Université de Montréal

Processus post-transcriptionnels inédits dans la mitochondrie des diplonémides

par
Georgette N. Kiethega

Département de biochimie
Faculté de médecine

Thèse présentée à la Faculté de médecine
en vue de l'obtention du grade de *Philosophiae* Doctor (PhD)
en biochimie

mars 2013

Université de Montréal
Faculté des études supérieures et postdoctorales

Cette thèse intitulée

Processus post-transcriptionnels inédits dans la mitochondrie des diplonémides

Présentée par :
Georgette N. Kiethega

a été évaluée par un jury composé des personnes suivantes :

Dr Normand Brisson, président-rapporteur
Dr Gertraud Burger, directeur de recherche
Dr Gerardo Ferbeyre, membre du jury
Dr Linda Bonen, examinateur externe
Dr Marc Drolet, représentant du doyen de la Faculté
des études supérieures et postdoctorales

Résumé

Notre laboratoire a récemment découvert un mode d'expression des gènes mitochondriaux inédit chez le protozoaire biflagellé *Diplonema papillatum*. Outre son ADNmt formé de centaines de chromosomes circulaires, ses gènes sont fragmentés. Le gène *cox1* qui code pour la sous unité I de la cytochrome oxydase est formé de neuf modules portés par autant de chromosomes. L'ARNm de *cox1* est obtenu par épissage en *trans* et il est également édité par insertion de six uridines entre deux modules.

Notre projet de recherche a porté sur une étude globale des processus post-transcriptionnels du génome mitochondrial de diplonémides.

Nous avons caractérisé la fragmentation de *cox1* chez trois autres espèces appartenant aux deux genres du groupe de diplonémides à savoir : *Diplonema ambulator*, *Diplonema* sp. 2 et *Rhynchopus euleeides*. Le gène *cox1* est fragmenté en neuf modules chez tous ces diplonémides mais les modules sont portés par des chromosomes de taille et de séquences différentes d'une espèce à l'autre.

L'étude des différentes espèces a aussi montrée que l'édition par insertion de six uridines entre deux modules de l'ARNm de *cox1* est commune aux diplonémides. Ainsi, la fragmentation des gènes et l'édition des ARN sont des caractères communs aux diplonémides.

Une analyse des transcrits mitochondriaux de *D. papillatum* a permis de découvrir quatre autres gènes mitochondriaux édités, dont un code pour un ARN ribosomique. Donc, l'édition ne se limite pas aux ARNm.

De plus, nous avons montré qu'il n'y a pas de motifs d'introns de groupe I, de groupe II, de type ARNt ou d'introns impliqués dans le splicéosome et pouvant être à l'origine de l'épissage des modules de *cox1*. Aucune complémentarité significative de séquence n'existe entre les régions flanquantes de deux modules voisins, ni de résidus conservés au sein d'une espèce ou à travers les espèces. Nous avons donc conclu que l'épissage en *trans* de *cox1* chez les diplonémides fait intervenir un nouveau mécanisme impliquant des facteurs *trans* plutôt que *cis*.

L'épissage et l'édition de *cox1* sont dirigés probablement par des ARN guides, mais il est également possible que les facteurs *trans* soient des molécules protéiques ou d'ADN.

Nous avons élucidé les processus de maturation des transcrits mitochondriaux de *D. papillatum*. Tous les transcrits subissent trois étapes coordonnées et précises, notamment la maturation des deux extrémités, l'épissage, la polyadénylation du module 3' et dans certains cas l'édition. La maturation des extrémités 5' et 3' se fait parallèlement à l'épissage et donne lieu à trois types d'intermédiaires. Ainsi, un transcrit primaire avec une extrémité libre peut se lier à son voisin. Cet épissage se fait apparemment sans prioriser un certain ordre temporel alors que dans le cas des transcrits édités, l'édition précède l`épissage.

Ces études donnant une vue globale de la maturation des transcrits mitochondriaux ouvrent la voie à des analyses fonctionnelles sur l'épissage et l'édition chez *D. papillatum*. Elles sont le fondement pour finalement élucider les mécanismes moléculaires de ces deux processus post-transcriptionnels de régulation dans ce système intriguant.

Mots-clés : fragmentation des gènes, transcription, maturation des extrémités des ARN, épissage, édition par addition d'uridines, diplonémides, euglénozoaires, mitochondrie.

Abstract

Our laboratory has recently discovered an unprecedented mode of expression of mitochondrial genes in *D. papillatum*, a biflagellate protozoan. In addition to its mtDNA formed of hundreds of circular chromosomes, genes are fragmented. For example, the *cox1* gene which encodes the subunit one of the cytochrome oxidase complex, comprises nine modules carried by nine chromosomes. The *cox1* mRNA is obtained by trans-splicing and is also edited by the insertion of six uridines between two modules.

My thesis project focused on the study of post-transcriptional processes in diplonemid mitochondria. We characterized the fragmentation of *cox1* in three other species belonging to two diplonemids genera: *Diplonema ambulator*, *Diplonema* sp. 2 and *Rhynchopus euleeides*.

The *cox1* gene is fragmented into nine modules in all species but the modules are carried by chromosomes of different size and sequences from one species to another. We have shown that there are no motifs for classical introns, including spliceosomal and archaeal introns, as well as introns of group I and II, that might be implicated in the trans-splicing of *cox1* modules. No significant complementarity exists between the flanking regions of two neighboring modules, nor are any conserved residues within a species or across species. We therefore concluded that the trans-splicing of *cox1* in diplonemids involves a novel mechanism implicating *trans* rather than cis-factors. Trans-splicing and editing of *cox1* probably involve guide RNAs, but it is also possible that the trans-factors are proteins or DNA molecules.

The study of different species has also shown that the insertion of six uridines between two *cox1* modules in mRNA is a shared trait in these diplonemids. We discovered that four other mitochondrial genes are also edited in *D. papillatum* and that RNA editing is not limited to mRNA. So, fragmented genes and RNA editing are common characteristics of diplonemids.

We elucidated *D. papillatum*'s mitochondrial transcript maturation steps. All transcripts undergo three coordinated and precise processes including end processing,

trans-splicing and / or editing and polyadenylation. The processing of the 5 'and 3' ends gives rise to three kinds of maturation intermediates. A primary transcript with one free end can bind to its neighbor and trans-splicing occurs without directionality. In the case of edited transcripts, editing precedes trans-splicing.

These studies have prepared the ground for functional studies of trans-splicing and RNA editing with the long term goal to elucidate the molecular mechanisms involved in post-transcriptional regulation in this intriguing system.

Keywords: gene fragmentation, transcription, ends processing, trans-splicing, U-addition RNA editing, diplonemids, euglenozoa, mitochondria.

Table des matières

Liste des tableaux

Liste des figures

Liste des abréviations

A : Adénine

ADN : Acide Désoxyribo Nucléique

ADNc : ADN complémentaire

ADNdb : ADN double brin

ADNk : ADN du kinétoplaste

ADNmt : ADN mitochondrial

ADNr : ADN ribosomal

ARN : Acide RiboNucléique

ARNg : ARN guide

ARNi : ARN interférent

ARNm : ARN messager

ARNr : ARN ribosomal

ARNt : ARN de transfert

Asp : Acde aspartique

ATCC : American Type Culture Collection

miRNA : micro RNA

siRNA : small interfering RNA

snRNA : small nuclear RNA

AMV : Avian Myeloblastosis Virus

ATP : Adénosine Tri Phosphates

[^{32}P]-ATP : ATP marqué au phosphate 32

BET : Bromure d'Ethidium

BLAST : Basic Local Alignment Search Tools

C : Cytidine

CAM : Chloramphenicol

DEPC : Di Ethyl Pyro Carbonate

DSN : Duplex Specific Nuclease

dsRBD : double strand RNA Binding Domain

DYW : domaine c terminal de PPR constitue d'aspartate, tyrosine et tryptophane

E, E+ : domaine c terminal de PPR constitue de glutamate

EBS : Exon Binding Site

EDTA : Ethylene Diamine Tetraacetic Acid

FASTA : Functional Analysis System Technique Algorithm

GDE : Genetic Data Environment

Glu : Acide glutamique

h : Heure

HS : Heavy Strand

HSP : Heavy Strand Promoter

G : Guanine

I : Inosine

IBS : Intron Binding Site

IPTG : Iso Propyl-βeta -Thio Galactoside

Kb : Kilo base

L : Litre

LB : Liberia Bertani

LS : Light Strand

LSP : Light Strand Promoter

LSU : Large SubUnit

M : Molaire

MEB : Microscopie Electronique à Balayage

mg : milligramme

mM : millimolaire

min : minute

NADH : Nicotinamide Adénine Di nucléotide

NCBI : National Center for Biotechnology Information's

ng : nanogramme

nt : nucléotide

ORF : Open Reading Frame

PAP: Poly (A) Polymerase

pb: Paire de base

PCR : Polymerase Chain Reaction

PPR : PentatricoPeptide Repeat

POLRMT : Mitochondrial DNA direct RNA Polymerase

RBP : RNA Binding Protein

RBS : RNA Binding Site

RCF : Relative Centrifugal Force

RNA-Seq : Séquençage haut débit d'ARN

RRM : RNA Recognition Motif

RT-PCR : Reverse Transcriptase-Polymerase Chain Reaction

S : Svedberg

SDS : Sodium Dodecyl Sulfate

SSC : Saline Sodium Citrate

SSU : Small SubUnit

STE : Sodium chlorid Tris EDTA

T : Thymidine

TAP : Tobacco Acid Pyrophosphate

T4PNK : T4 Poly Nucleotid Kinase

TCA : Tri Carboxylic Acid

TET : Tetracycline

TFAM : Mitochondrial Transcription Factor A

TFBM : Mitochondrial Transcription Factor B

TUTase : Terminal Uridyl Transferase

Trp : Tryptophane

Tyr : Tyrosine

U : Uridine

μg : microgramme

UTP : Uridine Tri Phosphate

UTR : UnTranslated Region

UV : Ultra Violet

Dédicace

Ninsaal noor yaa sugr yandre. A san fooge a lebsg lebga toogo.

Je dédie cette thèse à la nouvelle génération qui s’annonce.

Remerciements

Je suis très reconnaissante envers Gertraud Burger, ma directrice de thèse pour m'avoir accepté dans son laboratoire et permis de mener à bien mon projet de thèse. Je tiens à lui dire merci pour sa disponibilité, sa patience et sa rigueur dans mon encadrement. Merci de m'avoir soutenue moralement, financièrement surtout pendant ma rédaction et de m'avoir permis de participer à des congrès scientifiques.

Je remercie le Dr Marcel Turcotte pour sa collaboration dans mon projet de thèse.

Je remercie Shona Teijeiro pour avoir grandement facilité ma familiarisation à mon nouveau laboratoire. Merci de m'avoir encadrée à la paillasse au début de ma thèse, pour tes réflexions, critiques et suggestions et tes corrections. Merci pour tes qualités humaines.

Un grand merci aux stagiaires et étudiants qui ont travaillé avec moi sur ce projet : Alexandre, Christel, Tuana, Pavel et Yifei.

Je remercie les Drs Matus Valach avec qui j'ai travaillé sur le projet RNA-Seq et Sophie Breton qui a accepté de corriger mon manuscrit.

Je remercie tous les membres passés et présents des laboratoires du Dr Burger et Dr Lang, et tous ceux que j'ai côtoyés durant ma thèse (collègues, stagiaires, enseignants et employés) au département de biochimie.

Mes remerciements à tous ceux que j'ai côtoyés au laboratoire d'enseignement du département de biochimie et spécialement aux étudiants qui m'ont amenés à me dépasser tant sur le plan personnel que professionnel.

Je tiens à remercier les membres de mon comité de thèse les Drs Serguei Chteinberg, Muriel Aubry et Georges Szatmari pour leurs conseils avisés.

Je tiens également à remercier l'ensemble des membres du jury pour avoir accepté de juger ce travail.

Je remercie sincèrement le Pr Gerard Kientega, Pr Jacques Simpore, Pr Francis Fumoux et Dr Daniel Parzy pour m'avoir encouragé à faire la thèse.

Ma reconnaissance va à l'endroit de tous ceux qui m'ont soutenue dans cette aventure. Je pense à mes parents Solange et Jean Baptiste.

Une pensée spéciale pour ma très chère mère qui m'a toujours soutenue. Tu m'as souvent répétée cette phrase « on ne met pas son enfant au monde pour qu'il réalise nos rêves, mais pour l'aider à réaliser les siens ». Les mots me manquent pour t'exprimer ma reconnaissance. Merci de m'avoir transférée la vie et m'avoir inculquée la persévérance.

Je voudrais exprimer ma reconnaissance envers des frères, sœurs et amis sans qui cette aventure n'aurait pas été possible : Prosper, Jean François, Marie Andrea, Honoré, Claver, Olivia, Gertrude, Jocelyne, Geneviève, Sara, Nadine, Prisca, Flavien, et Bertrand. Merci d'être restés en contact avec moi malgré la distance.

Une pensée pour les deux ainées chez qui j'ai fait du bénévolat ces deux dernières années.

Ces soirées passées à vous écouter me raconter vos expériences de vie m'ont beaucoup enrichie.

Un gros merci à Shen, Jenny, Tamara, David et Jean de Capistran qui ont rendu la vie à Montréal plus agréable.

Finalement cette aventure n'aurait pas été possible sans la bourse d'excellence du Programme Canadien de Bourses de la Francophonie (P.C.B.F).

Chapitre 1. Introduction

Cette thèse a pour but la caractérisation des processus post-transcriptionnels uniques dans la mitochondrie d'un groupe d'eucaryotes peu connus, les diplonémides. Notre travail porte spécifiquement sur l`épissage et l'édition des transcrits des gènes fragmentés dans la mitochondrie des diplonémides. Pour circonscrire le contexte du projet, le chapitre I introduit la mitochondrie, organelle caractéristique des eucaryotes, de même que les grands groupes eucaryotes. Ensuite sont présentés les génomes mitochondriaux et les mécanismes d'expression du génome mitochondrial connus à l'heure actuelle.

1 Origine et rôle de la mitochondrie

La mitochondrie, usine énergétique de la cellule eucaryote, a pour origine une α-protéobactérie endosymbiotique (Gray, 1999). Durant la transition de l'endosymbiote en organite, la plupart des gènes de l'endosymbiote ont été perdus ou transférés au noyau de la cellule hôte (Adams & Palmer, 2003, Burger*, et al.*, 2003). Plusieurs phylogénies basées sur des ARN ribosomiques et des protéines mitochondriales démontrent ce lien entre la mitochondrie et les α- protéobactéries (Gray, 1999, Lang*, et al.*, 1999).

La mitochondrie héberge entre autre la phosphorylation oxydative, la biosynthèse impliquant le fer, le soufre, le cycle de l'acide citrique, le cycle de l'urée et la synthèse de l'hème, la rendant essentielle aux cellules eucaryotes (McBride*, et al.*, 2006). Elle est aussi impliquée dans l'apoptose, le vieillissement et dans certaines maladies (Graeber & Muller, 1998, Kroemer*, et al.*, 1998, Wei, 1998) (Figure 1).

Tous les eucaryotes possèdent une mitochondrie ou un organite dérivé de celle-ci. Parmi les eucaryotes vivants dans des milieux anaérobies, certains sont munis d'hydogénosomes et d'autres de mitosomes qui sont des reliques de mitochondries (Burger*, et al.*, 2003, Embley*, et al.*, 2003).

