single stem-loop, while nucleotides that are not in the stem-loop are available for guiding. (B) In clones dp6917, dp6932, dp6948, and dp7139. The guiding nucleotides are interrupted by two stem-loops. All secondary structures are calculated using Mfold, a free on-line program that calculates secondary structures of RNA.

Among sequences we amplified for other module junctions (RT-PCR set 1) we noticed that the RT-PCR targeting the M8/M9 junction amplified numerous un-processed modules of different mitochondrial genes (Table 1). Most of them followed one common pattern: a partial coding sequence and a non-coding sequence of a non-targeted gene were connected, sometimes included even an extra unknown sequence beyond the non-coding sequence. The observed complementarities offered one more possible way that gRNAs may be involved: the joining of two modules may be directed by the primary transcript of another module.

We detected modules of genes such as *cox2*, *cox3*, *nad5*, as well as the small subunit of the ribosome (*rns*) among the amplified sequences. According to the sequencing results, we propose that modules 8 and 9 joining may be carried out in the configuration shown in Fig. 23.

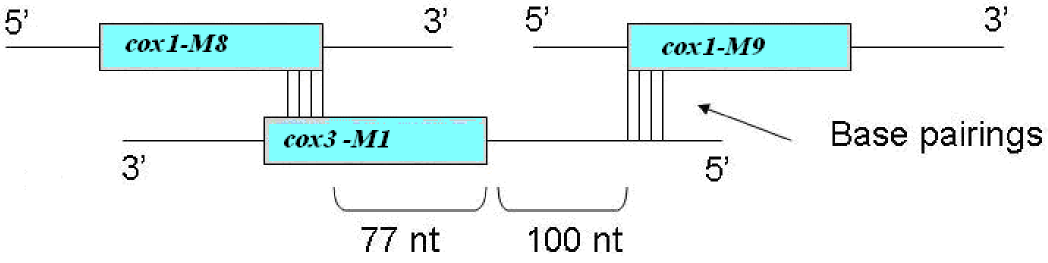


Figure 23. Modules of mRNA may act as gRNA
In clone dp9356, M1 of *cox3* gene was detected by RT-PCR targeting gRNAs for joining M8 and M9 of *cox1* mRNA. Primer sequences were partially (about 10 nt) identified at the ends of the amplified sequences of *cox3*-M1, indicating that pairings with M8 and M9 span about 10 nt.

The distance between two base-pairing regions in *cox3*-M1 is 177 nucleotides. For other gene modules identified, the distances are around 200 nucleotides. Joining of two modules around 200 nt apart seemed unlikely. We suspected whether the "guiding" modules were rich in secondary structure, as they are highly G-C rich, so that the two base pairing regions were brought close for trans-splicing.

Mfold was used to calculate the secondary structure of the sequences between the base-paired regions in M1 of *cox3* and the result showed that it was quite possible that the two paired regions are within 20 nt of each other (Fig. 24). The complementarity within *cox3*-M1 thus might bring M8 and M9 of *cox1* together and it could greatly facilitate the ligation of M8 and M9 of *cox1*.

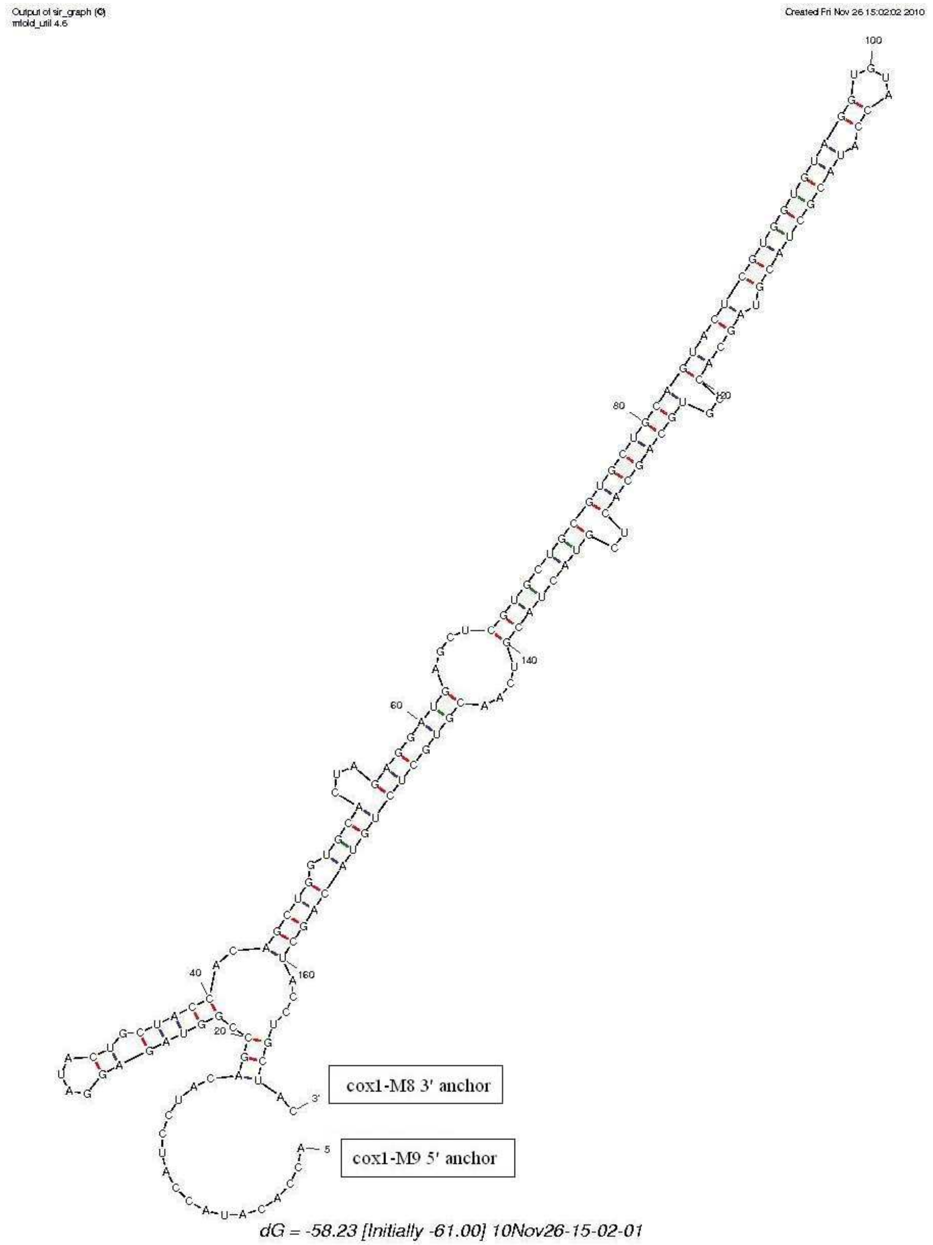


Figure 24. Secondary structure of *cox3*-module 1 (clone dp9356)
Secondary structures can be formed between regions that pair with M8 and M9 of the *cox1* gene. The 5' end of the module, which pairs with M9 of *cox1*, and the 3' end of the module, which pairs with M8 of *cox1*, can be brought close to each other within 20 nucleotides distance (single-stranded loop at the bottom).