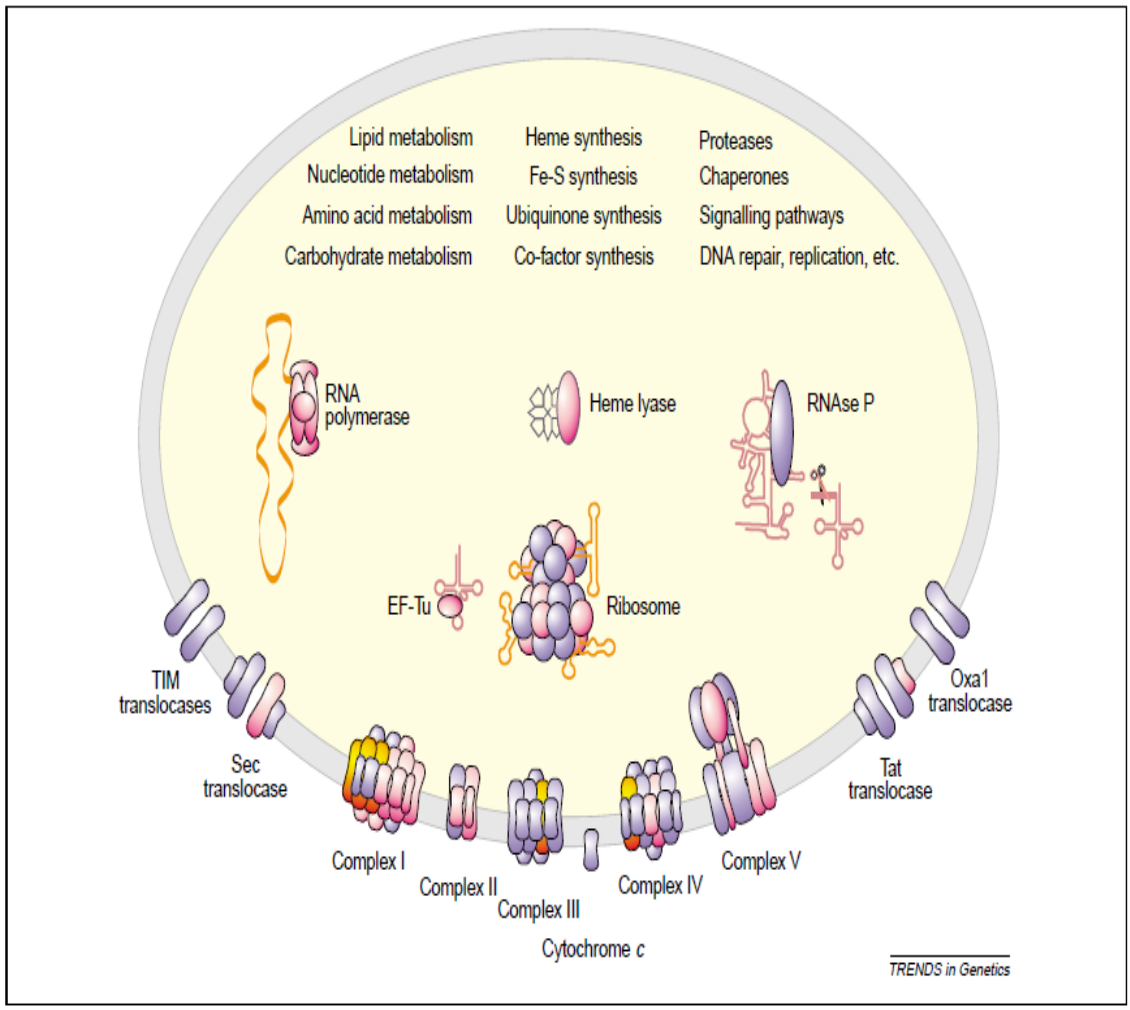


Figure 1. Implication de la mitochondrie dans des processus cellulaires (Burger*, et al.*, 2003). Les composantes mitochondriales encodées par le noyau sont en bleue. Celles qui sont encodées de facon invariable par l'ADNmt ou par quelques organismes sont respectivement en orange et rose.

2 Les grands groupes eucaryotes

Les eucaryotes sont classés généralement en cinq ou six grands groupes : *Les Archaeplastidae*, les Unikonts (*Opistokonta* et *Amoebozoa*), les *Rhizaria*, les *Chromalveolata* et les *Excavata* (Keeling*, et al.*, 2005, Parfrey*, et al.*, 2006) (Figure 2).

Les *Archaeplastidae* sont composés par les algues vertes, les plantes terrestres, les rhodophytes et les glaucophytes (Cavalier-Smith, 1998, Rodriguez-Ezpeleta*, et al.*, 2005). Le représentant le plus étudié est sans conteste *Arabidopsis thaliana* une angiosperme.

Les *Opisthokonta* regroupent l'homme, les animaux et champignons et leurs ancêtres unicellulaires. Ce groupe a été établi sur des bases moléculaires (Sogin*, et al.*, 1996). Ils forment avec les *Amoebozoa*, (myxomycètes) les Unikonts. Les myxomycètes incluent des organismes dont le cycle de vie passe par deux stades (plasmodium et spores flagellés) tel que *Physarum polycephalum*.

Le groupe des *Chromalveolata* a été contesté récemment et scindé en quatre sous groupes (les Alvéolates, les Stramenopiles, les Haptophytes et les Cryptophytes) et les eucaryotes ont été regroupés en cinq supers groupes : les *Amoebozoa,* les *Opistokonta,* les *Excavata,* les *Archaeplastidae* et les SAR désignant les Stramenopiles, les Alvéolates et les *Rhizaria* (Baurain*, et al.*, 2010, Adl*, et al.*, 2012). Le plus grand groupe, les Alvéolates, sont composés d'organismes photosynthétiques et non photosynthétiques : les apicomplexés, les dinoflagellés et les ciliés. Les relations entre les membres ont été établies à partir de phylogénies de gènes nucléaires (Keeling*, et al.*, 2005).

Les *Rhizaria* regroupent en leur sein des organismes qui ont une grande importance écologique tels que les cercomonades et les foraminifères. Ce groupe n'a été établi que récemment, en se basant sur des caractères moléculaires (Cavalier-Smith, 2002, Burki & Pawlowski, 2006).

Les *Excavata* sont un groupe monophylétique de protistes formé sur la base de caractères morphologiques et moléculaires (Simpson*, et al.*, 2006, Hampl*, et al.*, 2009).

On y trouve les jakobides, diplomonades et parabasalides ainsi que les euglénozoaires, notre groupe d'intérêt.

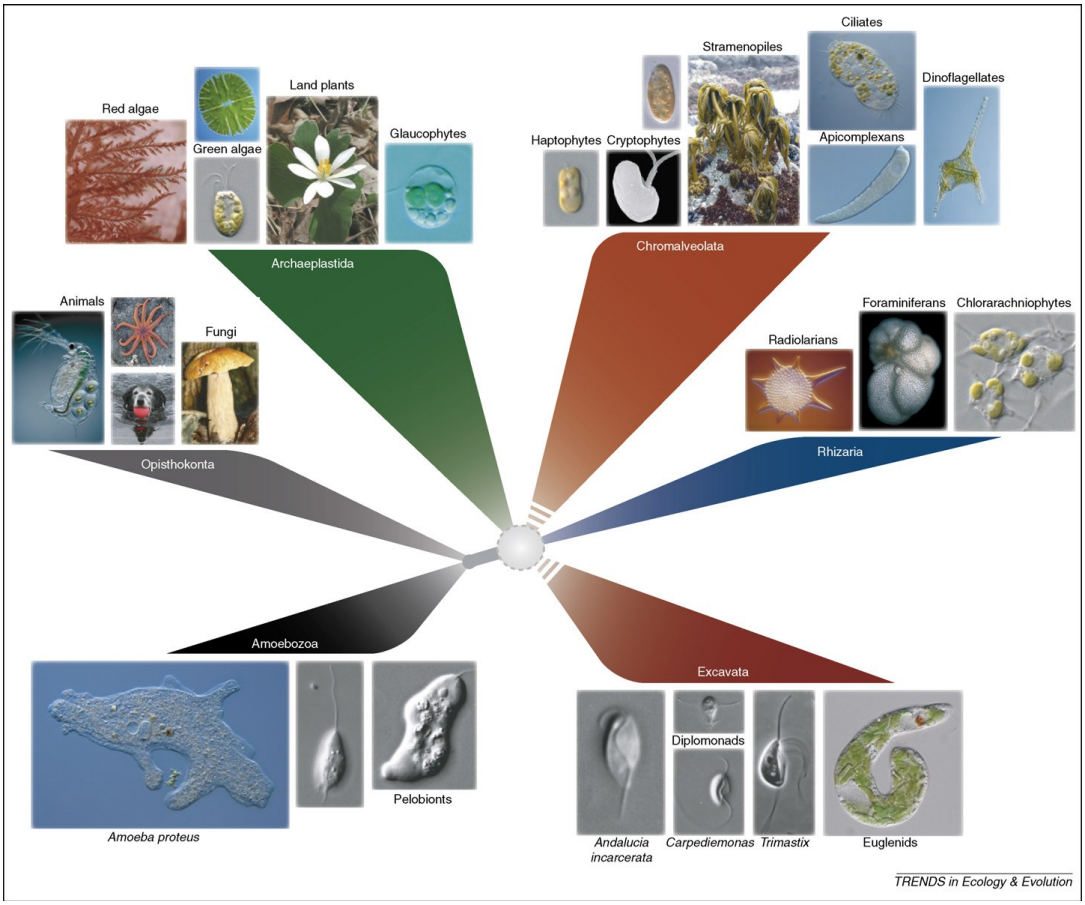


Figure 2. Les grands groupes eucaryotes (*Archaeplastida*, *Chromalveolata*, *Opisthokonta*, *Rhizaria*, *Amoebozoa* et *Excavata*). Cette figure est tirée de (Parfrey*, et al.*, 2006). A noter que le groupe *Chromalveolata* a été scindé en quatre groupes à appartenance inconnue : les Alvéolates, les Stramenopiles, les Haptophytes et les Cryptophytes.

3 Les euglénozoaires

Ce phylum, est formé de trois groupes de flagellés unicellulaires : les kinétoplastides, les diplonémides et les euglénides (Simpson, 1997, Cavalier-Smith, 1998, Simpson*, et al.*, 2002) (Figure 3). Dans ce groupe monophylétique (Maslov*, et al.*, 1999, Maslov*, et al.*, 2001, Preisfeld*, et al.*, 2001), la phylogénie moléculaire montre les euglénides à la base, les kinétoplastides et les diplonémides étant des groupes frères (Simpson & Roger, 2004).

Les diplonémides sont des protozoaires biflagellés des eaux marines, qui ont été précédemment décrites comme appartenant au genre *Isomena*. Certaines espèces de ce groupe infectent les homards, les palourdes, les diatomées et les *Cryptocorynae* (plantes aquatiques) (Kent*, et al.*, 1987). Les diplonémides sont constitués de deux genres: *Diplonema* et *Rhynchopus.* Le genre *Diplonema* décrit pour la première fois en 1913, est caractérisé par deux courts flagelles insérés à la base de la cellule. Plusieurs espèces ont été identifiées dont *D. breviciliata*, *D. metabolicum*, *D. negricans, D. ambulator, D. papillatum et D.* sp. 2. Les trois dernières espèces constituent avec *Rhynchopus euleeides*, les sujets de notre étude (Kent*, et al.*, 1987, Triemer & Ott, 1990, Roy*, et al.*, 2007) (Figure 4).

Le genre *Rhynchopus* est caractérisé par deux flagelles discrets qui peuvent se déployer ou non selon certaines conditions liées à la disponibilité de nutriments (Roy*, et al.*, 2007). Plusieurs espèces ont également été décrites dans ce genre: *R. amitus*, *R. coscino-discovorus, R.* sp. 2 et *R. euleeides*. La dernière a été formellement décrite par un membre de notre groupe (Roy*, et al.*, 2007).

Les kinétoplastides, le groupe frère des diplonémides, sont formés par deux ordres: les *Bodonida* et les *Trypanosomatida* (Simpson*, et al.*, 2002). Au sein des trypanosomatides, on retrouve des pathogènes de l'homme et des animaux dont les plus notoires sont *Trypanosoma brucei* responsable de la maladie du sommeil et

Trypanosoma cruzi responsable de la maladie de Chaggas. Ce qui fait des kinétoplastides le groupe le plus étudié au sein des euglénozoaires (Lukes*, et al.*, 2005).

Le dernier groupe de protistes faisant partie des euglénozoaires est formé par les euglénides lesquels sont constitués d'organismes photosynthétiques (*Euglena*) et non photosynthétiques (*Peranema*) (Leander*, et al.*, 2007).

Comme décrite dans le paragraphe 4, la structure du génome mitochondrial des euglénides est la moins connue des trois groupes d'euglénozoaires.

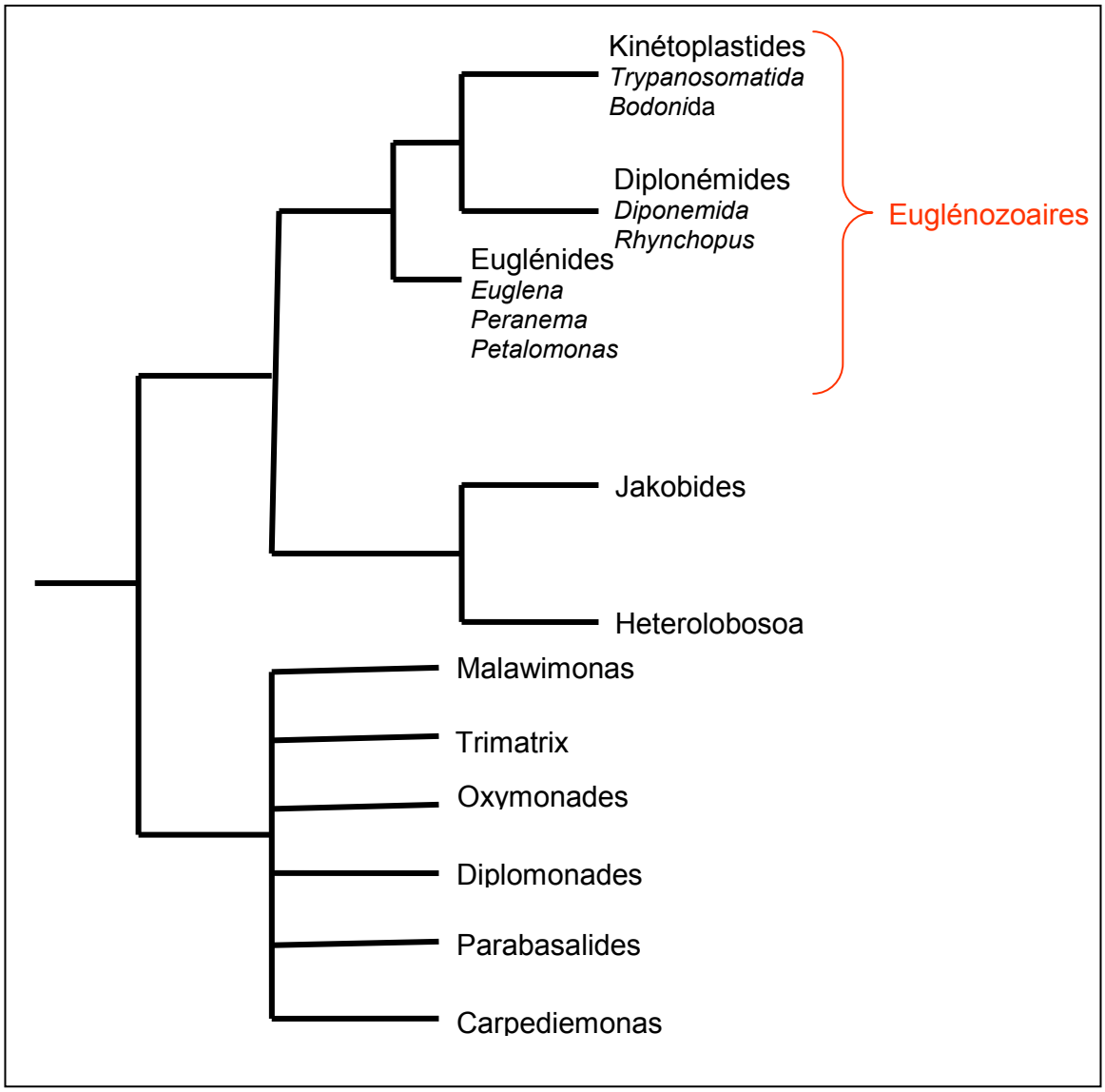


Figure 3. Phylogénie des Excavates. Les euglénozoaires (en rouge) basée sur (Simpson & Roger, 2004). Les différents sous groupes sont en noirs et les genres en italique. Chez les euglénides trois genres sont mentionnés.

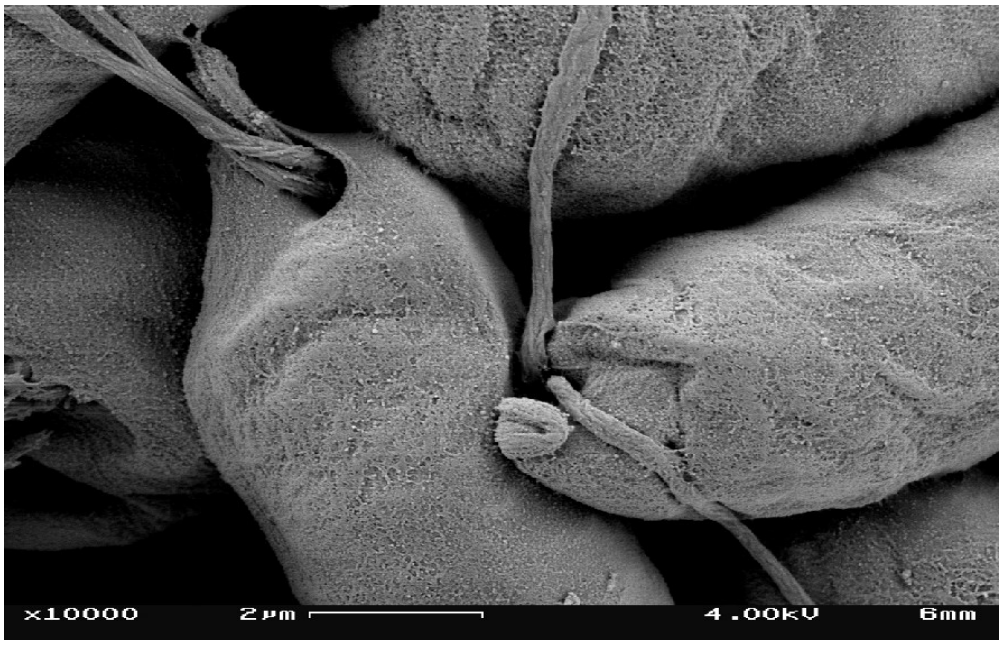


Figure 4. Image de *D. papillatum* obtenue par MEB (Microscopie Electronique à Balayage) (Burger*, et al.*, 2011). Grossissement x 10000. On voit les deux courts flagelles sub-apicaux et la poche d'ingestion.

4 Le génome mitochondrial

L'avancée de la génomique a donné lieu à une accumulation de données sur les génomes mitochondriaux (www.ncbi.nlm.nih.gov/genomes) et a mis en évidence une diversité inattendue des ADNmt concernant leur taille, leur architecture et leur contenu en gènes, ainsi que leur organisation et leur expression. Une brève revue de ces aspects sera effectuée dans les paragraphes suivants.

4.1Taille, architecture et organisation du génome mitochondrial

L'architecture des génomes mitochondriaux consiste en général en un chromosome circulaire chez les métazoaires, les champignons et les plantes; en plusieurs copies (Bendich, 1993, Nosek & Tomaska, 2003). Mais on trouve également des formes linéaires monomériques ou concatenées (Burger, *et al.*, 2003). La taille du génome mitochondrial varie de 6 kilobases (kb) chez *Plasmodium* à 11.3 Mb chez la plante *Silene conica* (Ward, *et al.*, 1981, Feagin, *et al.*, 1991, Alverson, *et al.*, 2010, Sloan, *et al.*, 2012). La plupart des eucaryotes ne possèdent qu'un seul type de chromosome ; chez l'homme par exemple il s'agit d'un chromosome circulaire de 16.5 kb. Il existe aussi des organismes avec plusieurs types de chromosomes mitochondriaux comme certaines algues vertes et animaux (Burger, *et al.*, 2003).

Les cas d'organisation de l'ADNmt les plus exceptionnels se retrouvent chez les euglénozoaires. Les diplonémides témoignent de cette plasticité inouïe des génomes mitochondriaux. En effet, les travaux de notre laboratoire ont démontré que l'ADNmt de *D. papillatum* est formé de centaines de chromosomes circulaires de deux classes de tailles: A et B, de 6 et 7 kb respectivement (Marande, *et al.*, 2005).

L'organisation de l'ADNmt chez les kinétoplastides, le groupe frère des diplonémides, est plus variable. Ainsi, on trouve de l'ADNmt constitué de cercles dispersés ou concaténés chez les bodonides (Hajduk, *et al.*, 1986, Lukescaron, *et al.*, 1998); chez *Cryptobia helicis*, l'ADNmt appelé pankinétoplaste est formé de mini-cercles de 4.2 kb et de maxi-cercles de 43 kb non concaténés (Lukescaron, *et al.*, 1998);

chez les trypanosomatides, le génome mitochondrial ou kinétoplaste est formé par un réseau compact de milliers de mini-cercles et de maxi-cercles (Chen*, et al.*, 1995, Liu*, et al.*, 2005) . Les mini-cercles de 1 kb chez *Trypanosoma brucei* encodent au moins 1200 ARNg (Shapiro*, et al.*, 1999, Stuart*, et al.*, 2005) et les maxi-cercles codent pour les gènes mitochondriaux. L'eukinétoplaste décrit chez *Crithidia fasciculata* est un réseau lâche formé de 5000 mini-cercles et 25 maxi-cercles (Shapiro*, et al.*, 1999, Lukes*, et al.*, 2002) (Figure 5).

En ce qui concerne le génome mitochondrial des euglénides, les premières descriptions datent des années 1970 et rapportent des molécules linéaires de 1-70 kb chez *Euglena gracilis* (Manning*, et al.*, 1971, Yasuhira & Simpson, 1997). Notre groupe a décrit un ADNmt circulaire de 40 kb chez *Petalomonas cantuscigny,* un membre des euglénides (Roy*, et al.*, 2007). Récemment, il a été rapporté que le génome mitochondrial d'*Euglena gracilis* est formé des molécules linéaires de 4 kb et 7.5 kb (Spencer & Gray, 2011).

4.2 Les gènes mitochondriaux

À travers les eucaryotes, le génome mitochondrial ne contient qu'entre 5 et 100 gènes, la majorité des protéines mitochondriales (1000 à 1500), étant encodées dans le noyau et ensuite importées dans la mitochondrie (Neupert & Herrmann, 2007, Bolender*, et al.*, 2008, Meisinger*, et al.*, 2008, Chacinska*, et al.*, 2009, Schmidt*, et al.*, 2010).

Le génome mitochondrial code pour des protéines de la chaine respiratoire et impliquées dans la phosphorylation oxydative: *cox* (cytochrome oxidase), *cob* (apocytochrome b), *nad* (NADH déshydrogénase), *atp* (ATP synthase) et *sdh* (succinate déshydrogénase) (Tableau I).

Le génome mitochondrial code aussi invariablement pour des ARN du mito-ribosome, la SSU-rRNA (« Small SubUnit ribosomal RNA ») et la LSU-rRNA (« Large subnit ribosomal RNA »), associées respectivement à la petite et à la grande sous-unité du ribosome; généralement pour des ARN de transfert (ARNt) (il est à noter que tous les génomes mitochondriaux ne codent pas pour des ARNt) (Burger*, et al.*, 2003). Des

protéines ribosomiques et l'ARNr 5S peuvent aussi être encodés par l'ADNmt (Gray*, et al.*, 2004).

Les gènes les plus rarement encodés par la mitochondrie sont ceux qui spécifient les protéines impliquées dans l'importation et la maturation des protéines et des ARN.

Des gènes mitochondriaux impliqués dans la transcription ont seulement été décrits chez les jakobides (Gray*, et al.*, 2004, Burger*, et al.*, 2011). En effet, le plus grand nombre de gènes mitochondriaux se trouve chez les jakobides avec 31 gènes d'ARN structuraux et 69 gènes protéiques (Burger & Nedelcu, 2011). Le Tableau I donne un aperçu du contenu en gènes chez les organismes eucaryotes en général et les euglénozoaires en particulier. Chez les euglénides, le contenu des génomes mitochondriaux n'est que très partiellement connu : On a des gènes codant pour des protéines (*cox1*, *cox2*, *cox3)* et pour des ARNr chez *E. gracilis* alors que chez *Diplonema*, la plupart des gènes ont été identifiés ces dix dernières années (Maslov*, et al.*, 1999, Vlcek*, et al.*, 2011).

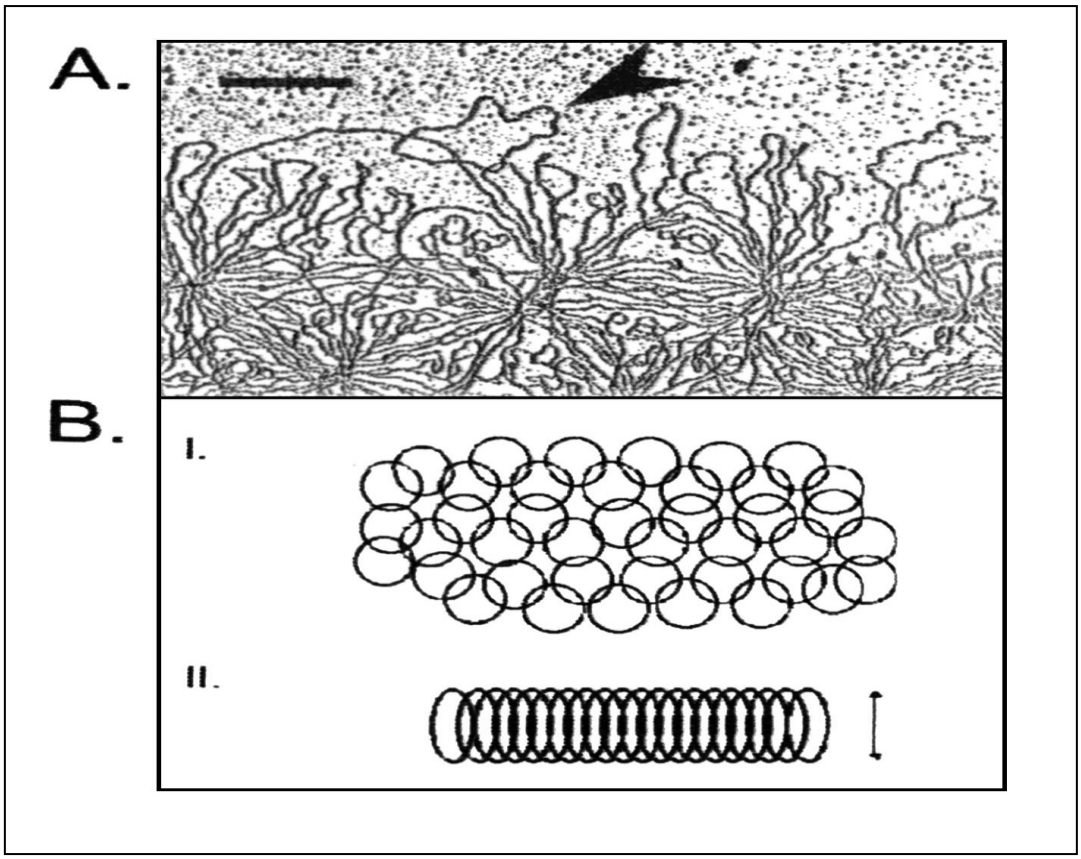


Figure 5**.** Structure de l'ADNmt de *Trypanosoma avium* (Lukes, et al., 2002). A. Image obtenue par microscopie électronique montrant des structures circulaires. La barre représente 500 nm; la flèche noire montre un mini-cercle. B. Organisation des chromosomes circulaires en un réseau formant le kinétoplaste.

Tableau I. Structures et fonctions de gènes mitochondriaux.

Groupes[c]	Structure de l'ADNmt	Gènes fragmentés	Complexes de la chaine respiratoire et composants ribosomiques
Animaux	circulaire, circular mapping	Non	I, III, IV, V, *rns, rnl, rps*, *rpl*
Plantes	linéaires, circulaires	Non	[a]I, II, III, IV, V, *rns, rnl, rrn5, rps*, *rpl*
Euglénozoaires	circulaires, linéaires, multiples	parfois	I, III, IV, V, *rns, rnl*, *rps*
Dinoflagellésés	linéaires, multiples	parfois	III, IV, *rns, rnl*
Myxomycètes	circulaires	Non	I, IV, V, *rns, rnl, rrn5, rps, rpl*
Jakobides[b]	circulaires	Non	I, II, III, IV, V, *rns, rnl, rrn5, rps, rpl*

[a]I, II, III, IV, V complexes NADH déshydrogénase, Succinate déshydrogénase, Cytochrome c oxidase, Cytochrome oxidase, ATP synthase; *rns*, *rnl*, *rrn5* sont les gènes de l'ARNr de la petite sous unité ribosomique, l'ARNr de la grande sous unité ribosomique et l'ARNr 5S; *rps*, *rpl* sont les protéines de la petite et grande sous unités du ribosome. [b]ADNmt code pour le gène de l'ARN polymérase. [c]Références : (Wolstenholme, 1992, Lang*, et al.*, 1997, Takano*, et al.*, 2001, Burger*, et al.*, 2003, Lukes*, et al.*, 2005, Marande*, et al.*, 2005, Jackson*, et al.*, 2007, Marande & Burger, 2007, Waller & Jackson, 2009, Burger*, et al.*, 2011, Burger & Nedelcu, 2011, Kayal*, et al.*, 2012, Lavrov*, et al.*, 2012, Smith*, et al.*, 2012).

5 Expression des gènes mitochondriaux

Tout comme la structure du génome mitochondrial, l'expression des gènes mitochondriaux est très diverse à travers les eucaryotes. On trouve en général une ARN polymérase de type phagique T3/T7 (Masters*, et al.*, 1987, Hedtke*, et al.*, 1997) encodée par le noyau sauf chez les jakobides où l'ADNmt encode une ARN polymérase de type bactérien (Lang*, et al.*, 1997). La transcription a été étudiée biochimiquement en détail chez les mammifères et les plantes mais demeure peu connue chez les autres eucaryotes.

5.1 La transcription mitochondriale

Chez les mammifères, l'initiation de la transcription des gènes mitochondriaux est décrite. Chez l'homme par exemple, l'ADN est transcrit en deux brins : LS (Light Strand) et HS (Heavy Strand) possédant respectivement un et deux promoteurs de la transcription (HSP1 et HSP2) (Asin-Cayuela & Gustafsson, 2007, Falkenberg*, et al.*, 2007). Le promoteur LSP (Light Strand Promoter) produit un seul transcrit tandis que le promoteur HSP2 (Heavy Strand Promoter 2) produit un précurseur polycistronique constitué de 12 ARNm et deux ARNr. Le promoteur HSP1 (Heavy Strand Promoter 1) produit lui deux ARNr. Tous les ARNm et ARNr sont flanqués par au moins un ARNt et ces derniers seront clivés dans le précurseur polycistronique par la suite, pour libérer les ARNm et ARNr. Une ARN polymérase, la POLRMT (« Mitochondrial DNA direct RNA polymerase ») effectue cette transcription. Chez l'homme, la POLRMT a besoin de trois facteurs auxiliaires: TFAM (« Mitochondrial Transcription Factor A »), TFB1M (« Mitochondrial Transcription Factor B1 ») ou TFB2M (« Mitochondrial Transcription Factor B2 ») (Bonawitz*, et al.*, 2006, Asin-Cayuela & Gustafsson, 2007, Shutt*, et al.*, 2010). Un facteur de transcription du type B existe également chez la levure *Saccharomyces cerevisiae* où il est désigné TFB ou Mtf1p (Schafer, 2005).