If this proposed model can be confirmed by in vitro assays, it means that the expression of genes that undergo trans-splicing in the mitochondria of *D. papillatum* takes place in a coordinated fashion: trans-splicing of one mitochondrial gene requires the presence of primary transcripts of the others. This entails that the synthesis of one protein in the respiratory chain is regulated by the presence of the mRNA of others, and that mRNA may act both as a message and as a regulator of translation via trans-splicing.

5.5 Biological relevance of trans-splicing in *D. papillatum*

Trans-splicing may offer advantages that helped *D. papillatum* survive through natural selection. First, when a gene is broken into pieces and divided among different chromosomes, smaller pieces of the gene can be transcribed simultaneously, hence the time required to transcribe each gene will be much less. As we have seen previously, the neighboring modules of a gene were trans-spliced without any particular order. This means that all neighboring modules can be trans-spliced simultaneously; therefore, the time needed for trans-splicing of all nine modules would be the same as the trans-splicing of any two. The parallelization of module transcription and trans-splicing greatly reduce the time required for gene expression. Second, as we have mentioned in the previous section, trans-splicing may offer an extra control point of gene expression in the mitochondria. In cap-dependent translation in eukaryotes, the eukaryotic translation initiation factors (eIFs) control translation at the rate-limiting step of initiation [92, 93]. One crucial eIF that connects the extracellular stimuli to global translation rates is the cap-binding protein, eIF4E, which acts in the eIF4F complex and regulates the binding of 40S ribosomal subunit to the 5' UTR of mRNA [94]. The 5' cap structure is, however, absent in the mitochondria. To regulate translation initiation, *D. papillatum* mitochondria may fragment genes into pieces and only when the assembly of the full-length mRNA is completed, will the synthesis of a protein begin. This concept was first suggested by the Simpson's group regarding RNA-editing in the kinetoplastid mitochondria as a translational control mechanism [95]. With recent advances in the study of kinetoplastid RNA editing, physical evidence showed that special mechanisms exist for the ribosomes to select the fully edited mRNAs for translation, supporting Simpson's original hypothesis [96].

On the other hand, the emergence of trans-splicing may be initially random, neutral, and maybe even wasteful. We could only speculate that the "advantages" mentioned above may not be necessary until the environment changed, and subsequently the species that underwent the “wasteful” mutations turned out to be well adapted to its new niche.

5.6 Future studies

In our experimental approach employing RT-PCR, we might have detected the hypothesized gRNAs according to the profiles established based on kinetoplastid editing. New methods need to be employed to identify gRNAs in a more efficient and non-biased fashion.

Recent advancement in parallel sequencing combined with shotgun cloning has offered a powerful tool to identify small RNAs (sRNAs) [97]. Bioinformatics were applied to filter the sequencing results, and then used to characterize the final list of gRNA candidates. The top “hits” were confirmed for their targeted genes, using Northern blot. The advantage of this method is that it is a genome-wide approach and sRNAs of unknown functions are more likely to be identified. Alternatively, high-density oligonucleotide probe arrays can be used to detect sRNAs transcripts [98]. For both methods, the greatest advantage is that they are less biased than an RT-PCR approach and they are both high-throughput.

Eventually we will need to confirm the biological functions of the gRNAs that we identified. Though the pilot in vitro system was not successful, we understand now that a mitochondrial purification and the substrates need to be optimized and all experimental conditions tested systematically.

6. REFERENCES

1. Marande, W., Lukes, J., and Burger, G., *Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids.* Eukaryot. Cell, 2005. **4**: p. 1137-1146.
2. Reinhold-Hurek, B. and D.A. Shub, *Self-splicing introns in tRNA genes of widely divergent bacteria.* Nature, 1992. **357**(6374): p. 173-6.
3. Abelson, J., C.R. Trotta, and H. Li, *tRNA splicing.* J Biol Chem, 1998. **273**(21): p. 12685-8.
4. Reyes, V.M. and J. Abelson, *Substrate recognition and splice site determination in yeast tRNA splicing.* Cell, 1988. **55**(4): p. 719-30.
5. Wich, G., W. Leinfelder, and A. Bock, *Genes for stable RNA in the extreme thermophile Thermoproteus tenax: introns and transcription signals.* EMBO J, 1987. **6**(2): p. 523-8.
6. Kjems, J., et al., *A unique tRNA intron in the variable loop of the extreme thermophile Thermofilum pendens and its possible evolutionary implications.* J Biol Chem, 1989. **264**(30): p. 17834-7.
7. Thompson, L.D., et al., *Transfer RNA intron processing in the halophilic archaebacteria.* Can J Microbiol, 1989. **35**(1): p. 36-42.
8. Thompson, L.D. and C.J. Daniels, *A tRNA(Trp) intron endonuclease from Halobacterium volcanii. Unique substrate recognition properties.* J Biol Chem, 1988. **263**(34): p. 17951-9.
9. Thompson, L.D. and C.J. Daniels, *Recognition of exon-intron boundaries by the Halobacterium volcanii tRNA intron endonuclease.* J Biol Chem, 1990. **265**(30): p. 18104-11.
10. Lambowitz, A.M. and S. Zimmerly, *Mobile group II introns.* Annu Rev Genet, 2004. **38**: p. 1-35.
11. Schafer, B., et al., *The mitochondrial genome of fission yeast: inability of all introns to splice autocatalytically, and construction and characterization of an intronless genome.* Mol Gen Genet, 1991. **225**(1): p. 158-67.
12. Gampel, A., M. Nishikimi, and A. Tzagoloff, *CBP2 protein promotes in vitro excision of a yeast mitochondrial group I intron.* Mol Cell Biol, 1989. **9**(12): p. 5424-33.
13. Lambowitz, A.M. and P.S. Perlman, *Involvement of aminoacyl-tRNA synthetases and other proteins in group I and group II intron splicing.* Trends Biochem Sci, 1990. **15**(11): p. 440-4.
14. Lang, B.F., M.J. Laforest, and G. Burger, *Mitochondrial introns: a critical view.* Trends Genet, 2007. **23**(3): p. 119-25.
15. Bonen, L., *Cis- and trans-splicing of group II introns in plant mitochondria.* Mitochondrion, 2008. **8**(1): p. 26-34.
16. Burger, G., et al., *Group I-intron trans-splicing and mRNA editing in the mitochondria of placozoan animals.* Trends Genet, 2009. **25**(9): p. 381-6.
17. Cech, T.R., *Self-splicing of group I introns.* Annu Rev Biochem, 1990. **59**: p. 543-68.
18. Bonen, L. and J. Vogel, *The ins and outs of group II introns.* Trends Genet, 2001. **17**(6): p. 322-31.