La machinerie de transcription mitochondriale chez les plantes a été également analysée (Liere*, et al.*, 2011). Elle implique deux ARN polymérases dont une spécifique

à la mitochondrie (RpoTm) et une autre (RpoTmp) qui est utilisée également par le plastide.

Contrairement aux organismes modèles appartenant aux animaux, champignons et plantes, la transcription des gènes mitochondriaux est peu étudiée chez les autres groupes eucaryotes (protistes). On sait qu'une polymérase de type T3/T7 y est impliquée (Lang*, et al.*, 1997, Shutt & Gray, 2006) mais il n'est pas exclu que plusieurs types de polymérases coexistent chez ces organismes (Barbrook*, et al.*, 2010). Chez les trypanosomes, de longs transcrits polycistroniques subissent une maturation pour donner des transcrits mono ou polycistroniques. Mais il faut noter que pour la transcription des ARN guides des trypanosomes, l'initiation se fait au niveau de sites situés sur leur chromosome (Lukes*, et al.*, 2005). La transcription mitochondriale peut être bidirectionnelle comme dans le noyau et le chloroplaste mais aussi dans les bactéries et virus (Beck & Warren, 1988). Ces dernières années, les nombreuses études ont montré que la plupart des promoteurs sont bidirectionnels (Wei*, et al.*, 2011).

5.2 Les étapes de maturation post-transcriptionnelle

Après la transcription, les ARN mitochondriaux passent par des étapes de maturation avant d'être traduits en protéines. Ces étapes incluent l'épissage des introns s'ils sont présents, occasionnellement l'édition des ARN et la maturation des extrémités 5' et 3'.

Elles constituent, avec la régulation traductionnelle, des processus de contrôle de l'expression des gènes mitochondriaux.

5.2.1 Épissage des introns

Deux types d'introns sont retrouvés dans les gènes mitochondriaux notamment ceux du groupe I et II. L'épissage des introns fait intervenir aussi bien la séquence *cis* que la structure secondaire des ARN concernés (Moreira*, et al.*, 2012). L'épissage se fait en *cis* lorsque la séquence intronique est entourée par des exons dans un seul précurseur

d'ARN. Dans le cas où le gène est fragmenté en plusieurs exons codés à différents endroits du génome, on a plusieurs précurseurs de transcrits qui doivent être liés par l'épissage en *trans*.

Des protéines sont également impliquées dans l'épissage. Ces dernières années les PPRs (« Penta trico Repeat Proteins »), protéines liant l'ARN, ont été identifiées comme jouant un rôle dans l'épissage des gènes d'organites chez les plantes (Schmitz-Linneweber & Small, 2008). Puisque le système d'expression étudié dans le contexte de cette thèse suggérait l'existence d'introns, nous présentons dans les paragraphes suivants un aperçu de types d'introns connus.

5.2.1.1 Les introns du splicéosome

Les introns du splicéosome sont trouvés exclusivement dans les gènes nucléaires des eucaryotes (Wachtel & Manley, 2009, Wahl*, et al.*, 2009, Valadkhan & Jaladat, 2010). On retrouve deux sous-types de ces introns (types U2 et U12) avec des splicéosomes distincts. L'intron de type U2 est le plus répandu et est caractérisé par les motifs GT en 5' et AG en 3'de l'intron tandis que l'intron de type U12 contient plutôt les di-nucléotides AT-AC.

5.2.1.2 Les introns des Archaea ou introns d'ARNt

Ces introns ont été retrouvés dans les gènes nucléaires eucaryotes et archaebactériens. La particularité de ce type d'introns réside dans son site d'excision caractérisé par le motif « Bulge-Helix-Bulge » (Figure 6). Aucun intron de ce type n'a été découvert à ce jour dans les organelles (Moreira*, et al.*, 2012).

5.2.1.3 Les introns de groupe I

Les introns de groupe I se trouvent majoritairement dans les organites mais aussi dans le noyau et chez certaines bactéries. Pour une revue récente voir (Nielsen & Johansen, 2009, Moreira*, et al.*, 2012). Ils ont une structure secondaire caractéristique et

essentielle à leur activité catalytique car cette structure spécifie le site d'épissage. La structure 2D d'un intron de groupe I est constituée de plusieurs épingles ou « stems loops » (segments P1-P10). Le site catalytique est formé par les domaines P3-P7-P8-P9 (Figure 7). L'épissage est réalisé par l'intron lui-même d'où le terme ribozyme sous lequel on les désigne. Les introns de groupe I peuvent inclure des ORFs (Open Reading Frame) encodant des endonucléases qui assistent les premiers dans leur mobilité ainsi que des maturases facilitant le repliement des introns. La structure typique, les ORFs introniques et le mécanisme d'épissage ont été décrits dans plusieurs revues (Stoddard, 2005, Lang*, et al.*, 2007, Edgell*, et al.*, 2011, Bonen, 2012, Moreira*, et al.*, 2012). Les introns du groupe I sont généralement impliqués dans l'épissage en *cis* mais récemment plusieurs auteurs ont rapporté l'existence d'introns de groupe I impliqués dans l`épissage en *trans* du gène mitochondrial *cox1* chez plusieurs organismes (Burger*, et al.*, 2009, Grewe*, et al.*, 2009, Pombert & Keeling, 2010, Hecht*, et al.*, 2011, Nadimi*, et al.*, 2012, Nishimura*, et al.*, 2012). L'ARNm de ce gène est obtenu après appariement des régions introniques de transcrits séparés en une structure type intron de groupe I.

5.2.1.4 Les introns de groupe II

La plupart des introns de groupe II connus se trouvent dans les organelles mais ils sont également retrouvés chez quelques bactéries et *Archaea.* Ils sont bien décrits dans les organelles des plantes et des champignons et pour quelques protistes et animaux (Bonen, 2008, Lambowitz & Zimmerly, 2011, Bonen, 2012). La structure secondaire des introns de groupe II est formée par six domaines hélicoïdaux (I-VI) connectés à un noyau central (Michel*, et al.*, 2009) (Figure 8). Le site catalytique de l'intron est formé par le domaine V qui est de ce fait utilisé pour l'identification in *silico* des introns de groupe II (Lang*, et al.*, 2007). Les introns de groupe II encodent généralement une endonucléase et une maturase mais parfois aussi une transcriptase inverse qui permet sa mobilité. L'épissage en *trans* impliquant des introns de groupe II a été beaucoup étudié dans la mitochondrie des plantes (Bonen, 2008, Bonen, 2012).

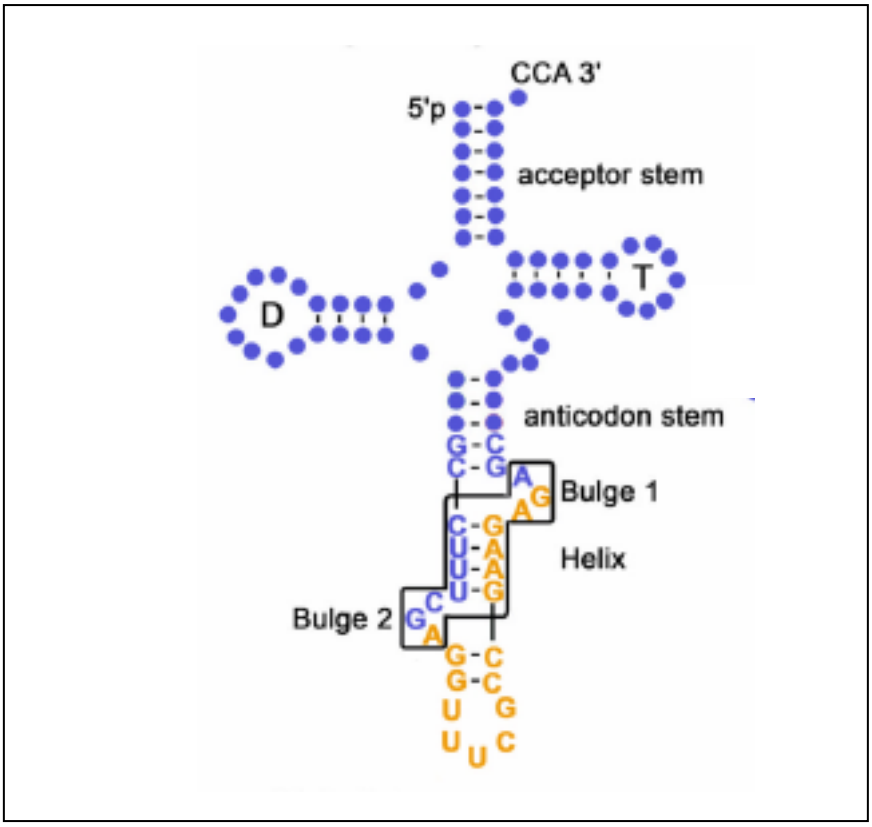


Figure 6. Structure secondaire de l'intron de l'ARNt-Glu chez *Archaeglobus fulgidus* (Heinemann*, et al.*, 2010). En bleu et orange les séquences exoniques et introniques, respectivement.

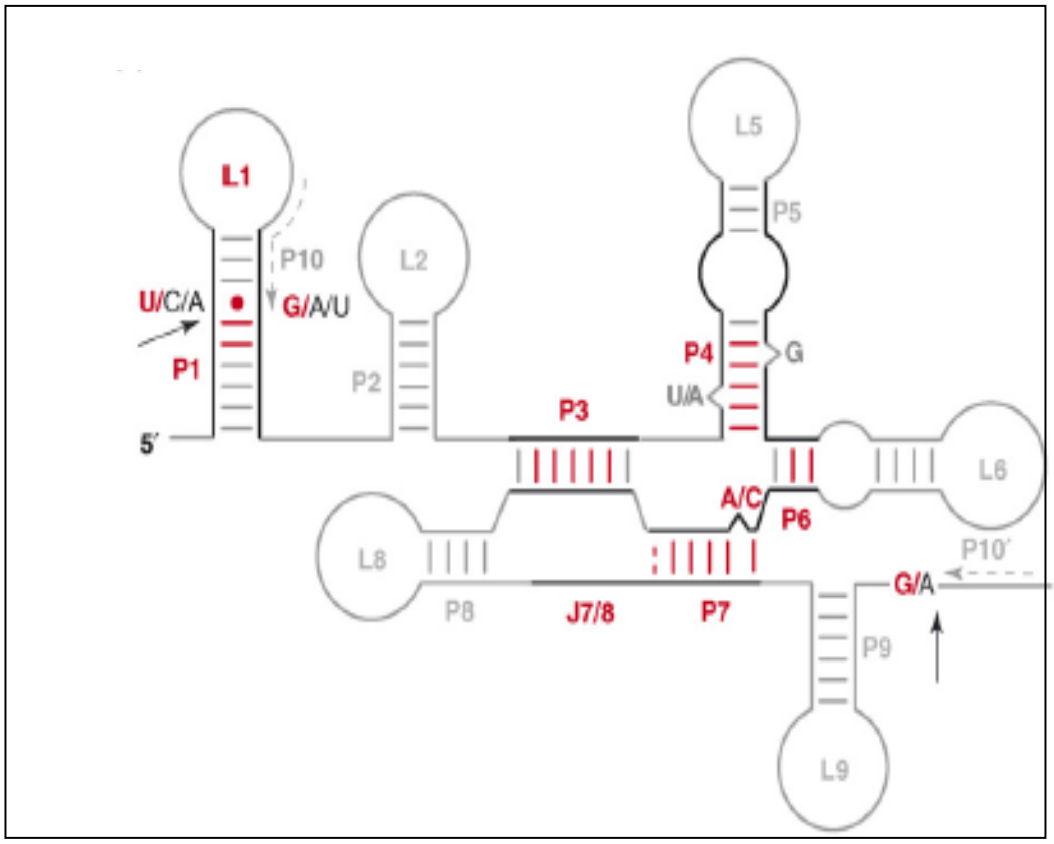


Figure 7. Structure secondaire consensus d'un intron de groupe I mitochondrial. On voit les segments appariés P1-P10, les séquences des boucles L (loop) et de jonctions entre domaines (J7/8) (Lang*, et al.*, 2007).

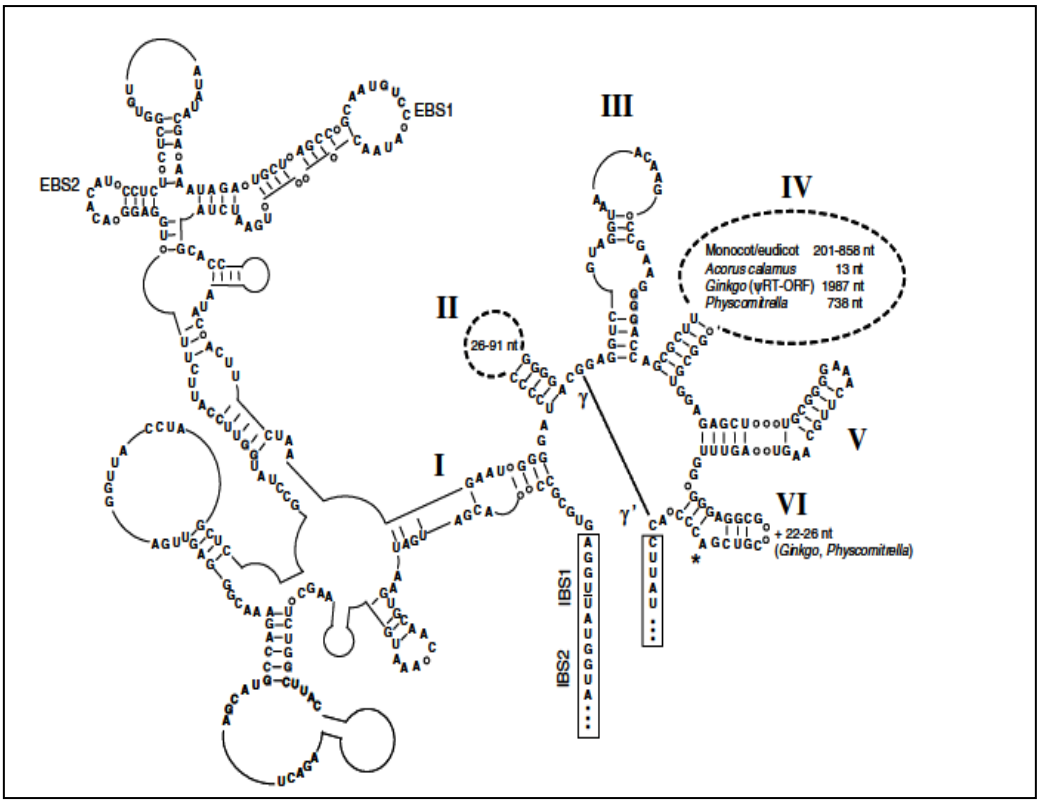


Figure 8. Structure secondaire d'un intron de groupe II mitochondrial chez les plantes. Les six domaines, les IBS (Intron Binding Site) et les EBS (Exon Binding Site) sont mis en évidence (Bonen, 2008).