19. Haugen, P., D.M. Simon, and D. Bhattacharya, *The natural history of group I introns.* Trends Genet, 2005. **21**(2): p. 111-9.

20. Davies, R.W., et al., *Making ends meet: a model for RNA splicing in fungal mitochondria.* Nature, 1982. **300**(5894): p. 719-24.

21. Michel, F. and B. Dujon, *Conservation of RNA secondary structures in two intron families including mitochondrial-, chloroplast- and nuclear-encoded members.* EMBO J, 1983. **2**(1): p. 33-8.

22. Konforti, B.B., et al., *Ribozyme catalysis from the major groove of group II intron domain 5.* Mol Cell, 1998. **1**(3): p. 433-41.

23. van der Veen, R., et al., *Excised group II introns in yeast mitochondria are lariats and can be formed by self-splicing in vitro.* Cell, 1986. **44**(2): p. 225-34.

24. Muller, M.W., et al., *Fate of the junction phosphate in alternating forward and reverse self-splicing reactions of group II intron RNA.* J Mol Biol, 1991. **222**(2): p. 145-54.

25. Morl, M. and C. Schmelzer, *Integration of group II intron bI1 into a foreign RNA by reversal of the self-splicing reaction in vitro.* Cell, 1990. **60**(4): p. 629-36.

26. Hall, S.L. and R.A. Padgett, *Conserved sequences in a class of rare eukaryotic nuclear introns with non-consensus splice sites.* J Mol Biol, 1994. **239**(3): p. 357-65.

27. Jackson, I.J., *A reappraisal of non-consensus mRNA splice sites.* Nucleic Acids Res, 1991. **19**(14): p. 3795-8.

28. Valadkhan, S., *Role of the snRNAs in spliceosomal active site.* RNA Biol. **7**(3): p. 345-53.

29. Mount, S.M., et al., *The U1 small nuclear RNA-protein complex selectively binds a 5' splice site in vitro.* Cell, 1983. **33**(2): p. 509-18.

30. Black, D.L., B. Chabot, and J.A. Steitz, *U2 as well as U1 small nuclear ribonucleoproteins are involved in premessenger RNA splicing.* Cell, 1985. **42**(3): p. 737-50.

31. Kramer, A., *The structure and function of proteins involved in mammalian pre-mRNA splicing.* Annu Rev Biochem, 1996. **65**: p. 367-409.

32. Ruskin, B., et al., *Excision of an intact intron as a novel lariat structure during pre-mRNA splicing in vitro.* Cell, 1984. **38**(1): p. 317-31.

33. Padgett, R.A., et al., *Lariat RNA's as intermediates and products in the splicing of messenger RNA precursors.* Science, 1984. **225**(4665): p. 898-903.

34. Toor, N., K.S. Keating, and A.M. Pyle, *Structural insights into RNA splicing.* Curr Opin Struct Biol, 2009. **19**(3): p. 260-6.

35. Dayie, K.T. and R.A. Padgett, *A glimpse into the active site of a group II intron and maybe the spliceosome, too.* RNA, 2008. **14**(9): p. 1697-703.

36. Valadkhan, S., *The spliceosome: a ribozyme at heart?* Biol Chem, 2007. **388**(7): p. 693-7.

37. Abelson, J., *Is the spliceosome a ribonucleoprotein enzyme?* Nat Struct Mol Biol, 2008. **15**(12): p. 1235-7.

38. Murphy, W.J., Watkins, K.P. & Agabian, N., *Identification of a novel branch structure as an intermediate in trypanosome mRNA processing: Evidence for trans-splicing.* Cell, 1986. **47**: p. 517–525.

39. Sutton, R.E.B., J.C., *Evidence for trans-splicing in trypanosomes.* Cell, 1986. **47**(4): p. 527-35.
40. Guthrie, C., *Messenger RNA splicing in yeast: clues to why the spliceosome is a ribonucleoprotein.* Science, 1991. **253**(5016): p. 157-63.
41. Blumenthal, T. and J. Thomas, *Cis and trans mRNA splicing in C. elegans.* Trends Genet, 1988. **4**(11): p. 305-8.
42. Tessier, L.H., et al., *Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in Euglena.* EMBO J, 1991. **10**(9): p. 2621-5.
43. Sturm, N.R., et al., *Diplonema spp. possess spliced leader RNA genes similar to the Kinetoplastida.* J Eukaryot Microbiol, 2001. **48**(3): p. 325-31.
44. Li, B., *Human acyl-CoA:cholesterol acyltransferase-1 (ACAT-1) gene organization and evidence that the 4.3-kilobase ACAT-1 mRNA is produced from two different chromosomes.* Biol Chem, 1999. **274**(16): p. 11060-71.
45. Pirrotta, V., *Trans-splicing in Drosophila.* BioEssays, 2002. **24**: p. 988-991.
46. Bonen, L., *Trans-splicing of pre-mRNA in plants, animals, and protists.* FASEB J., 1993. **7**(1): p. 40-6.
47. Kohchi, T., et al., *A nicked group II intron and trans-splicing in liverwort, Marchantia polymorpha, chloroplasts.* Nucleic Acids Res, 1988. **16**(21): p. 10025-36.
48. Koller, B., et al., *Evidence for in vivo trans splicing of pre-mRNAs in tobacco chloroplasts.* Cell, 1987. **48**(1): p. 111-9.
49. Konarska, M.M., Padgett, R. A. & Sharp, P.A., *Trans splicing of mRNA precursors in vitro.* Cell, 1985. **42**(1): p. 165-171.
50. Perron, K., *A factor related to pseudouridine synthases is required for chloroplast group II intron trans-splicing in Chlamydomonas reinhardtii.* EMBO J., 1999. **18**(22): p. 6481-90.
51. Grewe, F., et al., *A trans-splicing group I intron and tRNA-hyperediting in the mitochondrial genome of the lycophyte Isoetes engelmannii.* Nucleic Acids Res, 2009. **37**(15): p. 5093-104.
52. Pombert, J.F. and P.J. Keeling, *The mitochondrial genome of the entomoparasitic green alga Helicosporidium.* PLoS One, 2010. **5**(1): p. e8954.
53. Eul, J., M. Graessmann, and A. Graessmann, *Experimental evidence for RNA trans-splicing in mammalian cells.* EMBO J, 1995. **14**(13): p. 3226-35.
54. Triemer, R.E., and M. A. Farmer, *An altrastructural comparison of the mitotic apparatus, feeding apparatus, flagellar apparatus and cytoskeleton in euglenoids and kinetoplastids.* Protoplasma, 1991. **164**: p. 91-104.
55. Simpson, A.G.B., *The identity and composition of the Euglenozoa.* Arch. Protistenkd, 1997. **148**: p. 318-328.
56. Simpson, A.G. and A.J. Roger, *Protein phylogenies robustly resolve the deep-level relationships within Euglenozoa.* Mol Phylogenet Evol, 2004. **30**(1): p. 201-12.
57. Maslov, D.A., S. Yasuhira, and L. Simpson, *Phylogenetic affinities of Diplonema within the Euglenozoa as inferred from the SSU rRNA gene and partial COI protein sequences.* Protist, 1999. **150**(1): p. 33-42.

58. Simpson, A.G., J. Lukes, and A.J. Roger, *The evolutionary history of kinetoplastids and their kinetoplasts.* Mol Biol Evol, 2002. **19**(12): p. 2071-83.
59. Gray, M.W., B.F. Lang, and G. Burger, *Mitochondria of protists.* Annu Rev Genet, 2004. **38**: p. 477-524.
60. Roy, J., et al., *Unusual mitochondrial genome structures throughout the Euglenozoa.* Protist, 2007. **158**(3): p. 385-96.
61. Marande, W. and G. Burger, *Mitochondrial DNA as a genomic jigsaw puzzle.* Science, 2007. **318**(5849): p. 415.
62. Vlcek, C., et al., *Systematically fragmented genes in a multipartite mitochondrial genome.* Nucleic Acids Res.
63. Nowacki, M., *RNA-mediated epigenetic programming of a genome-rearrangement pathway Nature.* Nature, 2008. **451**(7175): p. 153-158.
64. Stuart, K.D., et al., *Complex management: RNA editing in trypanosomes.* Trends Biochem Sci, 2005. **30**(2): p. 97-105.
65. Sturm, N.R. and L. Simpson, *Kinetoplast DNA minicircles encode guide RNAs for editing of cytochrome oxidase subunit III mRNA.* Cell, 1990. **61**(5): p. 879-84.
66. Stuart, K., et al., *RNA editing in kinetoplastid protozoa.* Microbiol Mol Biol Rev, 1997. **61**(1): p. 105-20.
67. Simpson, L., S. Sbicego, and R. Aphasizhev, *Uridine insertion/deletion RNA editing in trypanosome mitochondria: a complex business.* RNA, 2003. **9**(3): p. 265-76.
68. Kapushoc, S.T. and L. Simpson, *In vitro uridine insertion RNA editing mediated by cis-acting guide RNAs.* RNA, 1999. **5**(5): p. 656-69.
69. Carnes, J., et al., *RNA editing in Trypanosoma brucei requires three different editosomes.* Mol Cell Biol, 2008. **28**(1): p. 122-30.
70. Ernst, N.L., et al., *TbMP57 is a 3' terminal uridylyl transferase (TUTase) of the Trypanosoma brucei editosome.* Mol Cell, 2003. **11**(6): p. 1525-36.
71. Aphasizhev, R., I. Aphasizheva, and L. Simpson, *A tale of two TUTases.* Proc Natl Acad Sci U S A, 2003. **100**(19): p. 10617-22.
72. Trotter, J.R., et al., *A deletion site editing endonuclease in Trypanosoma brucei.* Mol Cell, 2005. **20**(3): p. 403-12.
73. Niemann, M., et al., *Kinetoplastid RNA editing involves a 3' nucleotidyl phosphatase activity.* Nucleic Acids Res, 2009. **37**(6): p. 1897-906.
74. Gao, G. and L. Simpson, *Is the Trypanosoma brucei REL1 RNA ligase specific for U-deletion RNA editing, and is the REL2 RNA ligase specific for U-insertion editing?* J Biol Chem, 2003. **278**(30): p. 27570-4.
75. Harris, M.E., D.R. Moore, and S.L. Hajduk, *Addition of uridines to edited RNAs in trypanosome mitochondria occurs independently of transcription.* J Biol Chem, 1990. **265**(19): p. 11368-76.
76. Frech, G.C., et al., *In vitro RNA editing-like activity in a mitochondrial extract from Leishmania tarentolae.* EMBO J, 1995. **14**(1): p. 178-87.
77. Sabatini, R.S., et al., *Biochemical methods for analysis of kinetoplastid RNA editing.* Methods, 1998. **15**(1): p. 15-26.
78. Seiwert, S.D. and K. Stuart, *RNA editing: transfer of genetic information from gRNA to precursor mRNA in vitro.* Science, 1994. **266**(5182): p. 114-7.

79. Cruz-Reyes, J. and B. Sollner-Webb, *Trypanosome U-deletional RNA editing involves guide RNA-directed endonuclease cleavage, terminal U exonuclease, and RNA ligase activities.* Proc Natl Acad Sci U S A, 1996. **93**(17): p. 8901-6.

80. Kable, M.L., et al., *RNA editing: a mechanism for gRNA-specified uridylate insertion into precursor mRNA.* Science, 1996. **273**(5279): p. 1189-95.

81. Byrne, E.M., G.J. Connell, and L. Simpson, *Guide RNA-directed uridine insertion RNA editing in vitro.* EMBO J, 1996. **15**(23): p. 6758-65.

82. Connell, G.J., E.M. Byrne, and L. Simpson, *Guide RNA-independent and guide RNA-dependent uridine insertion into cytochrome b mRNA in a mitochondrial lysate from Leishmania tarentolae. Role of RNA secondary structure.* J Biol Chem, 1997. **272**(7): p. 4212-8.

83. Brown, L.M., et al., *A cis-acting A-U sequence element induces kinetoplastid U-insertions.* J Biol Chem, 1999. **274**(10): p. 6295-304.

84. Rusche, L.N., et al., *Purification of a functional enzymatic editing complex from Trypanosoma brucei mitochondria.* EMBO J, 1997. **16**(13): p. 4069-81.

85. Pelletier, M., L.K. Read, and R. Aphasizhev, *Isolation of RNA binding proteins involved in insertion/deletion editing.* Methods Enzymol, 2007. **424**: p. 75-105.

86. J. Sambrook, E.F.F., and T. Maniatis, *Molecular Cloning: A Laboratory Manual*. 2 ed. Vol. 1. 1989: Cold Spring Harbor Laboratory Press.

87. Pollard, V.W., et al., *Organization of minicircle genes for guide RNAs in Trypanosoma brucei.* Cell, 1990. **63**(4): p. 783-90.