5.2.2 L'édition des ARN

Le terme édition de l'ARN regroupe tous les processus co- ou post-transcriptionnels altérant la séquence d'un transcrit et produisant une séquence différente de celle codée par le génome (Price & Gray, 1998). Le terme exclut l'ajout d'une coiffe en 5' de l'ARNm, d'une queue poly (A) à l'extrémité 3' et du CCA au 3' de l'ARNt. L'édition consiste en l'ajout, la suppression ou la substitution d'un ou de plusieurs nucléotides. L'édition peut avoir lieu dans le noyau ou les organelles et peut affecter les transcrits messagers et ceux des ARN structuraux, les ARN de certains virus et les micros ARN (Kolakofsky*, et al.*, 2005, Nishikura, 2010, Knoop, 2011). Elle a été décrite pour la première fois par l'équipe de Rob Benne (Benne*, et al.*, 1986) dans le gène *cox2* de *Trypanosoma brucei* et il s'est avéré par la suite qu'elle était très répandue chez les eucaryotes et certains virus. Aujourd'hui, on connait plusieurs types d'édition d'ARN qui diffèrent par rapport aux mécanismes moléculaires et aux enzymes impliqués. Nous décrirons ces différents types dans les paragraphes suivants.

5.2.2.1 Édition par conversion de Cytidine en Uridine et Adénine en Inosine des transcrits nucléaires des animaux

L'édition des transcrits des gènes nucléaires a été étudiée chez les mammifères. Elle existe sous deux formes : conversion d'une Cytidine en Uridine (C→U) et conversion d'une Adénine en Inosine (A→I).

La conversion de C→U affecte principalement les transcrits du gène apoB, qui est impliqué dans le métabolisme des lipides chez l'homme. Le transcrit de ce gène est édité par conversion du C en position 6666 en U, changeant de ce fait le codon CAA en un codon stop (UAA). Il en résulte la formation d'un ARNm traduit en une protéine plus courte, l'apo48, la forme normale non éditée étant l'apo100. Le choix de la Cytidine à éditer est très spécifique et dépend aussi bien de la structure primaire que secondaire de l'ARN (Blanc & Davidson, 2010) (Figure 9). La machinerie de l'édition est un grand complexe protéique renfermant Apo BEC-1 (ApoB mRNA Editing catalytic Component), une cytidine désaminase et aussi d'autres facteurs comme ACF (« Apo

BEC-1 Complementation Factor ») (Blanc & Davidson, 2010). L'ACF contient des motifs RRMs (RNA recognition motifs) indispensables à sa liaison au transcrit du gène ApoB et à l'enzyme Apo BEC-1. La fonction biologique de l'édition par conversion de Cytidine en Uridine est la diversification des protéines (Blanc & Davidson, 2003).

L'édition de A→I est de loin la plus répandue dans les gènes nucléaires des animaux (Zinshteyn & Nishikura, 2009). Il s'agit des transcrits ARN double brin du cerveau dont l'édition est catalysée par les protéines ADARs (Adenosine Deaminase that Act on RNA). Le site de fixation de ces protéines est un domaine appelé dsRBD (double Strand RNA Binding Domain) essentiel, résultant de la structure bidimensionnelle de l'ARN double brin formé par des séquences introniques et exoniques (Zinshteyn & Nishikura, 2009) (Figure 10). Cette édition a été identifiée aussi bien dans les régions codantes que non codantes des ARNm et dans les précurseurs des micros ARN (Nishikura, 2010). Elle conduit comme dans le cas des transcrits des récepteurs du glutamate à une diversité de protéines nécessaires au fonctionnement du cerveau (Jepson & Reenan, 2008, Zinshteyn & Nishikura, 2009).

5.2.2.2 Édition de transcrits mitochondriaux

Dans la mitochondrie, on trouve globalement quatre types d'éditions : la substitution de nucléotides, l'insertion de mono ou de di-nucléotides, la conversion de C→U et U→C et l'insertion/délétion d'uridines chez les kinétoplastides (Gray, 2003, Chateigner-Boutin & Small, 2011). Certains organismes se servent de plusieurs types d'édition d'ARN dans leur mitochondrie (Tableau II).

Edition par substitution chez les dinoflagellés

Une édition par substitution concerne les transcrits ARNm et ARNr des dinoflagellés (Gray, 2003, Lin*, et al.*, 2008, Waller & Jackson, 2009, Lin, 2011). On trouve quasiment toutes les permutations possibles donnant lieu à neuf types d'édition. Le mécanisme moléculaire d'édition est inconnu de même que les enzymes impliquées.

La diversité suggère que plusieurs mécanismes différents sont à l'œuvre (Nash*, et al.*, 2008, Waller & Jackson, 2009).

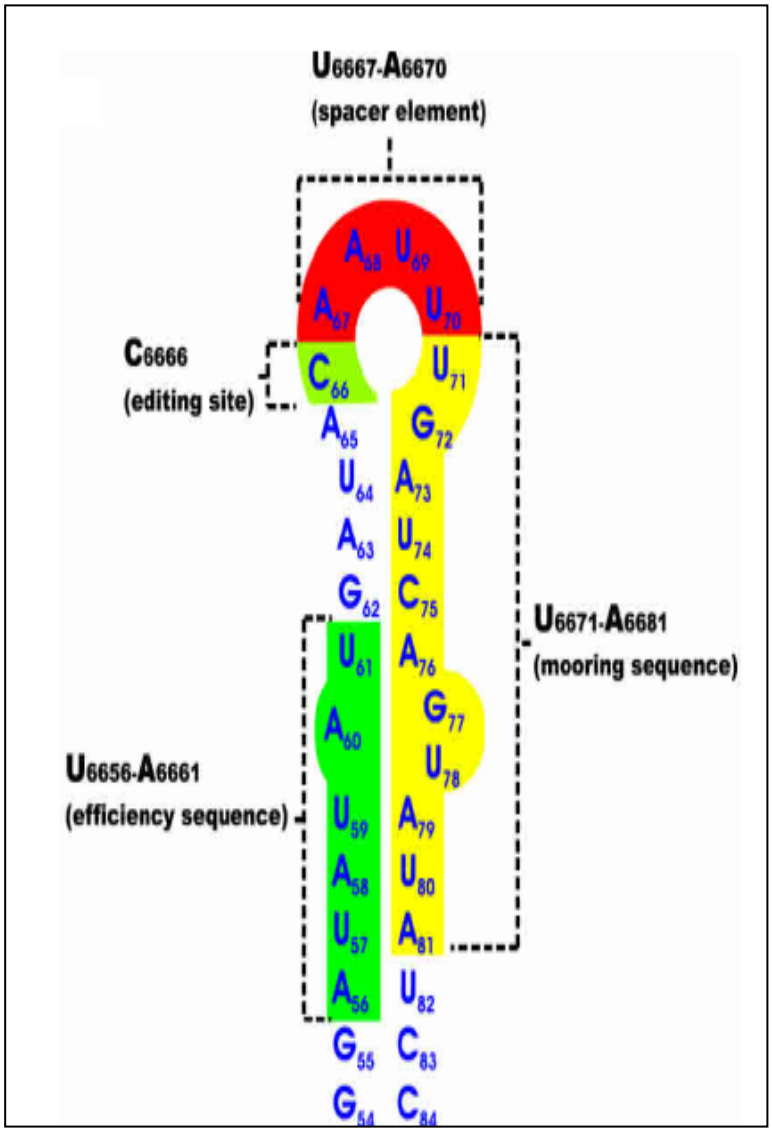


Figure 9. Structure tige-boucle de la séquence entourant la cytidine à éditer en position 6666 dans l'ARNm du gène *apoB* de l'homme (Maris*, et al.*, 2005).

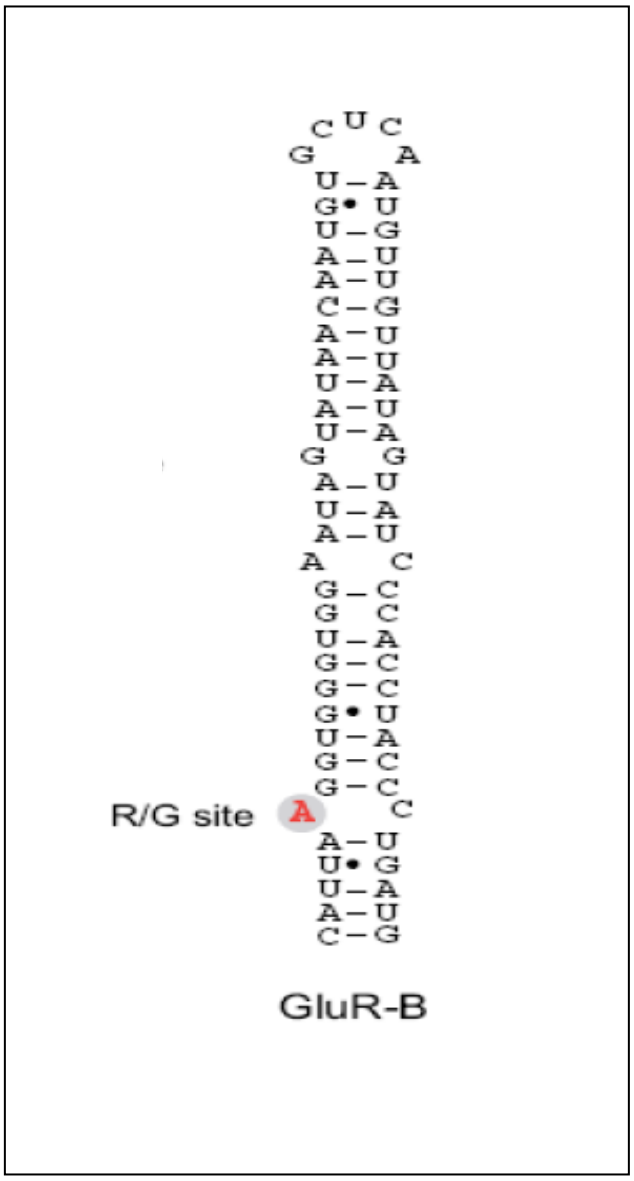


Figure 10. Structure secondaire de l'ARNm du gène de la sous-unité β du récepteur Glu chez l'homme. L'adénine éditée est en rouge (Ohman, 2007).

Tableau II. L'édition des ARN dans les organelles.

Types d'édition	ARN cibles	Localisation	Groupe
U insertion-délétion [b], **C→U**[a]	ARNm, ARNt	Mitochondrie	Kinétoplastides
C [b] **↔U**	ARNm, ARNt	Mitochondrie	Plantes
C [b] **↔U**	ARNm, ARNt	Chloroplaste	Plantes
Insertion [b] , délétion, addition de nucléotides, C→U	ARNm, ARNr, ARNt	Mitochondrie	Myxomycètes
A, C, U, G →A, C, U, G	ARNm, ARNr	Mitochondrie	Dinoflagellés
A, C, U, G →A, C, U, G	ARNm, ARNr	Chloroplaste	Dinoflagellés
U →C	ARNm	Mitochondrie	Placozoa
C→U	ARNm	Mitochondrie	*Naegleria gruberi*
U→A,G, A→G	5'ARNt	Mitochondrie	*A. castellanii, S. punctatus*
C,G,U →A,	3' ARNt	Mitochondrie	Animaux
N →A, G ,C, U	3' ARNt	Mitochondrie	*S. eucudoriensis*

[a]La conversion de C en U est représentée en gras. [b]Edition fréquente
Références : (Alfonzo*, et al.*, 1999, Knoop, 2011, Rudinger*, et al.*, 2011, Lavrov*, et al.*, 2012)

Édition « mixte » chez les myxomycètes

L'édition de transcrits dans la mitochondrie des myxomycètes (*Amoebozoa*) a été décrite pour la première fois en 1991 par le groupe de D.L. Miller, notamment l'insertion de cytidine dans 54 sites de l'ARNm du gène codant pour l'ATP synthase α de *Physarum polycepharum*. Aujourd'hui on sait que l'édition affecte aussi les transcrits ARNt et ARNr chez les myxomycètes (Gray, 2003). Elle consiste non seulement en des insertions de cytidines mais aussi d'uridines, d'adénines, de guanines et même de di-nucléotides mais aussi en des substitutions de C par U, ainsi que des délétions (Horton & Landweber, 2002, Gott*, et al.*, 2005, Bundschuh*, et al.*, 2011). Chez *P. polycepharum,* ces trois types d'édition cohabitent en plus de l'addition de nucléotides post-transcriptionnelle à l'extrémité 5' des ARNt (Gott*, et al.*, 2010). L'édition par insertion chez *P. polycepharum* a lieu pendant la transcription (Cheng & Gott, 2000) et la polymérase mitochondriale y est impliquée (Miller & Miller, 2008). Des séquences conservées de 9 nt localisées de part et d'autre de la cytidine affecteraient différemment l'édition par insertion de cytidine chez cet organisme (Rhee*, et al.*, 2009).

Edition des ARN dans les organites des plantes

L'édition des transcrits d'organites des plantes a été raportée dans la mitochondrie pour la première fois en 1989 par trois groupes de recherche (Covello & Gray, 1989, Gualberto*, et al.*, 1989, Hiesel*, et al.*, 1989, Hoch*, et al.*, 1991) et plus tard dans le chloroplaste (Hoch*, et al.*, 1991). La majorité des sites d'éditions se trouvent dans les organites de certaines plantes non vasculaires (Kugita*, et al.*, 2003). Chez *A. thaliana* par exemple, environ 500 sites d'éditions existent dans la mitochondrie contre 35 sites dans le chloroplaste (Shikanai, 2006, Chateigner-Boutin & Small, 2010). L'édition consiste en une conversion de C→U et quelque fois de U→C chez les plantes à fleurs (Gualberto*, et al.*, 1990, Schuster*, et al.*, 1990). L'édition chez les plantes concerne principalement les transcrits des gènes codants pour des protéines mais quelques exemples existent dans les ARN structuraux et les introns. Dans les deux organelles, il

s'agit de séquences *cis* situées en 5' et 3' du site qui guident l'édition, notamment de 20-30 nt en amont et cinq en aval du site d'édition (Shikanai, 2006). Ces éléments *cis* sont des sites de fixation de facteurs protéiques facilitant l'accès au site d'édition aux enzymes qui catalysent la réaction. Une cytidine désaminase serait impliquée dans l'édition de C→U des transcrits chloroplastiques et mitochondriaux, mais la machinerie de l'édition dans les organelles des plantes est encore largement inconnue (Castandet & Araya, 2011). Les PPRs ont été proposées très tôt comme candidats dans cette reconnaissance des sites d'édition mais aucune preuve du mécanisme n'a été apportée jusqu'à présent (Castandet & Araya, 2011). Les PPRs sont des protéines caractérisées par la répétition en tandem de motifs de 35 d`acides aminés (Figure 11). Les PPRs avec des domaines C terminaux DYW seraient des candidats potentiels vu leurs similarités avec les cytidines désaminases (Schmitz-Linneweber & Small, 2008) (Figure 12). Cette hypothèse est appuyée par le fait que ces PPRs sont présents chez les plantes où il ya de l'édition dans les organelles (Salone*, et al.*, 2007, Rudinger*, et al.*, 2008, Rudinger*, et al.*, 2011). Mais il est démontré que des PPRs avec des domaines C terminaux E, E+, sont aussi impliqués dans l'édition (Sung*, et al.*, 2010, Takenaka, 2010).

Pour résumer, le mécanisme enzymatique à l'origine de l'édition chez les plantes demeure inconnu. De plus, s'il est admis que les PPRs reconnaissent les séquences cis proches du site d'édition, la nature exacte de la reconnaissance reste à déterminer (Kobayashi*, et al.*, 2012, Nakamura*, et al.*, 2012, Okuda & Shikanai, 2012, Verbitskiy*, et al.*, 2012, Chateigner-Boutin*, et al.*, 2013, Yagi*, et al.*, 2013). Récemment une famille de protéines, les MORF (« Multiple Organellar RNA editing Factor ») a été identifiée, comme nécessaire à l'édition contribuant à la connaissance de l'éditosome dans les organelles des plantes (Takenaka*, et al.*, 2012, Verbitskiy*, et al.*, 2012).