88. Shuman, S., *What messenger RNA capping tells us about eukaryotic evolution.* Nat Rev Mol Cell Biol, 2002. **3**(8): p. 619-25.

89. Cowling, V.H., *Regulation of mRNA cap methylation.* Biochem J, 2009. **425**(2): p. 295-302.

90. Sturm, N.R. and L. Simpson, *Partially edited mRNAs for cytochrome b and subunit III of cytochrome oxidase from Leishmania tarentolae mitochondria: RNA editing intermediates.* Cell, 1990. **61**(5): p. 871-8.

91. Zuker, M., *Mfold web server for nucleic acid folding and hybridization prediction.* Nucleic Acids Res, 2003. **31**(13): p. 3406-15.

92. Kapp, L.D. and J.R. Lorsch, *The molecular mechanics of eukaryotic translation.* Annu Rev Biochem, 2004. **73**: p. 657-704.

93. Gebauer, F. and M.W. Hentze, *Molecular mechanisms of translational control.* Nat Rev Mol Cell Biol, 2004. **5**(10): p. 827-35.

94. Sonenberg, N., *eIF4E, the mRNA cap-binding protein: from basic discovery to translational research.* Biochem Cell Biol, 2008. **86**(2): p. 178-83.

95. Simpson, L. and J. Shaw, *RNA editing and the mitochondrial cryptogenes of kinetoplastid protozoa.* Cell, 1989. **57**(3): p. 355-66.

96. Aphasizheva, I., et al., *Pentatricopeptide repeat proteins stimulate mRNA adenylation/uridylation to activate mitochondrial translation in trypanosomes.* Mol Cell, 2011. **42**(1): p. 106-17.

97. Liu, J.M., et al., *Experimental discovery of sRNAs in Vibrio cholerae by direct cloning, 5S/tRNA depletion and parallel sequencing.* Nucleic Acids Res, 2009. **37**(6): p. e46.

98. Wassarman, K.M., et al., *Identification of novel small RNAs using comparative genomics and microarrays.* Genes Dev, 2001. **15**(13): p. 1637-51.

99. Yanagiya, A., et al., *Requirement of RNA binding of mammalian eukaryotic translation initiation factor 4GI (eIF4GI) for efficient interaction of eIF4E with the mRNA cap.* Mol Cell Biol, 2009. **29**(6): p. 1661-9.

7. APPENDICES

Appendix 1: Hypothetical gRNA sequences for *cox1* trans-splicing

For the *cox1* gene, nine gene modules give rise to eight module junctions in the mature mRNA. The gRNAs that direct the joining of two modules should contain sequences that perfectly pair with their junction (including G:U pairing). Since the anchor region of known gRNAs are usually 5-25 nt long, six nt of anchor sequences flanking the junction are included in the profile. Below, the ends of the modules are labeled with "M1" to "M9" on top and "|" indicates the junction between the modules. Alternative nucleotides that satisfy the profile are written in the same column. The "|"s between two rows of nucleotides represent base-pairing. A regular expression of the profile of a gRNA is written in the 5' to 3' direction below the representation of the junction. The regular expressions are used in the Motsearch program (Burger et al. unpublished). In regular expressions, square brackets ([]) enclose nucleotides that are all possible for a given position. Curved brackets ({}) enclose the number of repetitions that the nucleotide preceding the brackets occurs. After searching the regular expression against the sequences obtained from RT-PCR targeting the gRNA of the junction, we recorded the number of hits (indicated in brackets).

1. M1/M2:

```
           M1 | M2
mRNA:  ACGACG  CATGGC
       ||||||  ||||||
gRNA : TGCTGC  GTACCG
         T  T    GTT

gRNA regular expression 5'->3':
G[CT]{2}[AG]TG  [CT]GT[CT]GT
(0 hits in Motsearch, see Appendix 3)
```

2. M2/M3:

```
           M2 | M3
mRNA:  TCGGGA  CATGGC
       ||||||  ||||||
gRNA:  AGCCCT  GTACCG
       G TTT     GTT
```

```
gRNA regular expression 5’->3’:
G[CT]{2}[AG]TG  T[CT]{3}G[AG]
(6 hit in Motsearch)
```

3. M3/M4:

```
          M3 | M4
mRNA: GTGGCA  ATCACA
      ||||||  ||||||
gRNA: CACCGT  TAGTGT
      TGTT     G

gRNA regular expression 5’->3’:
TGTG[AG]T  TG[CT]{2}[AG][CT]
(0 hit in Motsearch)
```

4. M4/M5

```
          M4|      | M5
mRNA: GAGGAC TTTTTT CGCTCT
      |||||| |||||| ||||||
gRNA: CTCCTG AAAAAA GCGAGA
      T TT   GGGGGG  T G G

gRNA regular expression 5’->3’:
[AG]G[AG]G[CT]G[AG]{6} GT[CT]{2}T[CT]
(13 hit in Motsearch)
```

5. M5/M6

```
          M5 | M6
mRNA: ATGGTG  GGACTG
      ||||||  ||||||
gRNA: TACCAC  CCTGAC
       GTTGT  TT  GT

gRNA regular expression 5’->3’:
[CT][AG]GT[CT]{2}  [CT][AG][CT]{2}[AG]T
```

(5 hits in Motsearch)

6. M6/M7

```
          M6 | M7
mRNA: TCTCAT  TAGGAG
      ||||||  ||||||
gRNA: AGAGTA  ATCCTC
      G G  G  G TT T

gRNA regular expression 5’->3’:
[CT]T[CT]{2}T[AG] [AG]TG[AG]G[AG]
(0 hit in Motsearch)
```

7. M7/M8

```
          M7 | M8
mRNA: CGTGTA  CAGGTG
      ||||||  ||||||
gRNA: GCACAT  GTCCAC
       TGTG     TTGT

gRNA regular expression 5’->3’:
[CT][AG][CT]{2}TG  T[AG][CT][AG][CT]G
(0 hit in Motsearch)
```

8. M8/M9

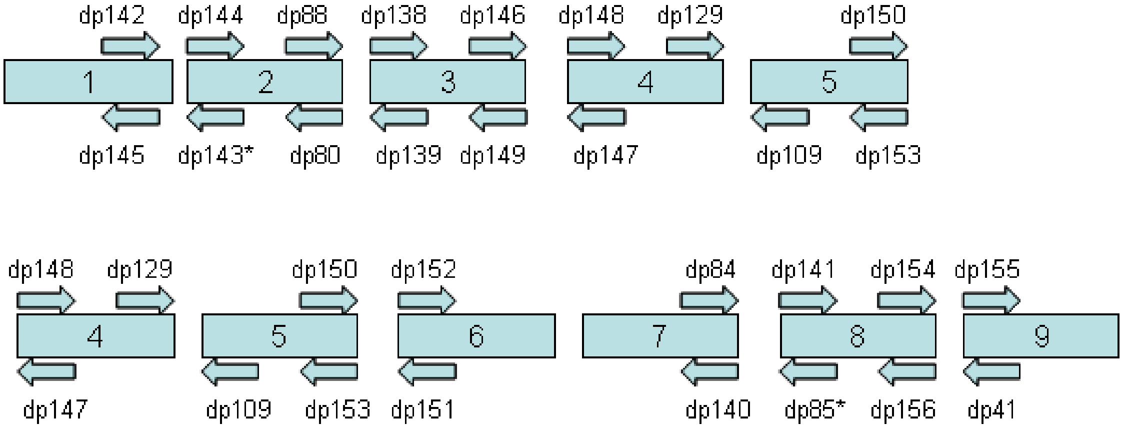
```
          M8 | M9
mRNA: CCTAGG  TACAGT
      ||||||  ||||||
gRNA: GGATCC  ATGTCA
        G TT  G   TG

gRNA regular expression 5’->3’:
[AG][CT]TGT[AG]  [CT]{2}T[AG]GG
(1 hit in Motsearch)
```

Appendix 2a. Primers used in the RT-PCR

```
>SMART-IV (Clontech SMART cDNA kit primer) N-1 (N-1 = A, G, or C)
AAGCAGTGGTATCAACGCAGAGTGGCCATTACGGCCGGG
>CDS-III (Clontech SMART cDNA kit primer)
ATTCTAGAGGCCGAGGCGGCCGACATG
>dp35 (cox1.3-r): 5' 3' Tm=60
GGTATCCATCAGGTGCATCT
>dp67:
CCTGGTACGTTGTGGCCTTC
>dp80:  Rev module 3 cox1 (Tm 46)
CCACTAGCAGCCATG
>dp84 (cox1-F6r-rev):
GCTGCATGGTTACTCCGTGTA
>dp85(cox1-F7f-rev):
CCAAGAGGACCACCTGAGTCA
>dp88 (cox1-F1r_rev)
CCCTAAGGTGAACAACGTCGG
>dp109 (cox1.2f-rev):
CGTACACCATGCCCAGCATGTTGTAGAGC
>dp129: 3'end of Module 4's coding strand; for hypothetical gRNA;
ATTCCCTACATCGAGGAGGA
>dp138(rev. compl. of dp84 + 6nt extension )
ATGGCTGCTAGTGGATGGCT
>dp139(rev. compl. of dp88)
CCGACGTTGTTCACCTTAGGG
>dp140(rev. compl. of dp84)
TACACGGAGTAACCATGCAGC
>dp141 (rev. compl. of dp85)
TGACTCAGGTGGTCCTCTTGG
>dp142 (m1 3' fwd): GTACAACGTCCTCACTACGA
>dp145 (m1 3' rev): TCGTAGTGAGGACGTTGTAC

>dp143 (m2 5' fwd): AAGAGCATGAGGATGGCTAC
>dp144 (m2 5' rev): ATGGCCATCCTCATGCTCTT

>dp146 (m3 3' fwd): TTGTATCCATGGAGCGTGGC
>dp149 (m3 3' rev): GCCACGCTCCATGGATACAA

>dp148 (m4 5' fwd): TCACAGGTGCACTGCTAGTA
>dp147 (m4 5' rev): TACTAGCAGTGCACCTGTGA

>dp150 (m5 3' fwd): AAGGTGTTCTCCTGGATGGT
>dp153 (m5 3' rev): ACCATCCAGGAGAACACCTT

>dp152 (m6 5' fwd): GACTGCGTAGGATGGCACTA
>dp151 (m6 5' rev): TAGTGCCATCCTACGCAGTC

>dp154 (m8 3' fwd): TGCACCTGATGGATACCTAG
>dp155 (m9 5' fwd): AGTGGTATCCACCACCTGTG
>dp156 (m8 3' rev): CTAGGTATCCATCAGGTGCA
>dp157;T7 RNA polymerase promoter(underlined) + 5' 20 nt of M3, Ta = 57
°C.
TAATACGACTCACTATAGGGCATGGCTGCTAGTGGATGGC
>dp158; reverse complement of 3' of M3, Ta = 55 °C.
TGCCACGCTCCATGGATACA
```

```
>dp159; T7 RNA polymerase promoter (underlined) + 5' 17 nt of M5, Ta =
47 deg. Cel.
TAATACGACTCACTATAGGGCGCTCTACAACATGCTG
>dp160; reverse complement of 3' of M5.
CACCATCCAGGAGAACACCT
>dp161; M4 forward primer, with T7 promoter (underlined) at the 5' end.
TAATACGACTCACTATAGGGATCACAGGTGCACTGCTAGT
>dp162; M4 reverse primer (without 6Ts).
GTCCTCCTCGATGTAGGGAA
>dp163; M4+6T reverse primer (with 6Ts).
AAAAAAGTCCTCCTCGATGT
```

Appendix 2b. A summary map of the primers used in convergent and divergent RT-PCRs



* Primers dp143 and dp85 contain erroneous mismatches.

Appendix 3. Anti-sense sequences detected with convergent RT-PCR on *cox*1 module junctions

Junction	frequency	% freq.	Primer 1 (RT primer)	Primer 2	clone ID	sequencing
M1/M2	--	-- (1)	dp142	dp143		
M2/M3	6 in 36	16.7	dp88	dp80	dp9230	i94
					dp9241	i94
					dp9244	i94
					dp9245	i94
					dp9264	i94
					dp9268	i94
M3/M4	n.d.	n.d. (2)	dp146	dp147		
M4/M5	13 in 96	14.0	dp129	dp109	dp7105	i62
					dp7111	i62
					dp7116	i62
					dp7119	i62
					dp7123	i62
					dp7128	i62
					dp7133	i62
					dp7136	i62
					dp7143	i62
					dp7145	i62
					dp7147	i62
					dp7180	i68
					dp7190	i68
M5/M6	5 in 24	20.8	dp152	dp151	dp9285	i94
					dp9286	i94
					dp9287	i94
					dp9288	i94
					dp9289	i94
M6/M7	n.d.	n.d. (3)				
M7/M8	--	-- (4)	dp84	dp85		
M8/M9	1 in 60	1.7	dp154	dp41	dp9387	i95
Total	25 in 252					

(1) No significant match due to incorrect dp143 sequence: AAGAGCATGAGGATGGCTAC. This primer should be: AAGAGCATGAGGATGCCATG.
(2) Not determined, because PCR did not yield a specific product.
(3) Not determined.
(4) No specific product, because primer dp85 sequence extends beyond module 8 and into the flanking region.