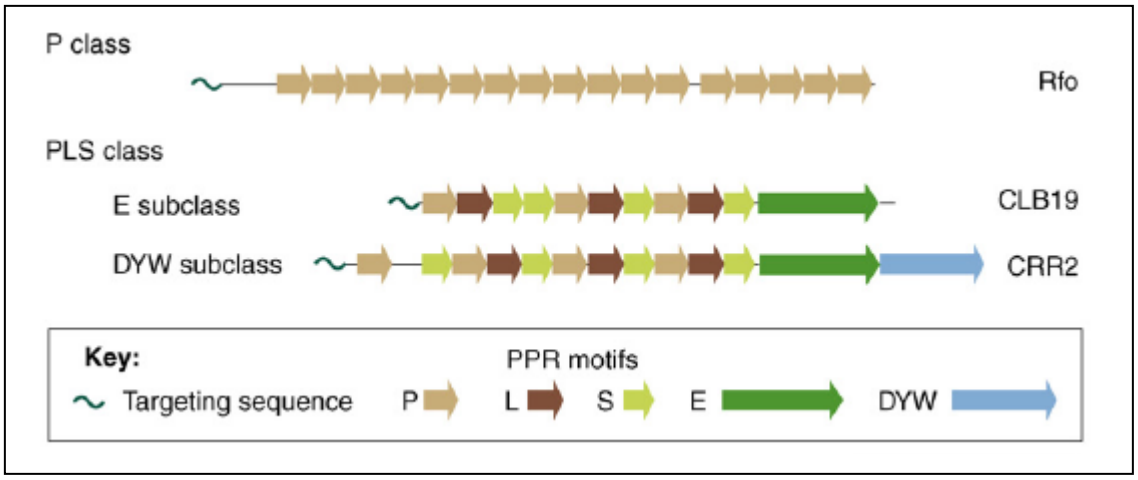


Figure 11. Les différentes classes et sous-classes des PPRs (Schmitz-Linneweber & Small, 2008). La classe P est formée de la répétition d'un motif de 35 acides aminés en tandem. La classe PLS est formée de la répétition de motifs L (« long sequence ») et S (« short sequence »). Les sous-classes E et DYW sont des PPRs de la classe PLS avec des motifs C terminaux formés respectivement par Asp (E) et par Glu-Tyr-Trp (DYW).

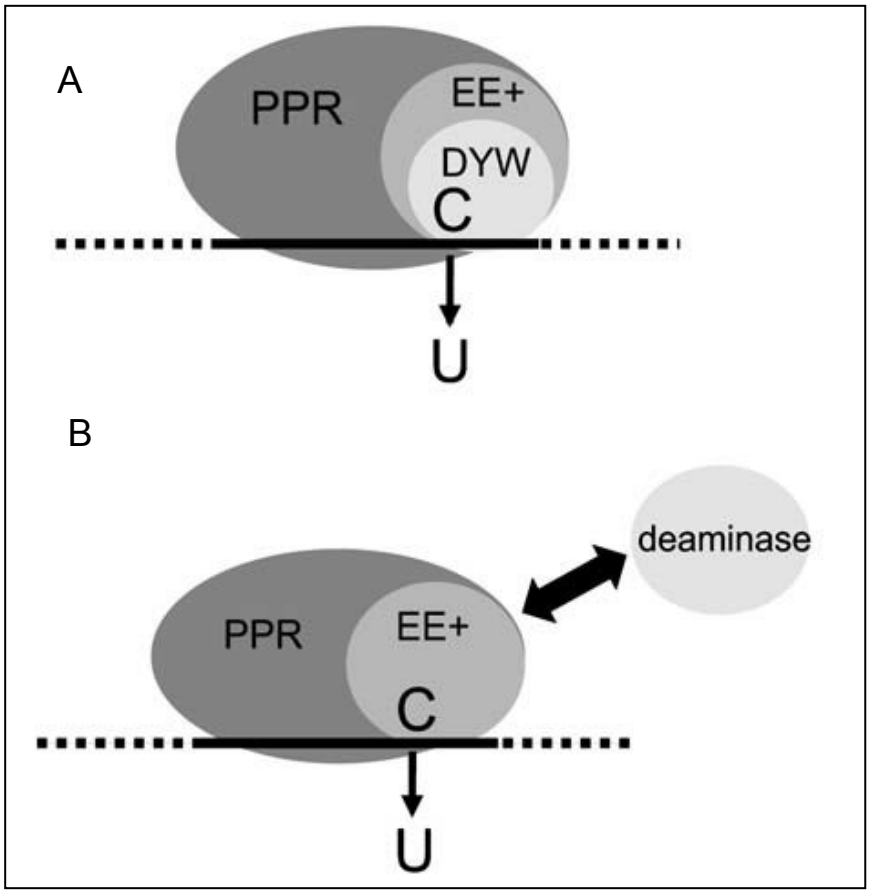


Figure 12. Hypothèses sur le rôle des PPRs dans l'édition chez les plantes (Castandet & Araya, 2011). A. Hypothèse 1 : le PPR reconnait l'ARNm à éditer et son domaine DYW catalyse la réaction de désamination de C en U (Salone*, et al.*, 2007). B. Hypothèse 2 : le PPR reconnaît l'ARNm à éditer, et son domaine E participe au recrutement d'une C désaminase (Chateigner-Boutin & Small, 2010).

L'édition des transcrits mitochondriaux chez les kinétoplastides

La majorité des transcrits des gènes mitochondriaux chez les kinétoplastides sont édités (Estevez & Simpson, 1999). Les transcrits de ces gènes, dits cryptiques, subissent une édition massive par insertion d'uridines et souvent par délétion. Par exemple, la moitié de la séquence du transcrit du gène *cox3* de *Trypanosoma brucei* est le produit de l'édition (Feagin*, et al.*, 1988). Outre cette édition massive, de l'édition limitée à 2 ou 4 sites a également été décrite chez certaines espèces (Kim*, et al.*, 1994, Lukes*, et al.*, 1994, Blom*, et al.*, 1998). L'édition résulte de la collaboration avec des ARN guides généralement encodés par les mini-cercles (Blum*, et al.*, 1990, Pollard*, et al.*, 1990, Sturm & Simpson, 1990). Cependant dans l'édition du gène *cox2* de *T. brucei*, le 3'UTR de l'ARNm sert de guide (Golden & Hajduk, 2005). Cette maturation post-transcriptionnelle des transcrits des kinétoplastides est essentielle pour générer des protéines fonctionnelles. L'édition comporte quatre étapes principales et fait intervenir plusieurs enzymes (Figure 13).

L'étape préalable à l'édition est l'hybridation entre l'ARN guide (ARNg) et le pré-ARNm à éditer. En effet l'ARNg possède dans sa région 5' une séquence complémentaire à la région du pré-ARNm qui se trouve directement en aval du site à éditer. L'ARNg a un 3' oligo-U dont le rôle est de stabiliser l'interaction ARNg - pré-ARNm (Lukes*, et al.*, 2005). L'appariement imparfait entre l'ARNg et le pré-ARNm est reconnu par la machinerie effectuant l'édition. Le clivage se fait ensuite par l'endonucléase au niveau de la première base du pré-ARNm mal appariée avec l'ARNg. La structure du complexe ARNg - pré-ARNm joue un rôle dans la reconnaissance du site d'édition par l'ARNg (Leung & Koslowsky, 2001, Leung & Koslowsky, 2001, Golden & Hajduk, 2006).

Après la coupure du pré-ARNm par une endonucléase, une exonucléase enlève les uridines qui ne sont pas appariés à l'ARNg. Une exonucléase 3' →5' spécifique pour l'uridine a été identifiée chez *Leishmania tarentolae* (Aphasizhev & Simpson, 2001) et plus tard deux exonucléases ont été décrites chez les trypanosomatides et leurs rôles

dans l'édition démontrés (Schnaufer, *et al.*, 2003, Kang, *et al.*, 2005, Rogers, *et al.*, 2007).

Après le clivage par l'endonucléase, l'ARNg joue le rôle de pont pour maintenir les deux fragments 5' et 3'du pré-ARNm en place. En effet la région 5' ou « anchor » est liée au fragment 3'du pré-ARNm et la queue poly (U) maintient le fragment 5' du pré-ARNm en se fixant à une région riche en purines.

L'insertion d'uridines se fait du côté 3' du pré-ARNm clivé et en direction 5'. C'est l'UTP libre qui est utilisé pour l'édition (Kable, *et al.*, 1996) et le nombre de U ajouté dépend du nombre de A ou de G dans l'ARN guide, situés juste après le site de coupure. En général, plusieurs ARN guides sont nécessaires pour l'édition complète d'un ARNm. L'ARN guide sert de matrice et l'insertion de nucléotides est catalysée par une TUTase. Deux TUTases ont été purifiées chez *T. brucei* et *L. tarentolae*. Il s'agit des protéines KRET1 et KRET2 qui ont un domaine de fixation de l'UTP et un domaine de fixation à l'ARN. Seule KRET2 est impliquée dans l'insertion d'uridines (Aphasizhev & Aphasizheva, 2008).

Après l'édition par insertion ou par délétion, les deux brins du pré-ARNm sont reliés pour donner le transcrit édité. La ligation se fait seulement après l'intervention d'une activité 3'phosphatase (Niemann, *et al.*, 2009). Deux ligases sont impliquées dans l'édition chez les kinétoplastides chacune étant spécifique à un type d'édition (Lukes, *et al.*, 2005, Stuart, *et al.*, 2005).

Le complexe effectuant l'édition mitochondriale a été étudié en détail chez *T. brucei* et *L. tarentolae.* Plusieurs groupes de recherche ont utilisé des systèmes d'édition *in vitro*, la génétique inverse avec ARN interférents, la purification de complexes et la spectrométrie pour dessiner la carte de l'éditosome chez ces organismes (Lukes, *et al.*, 2005). Il s'agit d'un complexe ribonucléoprotéique de 20S (Aphasizhev, *et al.*, 2003, Lukes, *et al.*, 2005, Stuart, *et al.*, 2005). Trois types d'éditosomes d'une composition similaire, caractérisés par des endonucléases différentes ont été identifiés chez *T. brucei* (Panigrahi, *et al.*, 2006, Carnes, *et al.*, 2008, Carnes, *et al.*, 2011) (Figure 14). Ces trois éditosomes ont en commun deux sous-complexes responsables de l'insertion et de la

délétion. Ils partagent aussi plusieurs autres protéines. L'un des trois éditosomes catalyse exclusivement l'édition par délétion tandis que les deux autres sont impliqués dans l'édition par insertion (Carnes*, et al.*, 2005, Carnes*, et al.*, 2008).

Pour résumer, les études dans le domaine ont catalogué les protéines catalytiques associées aux éditosomes (Tableau III). Par exemple chez *T. brucei*, on trouve des endonucléases, des exonucléases, une 3' phosphatase, une TUTase, deux ARN ligases et une hélicase ainsi que d'autres protéines supplémentaires nécessaires à l'intégrité de ces complexes ribo-protéiques (Stuart*, et al.*, 2005, Ernst*, et al.*, 2009).

Plusieurs facteurs protéiques qui ne font pas partie de l'éditosome, participent indirectement à l'édition d'ARNm chez *T. brucei* (Aphasizhev & Aphasizheva, 2011) (Tableau III). Ce sont par exemple des protéines de liaison à l'ARN qui facilitent l'association entre l'ARNg et le pré-ARNm et régulent non seulement l'efficacité de l'édition mais aussi la stabilité des transcrits édités et du duplex ARNg - pré-ARNm. Le complexe MRB1 par exemple (« Mitochondrial RNA Binding complex 1 »), stabilise les transcrits édités ainsi que le duplex ARNg - pré-ARNm. Ce complexe ne s'associe que transitoirement avec l'éditosome.

Le taux d'expression des transcrits édités varie au cours du cycle de vie de *T. brucei* et cette variation n'est pas liée aux ARNg (Lukes*, et al.*, 2005). L'édition est maximale au stade pro-cyclique et réduite au stade sanguin, permettant ainsi de réguler la production d'ATP au cours du cycle de vie du parasite. L'existence d'une édition alternative à l'origine de la diversité protéique, a été proposé par le groupe de Hadjuk (Hajduk & Ochsenreiter, 2010).

Il y a beaucoup de types d'édition dans la mitochondrie mais la reconnaissance du site d'édition est connue pour seulement deux, notamment l'édition par insertion/délétion d'uridines chez les kinétoplastides et l'édition par conversion de C→U chez les plantes. Ces deux types d'édition diffèrent d'une façon fondamentale. L'un utilise comme guides des molécules d'ARN tandis que l'autre utilise des liaisons avec des protéines qui reconnaissent le site sur le pré-ARNm à éditer.

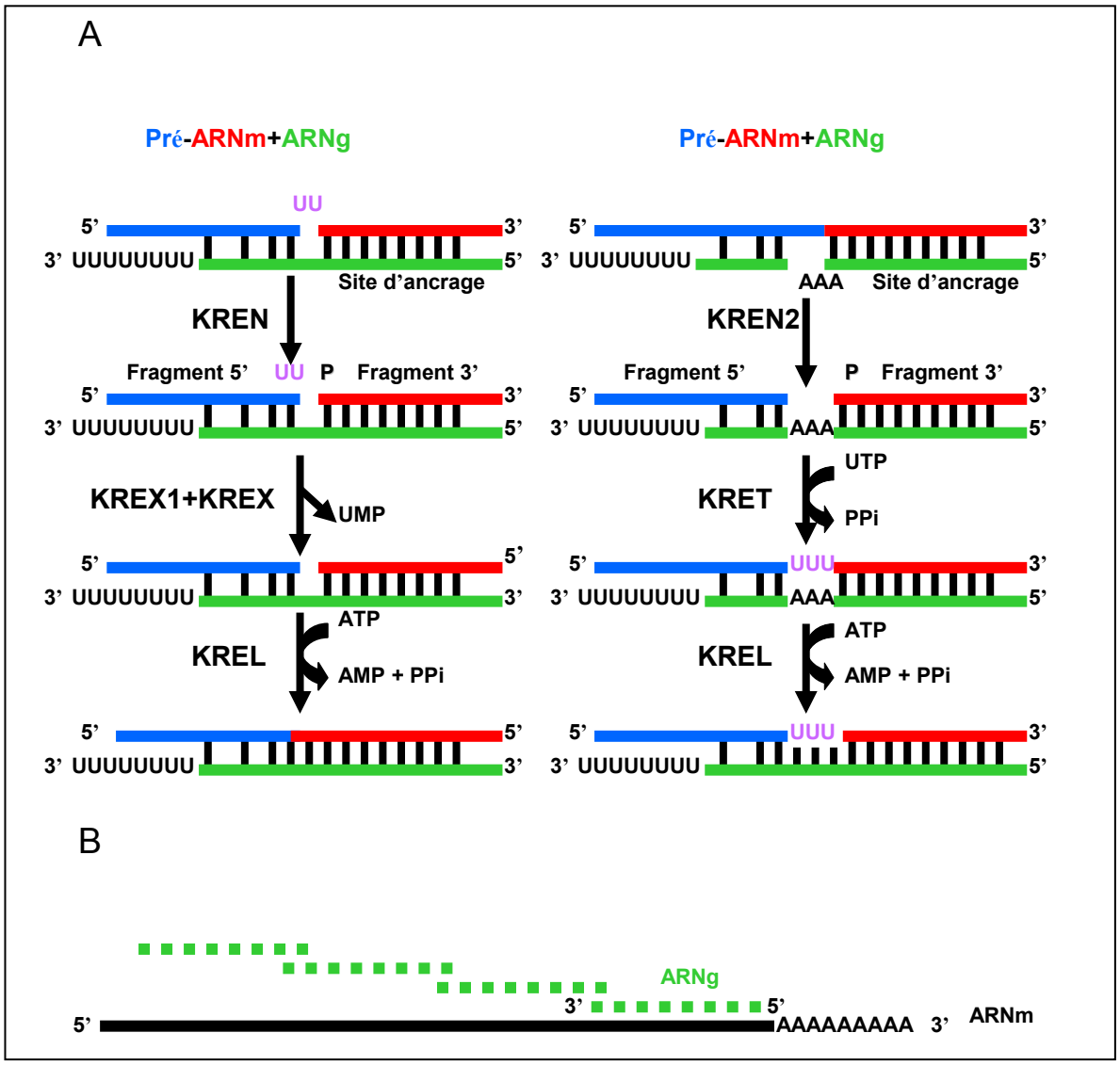


Figure 13. Edition par insertion/délétion d'uridines chez les kinétoplastides (Aphasizhev & Aphasizheva, 2011). A. Étapes de l'édition et protéines impliquées dans l'édition. A gauche, l'édition par délétion d'uridines et à droite l'édition par insertion d'uridines. B. Plusieurs ARNg sont nécessaires pour l'édition d'un ARNm.