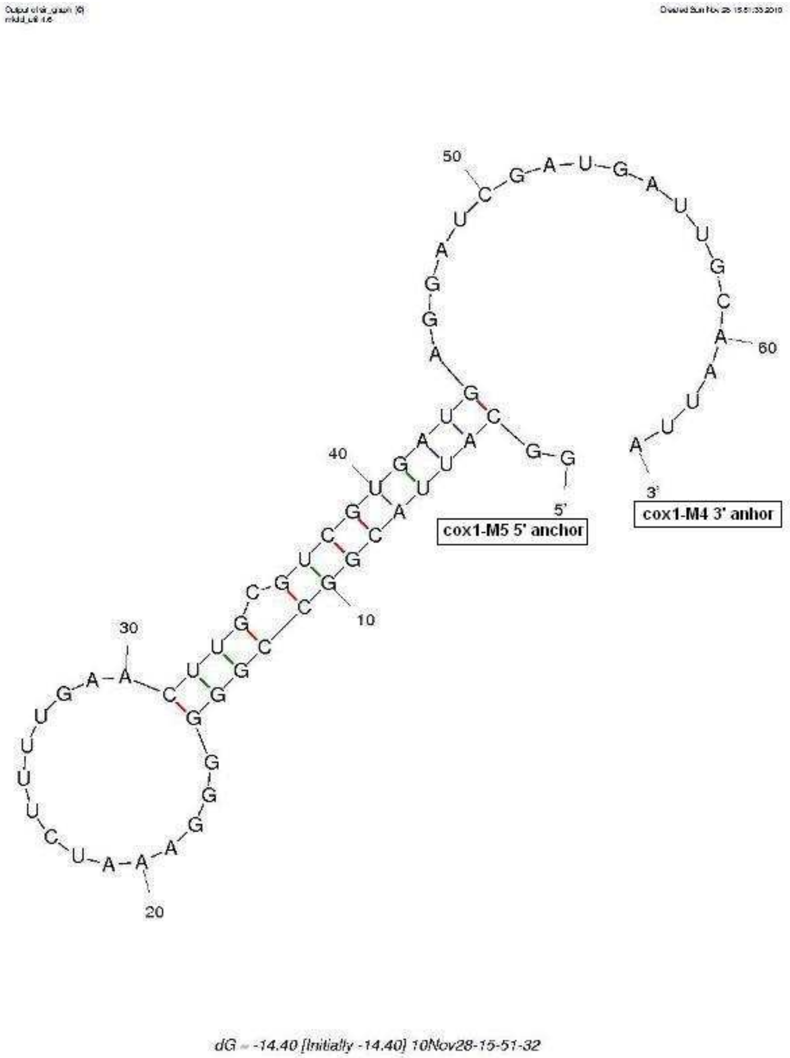
Appendix 4. Summary of unknown sequences from RT-RCR on circularized total RNA targeting gRNA

i80					
M2M3		M7M8	>50nt		<50nt
1	dp8013iai80	1	dp8049iai80	1	dp8073iai80
2	dp8014iai80	2	dp8051iai80	2	dp8074iai80
3	dp8017iai80	3	dp8053iai80	3	dp8075iai80
4	dp8018iai80	4	dp8060iai80	4	dp8077iai80
5	dp8021iai80	5	dp8062iai80	5	dp8078iai80
6	dp8022iai80	6	dp8063iai80	6	dp8079iai80
7	dp8043iai80	7	dp8064iai80	7	dp8081iai80
8	dp8044iai80	8	dp8065iai80	8	dp8082iai80
9	dp8045iai80	9	dp8068iai80	9	dp8083iai80
		10	dp8069iai80	10	dp8084iai80
		11	dp8070iai80	11	dp8085iai80
				12	dp8086iai80
				13	dp8087iai80
				14	dp8089iai80
				15	dp8091iai80
				16	dp8094iai80
i81					
M1/M2		M3/M4		M5/M6	
1	dp8144iai81	1	dp8161iai81	1	dp8188iai81
2	dp8145iai81	2	dp8162iai81	2	dpi8189iai81
3	dp8146iai81	3	dp8163iai81	3	dp8190iai81
4	dp8147iai81	4	dp8164iai81	4	dp8191iai81
5	dp8148iai81	5	dp8165iai81	5	dp8195iai81
		6	dp8166iai81	6	dp8196iai81
		7	dp8167iai81		
		8	dp8168iai81		11 distinct unknowns
		9	dp8169iai81		
		10	dp8170iai81		
		11	dp8171iai81		
		12	dp8172iai81		

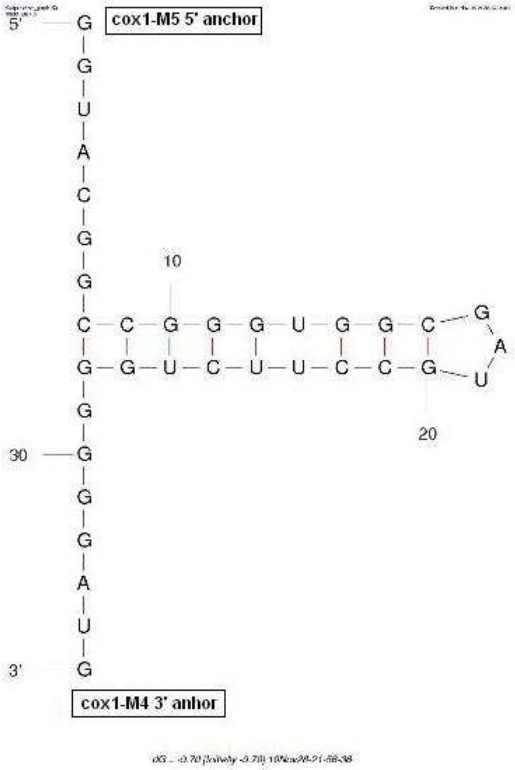
Appendix 5. Outputs of Mfold program for calculation of the secondary structures of selected clones

Secondary structures are calculated for sequences identified in the RT-PCR using a specific primers in M4 (dp129) and in M5 (dp109). A single-stranded stretch of six As or Gs are considered to be potential guiding nucleotides by pairing with six Us. The following sequences are calculated for their secondary structures where potential guiding nucleotides are identified. (A) clone dp6912; (B) clone dp6933; (C) clone dp6932; (D) clone dp7139. In (A) and (B), one stem-loop interrupts the guiding nucleotides, and in (C) and (D), the guiding nucleotides are spaced by two stem-loop structures. Boxed text indicates the locateions of the primer sequences and where they might bind to the modules.

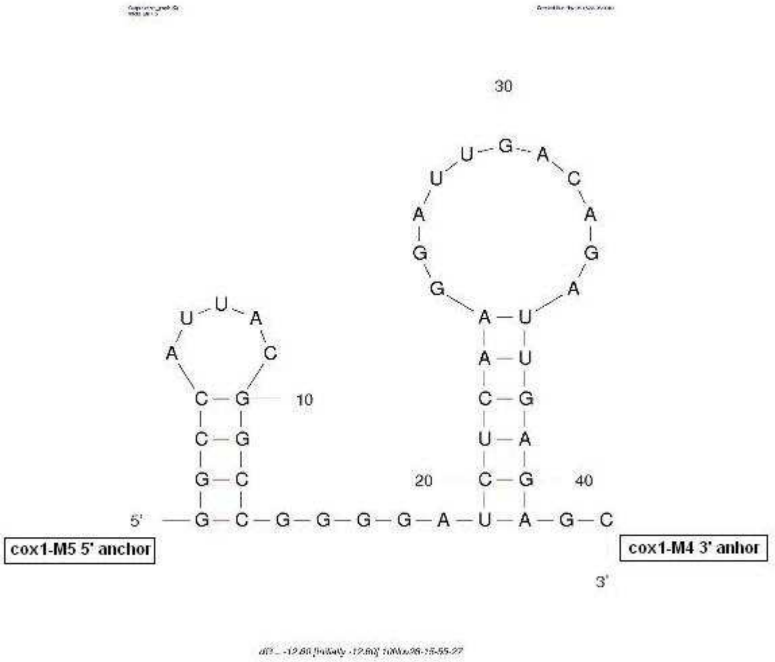
A. dp6912



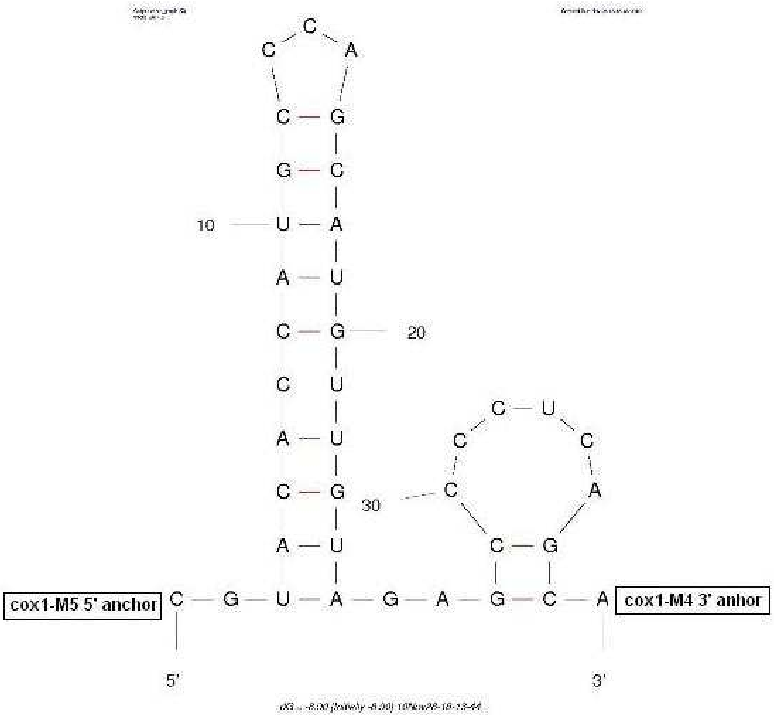
B. dp6933



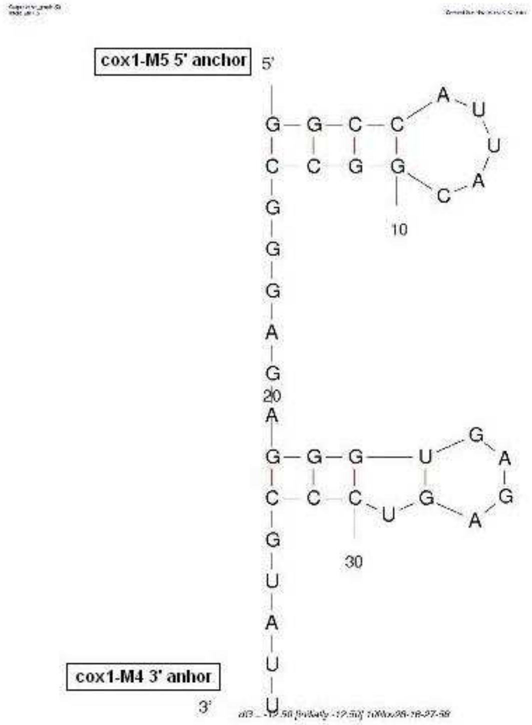
C. dp6932



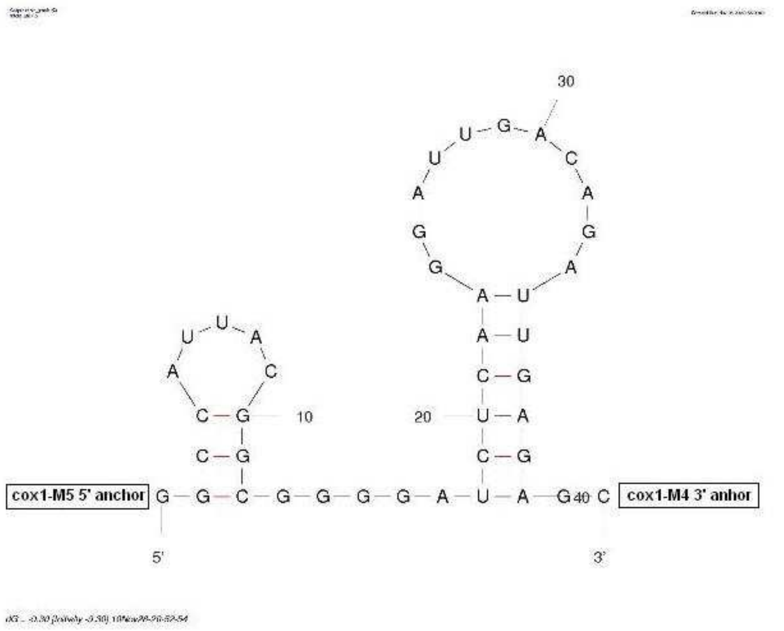
D. dp7139



E. dp6917



F. dp6948


cox1-M5 5' anchor
cox1-M4 3' anhor
5'
3'
10
20
30
40
dG = -3.30 [initially -3.30]

Appendix 6. Publication on Group I intron trans-splicing

TRENDS GENET. 2009 SEP;25(9):381-6. EPUB 2009 AUG 27.

Group I-intron trans-splicing and mRNA editing events in mitochondria of placozoan animals

Gertraud Burger, Yifei Yan, Pasha Javadi and B. Franz Lang