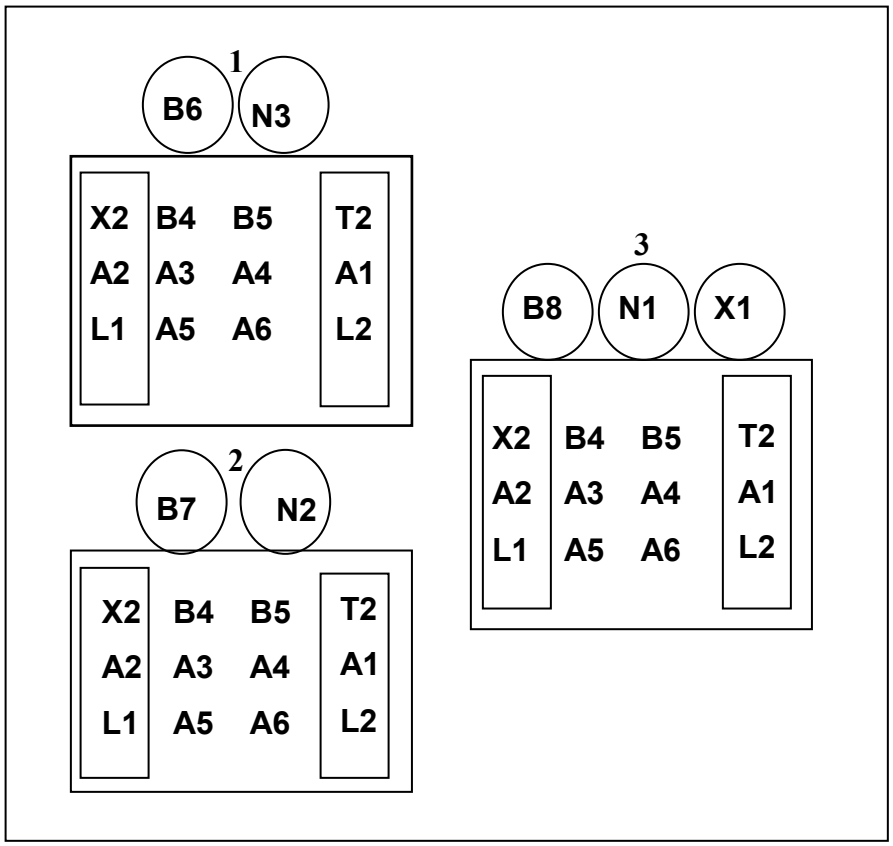


Figure 14. Composition des trois éditosomes de la mitochondrie de *T. brucei* (Ernst*, et al.*, 2009). 1; 2, Éditosomes spécifiques de l'insertion d'uridines. 3, Éditosome spécifique de la délétion d'uridines. N1, N2, N3, Endonucléases spécifiques à la délétion et à l'insertion d'uridines. X2, Exonucléase spécifique à l'édition par délétion. B6-8, Protéines de liaisons spécifiques à chaque éditosome. X2-A2-L1, T2-A1-L2, Sous complexes de l'édition par délétion et par insertion. B4, B5, A3-A6, Protéines communes aux trois éditosomes.

Tableau III. Protéines impliquées dans l'édition par insertion/délétion chez les kinétoplastides.

Protéines	Fonction	Références
KREN1	Endonucléase	(Panigrahi, *et al.*, 2006)
KREN2	Endonucléase	(Panigrahi, *et al.*, 2006)
KREN3	Endonucléase	(Carnes, *et al.*, 2008)
KREX1	Exonucléase spécifique de la délétion	(Kang, *et al.*, 2005)
KREX2	Exonucléase	(Schnaufer, *et al.*, 2003)
KREL1	RNA ligase pour l'édition par délétion	(Cruz-Reyes, *et al.*, 2002)
KREL2	RNA ligase pour l'édition par insertion	(Cruz-Reyes, *et al.*, 2002)
KRET1	TUTase (queue oligo(U) des ARNg	(Aphasizhev, *et al.*, 2002)
KRET2	TUTase (insertion d'uridines, édition)	(Ernst, *et al.*, 2003)
MEAT1	TUTase uridines spécifique	(Aphasizheva, *et al.*, 2009)
REH1	Hélicase	(Missel, *et al.*, 1997, Li, *et al.*, 2011)
REH2	Hélicase	(Hernandez, *et al.*, 2010)
KREPA1	Protéine de liaison	(Panigrahi, *et al.*, 2001)
KREPA2	Protéine de liaison	(Panigrahi, *et al.*, 2001)
KREPA3	Protéine de liaison	(Panigrahi, *et al.*, 2001)
KREPA4	Protéine de liaison	(Panigrahi, *et al.*, 2001)
KREPA5	Protéine de liaison	(Stuart, *et al.*, 2005)
KREPA6	Protéine de liaison	(Panigrahi, *et al.*, 2001)
KREPB4	Protéine de liaison	(Panigrahi, *et al.*, 2003)
KREPB5	Protéine de liaison	(Panigrahi, *et al.*, 2003)
KREPB6	Protéine de liaison	(Guo, *et al.*, 2012)
KREPB7	Protéine de liaison	(Guo, *et al.*, 2012)
KREPB8	Protéine accessoire	(Guo, *et al.*, 2012)
KREPB9	Protéine accessoire	(Lerch, *et al.*, 2012)
KREPB10	Protéine accessoire	(Lerch, *et al.*, 2012)
KREPC1	Phosphatase 3'	(Niemann, *et al.*, 2009)
KREPC2	Phosphatase 3'	(Niemann, *et al.*, 2009)
KPAP1	Protéine accessoire	(Etheridge, *et al.*, 2008)
MRB1	Protéine accessoire	(Panigrahi, *et al.*, 2008)
MRP1/2	Protéine accessoire	(Simpson, *et al.*, 2004)
RBP16	Protéine de liaison	(Pelletier & Read, 2003)
REAP-1	Protéine de liaison	(Hans, *et al.*, 2007)
TbRGG1	Protéine de liaison, stabilise les ARN édités	(Hashimi, *et al.*, 2008)
TbRGG2	Protéine accessoire	(Ammerman, *et al.*, 2010)

5.2.3 Maturation des extrémités des transcrits mitochondriaux

La maturation des transcrits mitochondriaux implique l'excision des extrémités 5' et 3'et est réalisée par des endonucléases et des exonucléases encodées par le noyau. Ces ARNases peuvent être spécifiques ou non à la mitochondrie. Si la maturation des extrémités 5' et 3' des autres transcrits mitochondriaux est peu connue, celle des transcrits immatures des ARNt a été bien étudiée chez beaucoup d'eucaryotes. L'excision de l'extrémité 5' du pré ARNt est réalisée par le complexe ribo-protéique mitochondrial, la RNase P. La maturation de l'extrémité 3' est catalysée par une RNase de type Z qui a été localisée dans la mitochondrie chez l'homme (Rorbach & Minczuk, 2012).

Les transcrits mitochondriaux n'ont pas la coiffe 7-méthyl guanine caractéristique des transcrits nucléaires. La maturation des transcrits mitochondriaux inclue parfois l'ajout de nucléotides aux extrémités 3' notamment d'un CCA aux ARNt et, chez certains eucaryotes, de poly (A) aux ARNm (Gagliardi*, et al.*, 2004). Les ARNr et ARNg des trypanosomes et certains transcrits des myxomycètes ont une queue poly (U) au 3' (Barbrook*, et al.*, 2010). La polyadénylation est réalisée par une Poly A Polymérase (PAP) dont celle de l'homme, la hmt PAP (Nagaike*, et al.*, 2008). Dans les cas plutôt rares d'ARNm mitochondriaux polyadénylés, cette queue joue des rôles très divers (Chang & Tong, 2012) : Elle complète les codons stop et stabilise l'ARNm chez les mammifères tandis que dans la mitochondrie des plantes elle est un signal de dégradation (Lange*, et al.*, 2009). Chez les trypanosomes dont les transcrits mitochondriaux sont édités, une queue poly A stabilise les transcrits durant et après l'édition tandis qu`une queue poly A/U ajoutée après l'édition sert de site de fixation du ribosome au cours de la traduction. Des protéines telles que KPAP1 et RET1 sont impliquées dans ces deux processus mais également des protéines PPRs (Aphasizhev & Aphasizheva, 2011).

6 Résumé de l'état des connaissances sur l'ADNmt et l'expression des gènes de *D. papillatum*

Le génome mitochondrial de *D. papillatum* a été étudié pour la première fois dans le groupe de Simpson et collaborateurs (Maslov*, et al.*, 1999). Ils ont décrit un ADNmt riche en GC ainsi qu'une structure du génome consistant probablement en plusieurs cercles. Il a fallu attendre des travaux dans notre laboratoire de recherche pour avoir des données précises sur la structure du génome mitochondrial (Marande*, et al.*, 2005). La microscopie électronique et par fluorescence a permis de dévoiler l'ultra-structure du génome mitochondrial de *D. papillatum* et la distribution de cet ADN dans l'organite. Il s'agit d'un ADNmt distribué de façon uniforme dans une seule grande mitochondrie et non compacté comme ce qui a été décrit chez les kinétoplastides, le groupe frère des diplonémides.

Ces études ont montré que l'ADNmt est multi-chromosomique, constitué d'une centaine de chromosomes circulaires monomériques de deux classes de taille: classe A (6 kb) et classe B (7 kb). Déjà en 2005, il avait été observé que le gène *cox1* est fragmenté et que chaque fragment est porté par un chromosome différent.

Depuis 2005, nous avons fait plusieurs découvertes. Ainsi, les chromosomes sont constitués d'une cassette et d'une région constante (Marande & Burger, 2007). La cassette (190 à 470 pb) comprend un seul fragment de gène (module de 60-350 pb) flanqué par environ 4-150 pb en 5' et en 3' (Marande & Burger, 2007, Vlcek*, et al.*, 2011). Les chromosomes d'une même classe partagent 95% de la région constante non codante avec plus de 97% d'identité de séquence. Une région de 2.6 kb est même partagée par les chromosomes de classes différentes avec 97% d'identité (Marande & Burger, 2007). Pour résumer nos recherches sur *D. papillatum*, plus de 250 kb du génome mitochondrial (la taille du génome est estimée à environ 650 kb) et 35 kb du transcriptome mitochondrial ont été séquencés. La portion de l'ADN mt séquencée jusqu'à présent code pour dix protéines de la chaine respiratoire et pour la grande sous

unité de l'ARN ribosomal et tous les gènes mitochondriaux identifiés jusque là sont fragmentés. Par exemple, le gène *cox1* qui code pour la sous unité un du complexe de la cytochrome oxidase, est en neuf morceaux encodés par des chromosomes différents (Marande & Burger, 2007, Burger*, et al.*, 2008). Chaque chromosome est constitué d'une région variable, une cassette et une région constante. La cassette est constituée du module (région codante) entourée de ses deux séquences flanquantes uniques. L'information sur l'ARNm de *cox1* a été obtenue grâce à des expériences de Northern Blot et l'identification de transcrits intermédiaires dans la banque ADNc ont permis de conclure que chaque module est transcrit de façon indépendante. Les transcrits des modules sont ensuite joints par épissage en *trans* dont le un mécanisme est inconnu pour donner finalement l'ARNm (Marande & Burger, 2007) (Figure 15). Une édition par insertion d'uridines entre deux modules a aussi été identifiée. Elle n'est pas sans rappeler l'édition par insertion/délétion d'uridines caractéristique des mitochondries chez les kinétoplastides. En tenant compte du nombre de modules, environ une centaine estimée dans la mitochondrie de *D. papillatum*, deux hypothèses ont été formulées :

Hypothèse 1 : L'épissage en *trans* se fait grâce à l'épissage d'introns discontinus inconnus permettant l'appariement des régions codantes ou flanquantes et la ligation de deux modules.

Hypothèse 2 : L'épissage en *trans* est médié par des ARNg ressemblant à ceux qui dirigent l'édition des trypanosomes (Stuart*, et al.*, 2005). Ces ARNg lieraient des modules adjacents mais seraient aussi à la base de l'édition par insertion d'uridines.

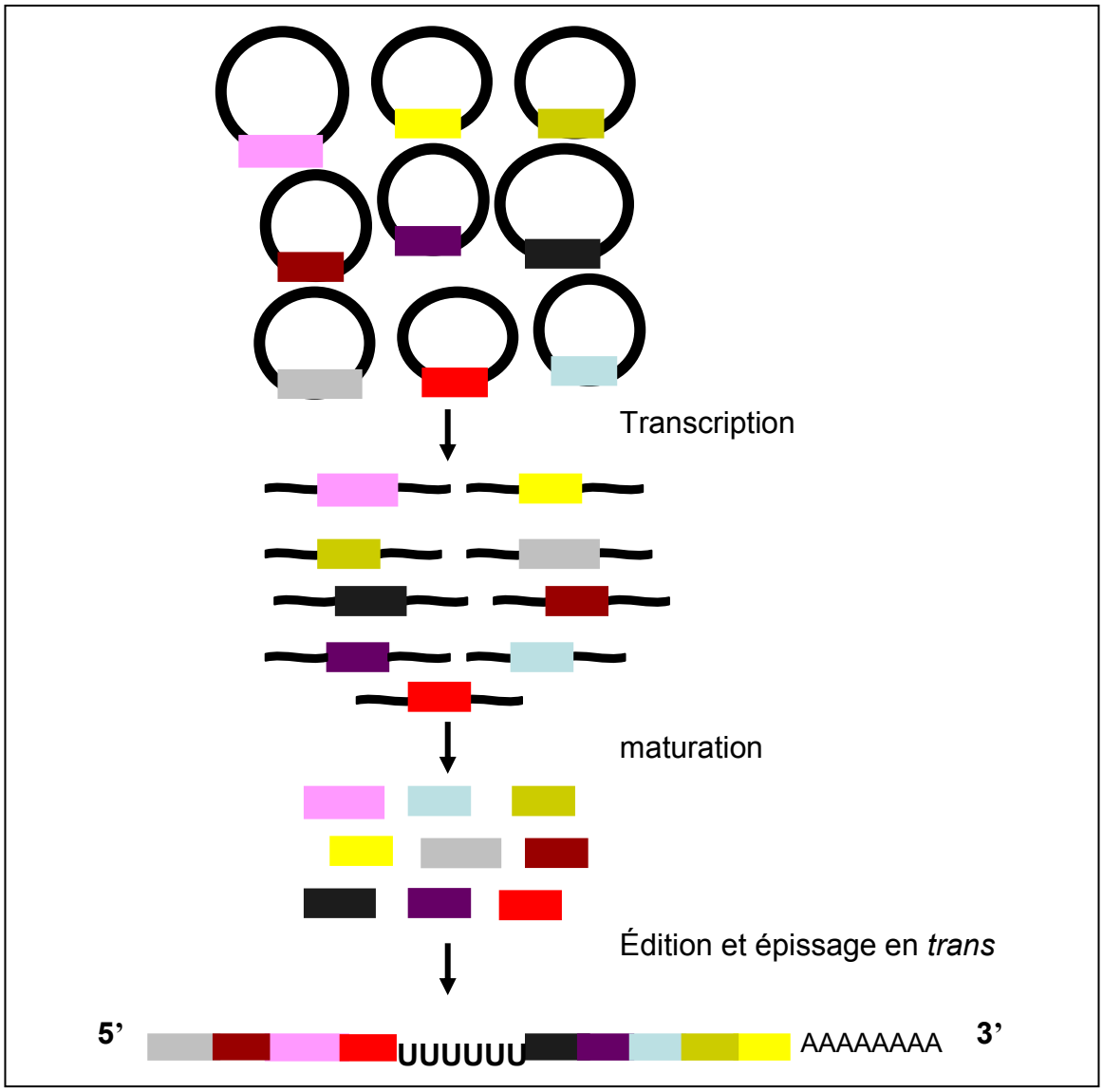


Figure 15. Expression de *cox1* chez *D. papillatum*. Les modules de *cox1* représentés dans des couleurs différentes sont encodés par des chromosomes différents. Ils sont transcrits, subissent une maturation et ensuite liés par l`épissage en *trans*. Les six uridines en gras sont ajoutés par l'édition dans l'ARNm.

7 Objectifs de cette thèse

La section précédente décrit les connaissances du système mitochondrial chez *Diplonema* avant le début de notre projet de thèse. Notre recherche a été axée sur les processus post-transcriptionnels dans la mitochondrie des diplonémides et comprenait trois objectifs spécifiques, à savoir:

1. Déterminer si les caractères du gène *cox1* (la fragmentation, l'épissage en *trans* et l`édition) identifiés chez *D. papillatum* sont communs au groupe des diplonémides (*Diplonema* et *Rhynchopus)*.

2. Caractériser l'épissage en *trans* des gènes chez les diplonémides.

3. Déterminer si à part l'édition du transcrit du gène *cox1*, on retrouve d'autres sites et d'autres types d'édition dans la mitochondrie de *D. papillatum*.

Chapitre 2. Matériels et Méthodes

1 Matériels

1.1 Matériels biologiques

Quatre espèces de diplonémides : *D. papillatum* (ATCC 50162), *D. ambulator* (ATCC 50223), *D.* sp. 2 (ATCC 50224) et *Rhynchopus euleeides* (ATCC 50226). La souche bactérienne *E. coli* DH5α a été utilisée pour les transformations par choc thermique.

1.2 Autres matériels

Les vecteurs de clonage pBLF6cat issu de pBluescript (Stratagene) du laboratoire du Dr Franz Lang et pDNR-LIB de CLONTECH ont été utilisés pour la construction des banques ADNc. Toutes les amorces utilisées ont été conçues en utilisant le site d'Integrated DNA TECHNOLOGIES (www.idtdna.ca) et commandées à la compagnie BIOCORP (www.biocorp.ca). Les marqueurs de poids moléculaires suivants ont été utilisés : 1kb+ d'Invitrogen, λHind III de Fermentas pour l'ADN, Ribo Ruler High Range et Low range RNA Ladder de Fermentas pour l'ARN et un marqueur double brin fait maison pour les plasmides.

2 Méthodes

2.1 Cultures de *D. ambulator D.* sp.2 et *R. euleeides*

Les quatre espèces ont été cultivées dans des flasques de culture (cell culture flask, de CORNING, cat. 430621) dans du milieu constitué de 3.3 % de sels Instant Ocean® et de 1-10 % de sérum de cheval décomplémenté (Wisent; cat. 065150). Ce milieu a été enrichi de 0.1 % de tryptone pour *D. papillatum*; 1% de tryptone pour *D. ambulator* ; 10% de tryptone, 1% de solution d'enrichissement (0.5 µg/ml EDTA.$2H_2O$, 0.5 µg/ml $NaNO_3$, 1 µg/ml $Na_2SiO_3.9H_2O$, 1 µg/ml de sodium glycérophosphate, 0.5 µg/ml H_3BO_3, 0.5 µg/ml $Fe(NH_4)_2(SO_4)_2.6H_2O$, 1 µg/ml $FeCl_3.6H_2O$, 0.5 µg/mg $MnSO_4.4H_2O$, 5

µg/ml $ZnSO_4.7H_2O$, 1 µg/ml $CoSO_4.7H_2O$) et 0.1% de vitamines (10 µg/ml Thiamine, 0.5 µg/ml Vitamine B12 et 0.5 µg/ml Biotine) pour *D.* sp. 2 et *R. euleeides.* Les cellules ont été prélevées au stade logarithmique de leur croissance et comptées à l'aide d'un hématocytomètre. Environ 3 $x10^8$ cellules ont été utilisées pour l'extraction de l'ADN ou de l'ARN (Figure 16).

2.2 Ouverture des cellules de diplonémides

Les cellules sont centrifugées à 3000 rpm (1100 RCF) avec le rotor GSA de Sorvall pendant 10 min et le culot cellulaire est lavé dans du tampon STE (0.6 M de Sorbitol, 50 mM Tris et 5 mM EDTA, pH 7.5). L'ouverture des cellules est réalisée mécaniquement pendant 2-5 min, à température ambiante en utilisant une seringue ou des billes de verre.

2.3 Purification de l'ADN mitochondrial

Le lysat cellulaire est purifié sur gradients de sucrose (80 g, 50 g, 25 g, et 18 g de sucrose dans 100 ml d'eau) (soit respectivement 10 ml de 80%, 10 ml de 50%, 5 ml de 25% et 5 ml de 18% dans des tubes Beckman (cat. 344058)) par ultracentrifugation à 20000 RCF avec le rotor SW28 Beckman 8373 pendant 45 min; les centrifugeuses ultra L8-70 M de Beckman et Sorvall discovery 100 SE. Les deux phases intermédiaires du gradient ont été récupérées pour refaire le même gradient de densité. La fraction mitochondriale (phases 50-25% sucrose) est récupérée à l'issue de cette seconde centrifugation et utilisée pour l'extraction ADN/ARN.

2.4 Extraction ADN/ARN total

L'ADN/ARN a ensuite été extrait à l'aide d'une solution de pH 5-7, constituée de 38% de phénol 12% de guanidine thiocyanate, 7.5% d'ammonium thiocyanate, 3.25% d'acétate de sodium 3 M et 10% de glycérol 50%. L'ADN/ARN est ensuite précipité avec de l'isopropanol 70% froid et lavé avec de l'éthanol 70% avant d'être solubilisé

dans de l'eau traitée au Diéthyle Pyrocarbonate (DEPC). Pour obtenir l'ADN ou l'ARN des traitements avec l'ARNase I de NEB (cat. M0243S) ou de Fermentas (cat. EN0601) ou avec l'ADNase de Roche (cat. 041672001) ou de Fermentas (EN0521) sont utilisés ou une purification sur colonne RNeasy de Qiagen (cat. 74104).

2.5 Préparation de l'ARN poly (A)

L'ARN poly (A) été obtenu après deux passages de l'ARN total sur une colonne oligo dT cellulose type 7 d'Amersham Biosciences en utilisant respectivement une solution de fixation 1X (10 mM Tris, pH 7, 1 mM EDTA, pH8, 0.05% SDS, 0.5 M NaCl), une solution de fixation 2X (20 mM Tris, pH 7, 2 mM EDTA, pH8, 0.1 % SDS, 1M NaCl), une solution de lavage (10 mM Tris, pH 7, 1mM EDTA, pH8, 0.05% SDS, 0.2 M NaCl) et une solution d'élution (0.5 M Tris EDTA, pH 8). Après précipitation dans de l'éthanol ammoniun acétate 95% (38.5 g d'ammonium acétate dans 1 L d'éthanol 95%), l'ARN poly (A) est lavé avec de l'éthanol 70% et solubilisé dans de l'eau DEPC.

2.6 Digestion enzymatique de l'ADNmt

Un microgramme d'ADNmt a été digéré en présence de plusieurs concentrations d'ADNase I de Roche (0.5, 1 et 2 unités respectivement). Le produit de la réaction a été visualisé sur gel agarose 0.8 % puis coloré au Bromure d'Ethidium (BET) à 1 μg/ml.

2.7 Southern Blot du gène *cox1* de diplonémides

Nous avons analysé l'ADNmt des diplonémides par Southern Blot. Pour cela dix μg d'ADN total sont séparés sur gel d'agarose et traités successivement avec des solutions de dénaturation (1.5 M NaCl, 0.5 M NaOH), neutralisation (1.5 M NaCl, 1M Tris, pH 8) et de transfert, solution saline de citrate de sodium (5X SSC) avant le transfert humide sur membrane Hybond-N d'Amersham toute la nuit. La membrane est ensuite rincée

dans du 6X SSC, séchée, fixée sous UV et hybridée avec une solution d'hybridation contenant du 6X SSC, 6X Denhardt (1% Ficoll 400, 1% polyvinylpyrrolidone, 1% d'albumine de sérum bovin), 0.8% SDS, 100 µg/ml d'ADN de sperme de saumon et les sondes ADN. Les sondes ADN ont été obtenues par amplification PCR du module terminal de *cox1*. Pour chaque espèce, on a utilisé les amorces suivantes: da1+da2, ds1+ds9 et re9+re10 (voir séquences page 90 Table S6 ou page 230 Table 1). Les sondes ont été ensuite marquées avec du [α-^{32}P]-dATP de Perkin Elmer (cat. BLU012H250UC) et utilisées pour l'hybridation. La membrane est ensuite lavée trois fois avec des solutions de SSC (2X SSC+0.5 M de Dodecyl Sulfate de Sodium (SDS), 1X SSC+0.1 M de SDS), séchée et le film d'Amersham (cat. 28906845) est exposé toute la nuit.

2.8 Analyse des ARN du gène *cox1* de diplonémides par Northern Blot

Les ARN totaux traités avec de l'ADNase et le poly (A) ont été séparés par électrophorèse sur gel d'agarose 1.5% dénaturant (37% formaldéhyde) et transférés sur membrane Hybond -N toute la nuit. La membrane est séchée et les ARN fixés aux ultra violet (UV) avant d'être hybridés aux sondes (constituées de l'ADN du module terminal de *cox1),* marquées par radioactivité avec du [α-^{32}P]-ATP de Perkin Elmer (cat. BLU003H250UC) dans une solution contenant 6X SSC, 6X Denhardt, 0.8% SDS et 100 µg/ml d'ADN de sperme de saumon. La membrane est ensuite lavée trois fois avec des solutions de SSC (1X SSC+0.1 M de SDS, 0.5X SSC+0.1 M de SDS), séchée et le film exposé toute la nuit.

2.9 Analyse des transcrits mitochondriaux du gène *rnl* de *D. papillatum* par Northern Blot

Dans le but d'identifier tous les transcrits du gène *rnl*, 10 µg d'ARN total (dans lequel les ARN poly (A) ont été enlevés) sont traités avec de l'ADNase et séparés sur gel 1.5 % agarose dénaturant (37% formaldéhyde) avant d'être transférés sur membrane

Hybond N d'Amersham. Le module terminal du gène *rnl* amplifié par PCR avec les amorces dp168 et dp169 est marqué radioactivement avec du [α-^{32}P]-ATP de Perkin Elmer (cat. BLU003H250UC) et utilisé comme sonde pour l'hybridation.

2.10 Recherche des chromosomes de *cox1* par amplification génique

Pour amplifier les différents chromosomes portant des modules de *cox1* chez les diplonémides, des amorces spécifiques et divergentes situées sur chaque module (Figure 17 A) ont été conçues à partir de la séquence d'ADN complémentaires (ADNc). Les produits PCR obtenus avec le kit Takara Bio (cat. RR001A) ont été fragmentés par nébulisation à 15 psi 3-5 mn avec de l'azote, clonés et séquencés (Figure 1, annexe). Deux autres amorces spécifiques ont été conçues dans les régions flanquantes de la séquence chromosomique portant le module partiel pour compléter la partie manquante ou cœur du module (Figure 17 B).

2.11 Amplification génique à travers plusieurs modules de *cox1*

Des amorces couvrant plusieurs modules de *cox1* (module 3-6 pour *D. ambulator*, module 6-9 pour *D*.sp. 2 et module 1-4 pour *R. euleeides*) ont été utilisées pour tenter d'amplifier plusieurs modules de *cox1* à partir de l'ADN. Ce sont respectivement da16+da21, ds12+ds1 et re27+re15 (Figure 18) (voir séquences page 90, Table S6 ou page 230, Table 1).

2.12 Recherche d'ARN antisens par transcription inverse

Plusieurs expériences de transcription inverse ont été réalisées pour amplifier des ARN antisens qui seraient à la base de l`épissage en *trans* chez les diplonémides, ou pour compléter la banque ADNc de *D. papillatum* avec des amorces spécifiques.

2.12.1 Long ARN antisens de *cox1*

De l'ARN total purifié par passage à travers une colonne RNeasy, a été traité avec de l'ADNase et utilisé pour amplifier un ADNc antisens de *cox1* chez *D. ambulator*, *D*.sp. 2 et *R. euleeides* en utilisant les amorces spécifiques da14+da1, ds34+ds1 et re27+re4 (Figure 19) (voir séquences page 90 Table S6 ou page 230 Table 1).

2.12.2 ARN antisens d'autres gènes mitochondriaux de *D. papillatum*

De l'ARN total obtenu à partir de fractions mitochondriales de *D. papillatum* et traité avec de l'ADNase a été utilisé pour amplifier des ADNc antisens des gènes *atp6*, *nad5*, *nad7*, *cob* et *cox2*. Les amorces utilisées pour la RT-PCR couvraient deux modules au moins (Tableau IV).

2.13 Détermination de la longueur d'hypothétique des ARN antisens par extension d`amorce

Nous avons utilisé la méthode de transcription inverse selon le protocole de Promega pour analyser la longueur des ARN antisens du gène *cox1*. Pour cette expérience nous avons marqué les extrémités 5' de toutes les amorces et les marqueurs utilisés avec du [γ-^{32}P]-ATP. Respectivement 4, 50, et 200 µg d'ARN poly (A), d'ARNmt enrichi et d'ARN total ont été incubés avec l'enzyme RTAMV de Roche en présence de 1 mM de désoxynucléotides triphosphates (dNTPs) et 40 µM de pyrophosphate. Les échantillons ont été ensuite analysés sur gel de polyacrylamide 8% et exposés.

2.14 Séquençage des extrémités 5' et 3' des ADNc de gènes mitochondriaux de *D. papillatum*

Pour avoir des informations sur les extrémités des ARNm mitochondriaux, nous avons utilisé des amorces spécifiques situées aux bouts 5' et 3' pour réaliser des RT-PCR sur de l'ARN circularisé (Tableau V). Dix µg d'ARN total traité avec de l'ADNase

ou d'ARN total RNeasy sont traités avec l'enzyme TAP (Tobacco Acid Pyrophosphatase) d'Epicentre Biotechnologies (cat. T19250) pour enlever la coiffe triphosphate des ARN-précurseurs. Après une précipitation avec de l'éthanol 95% ammonium acétate, les ARN sont solubilisés dans de l'eau traité au DEPC. Ils sont ensuite traités ainsi que de l'ARN poly (A) avec les T4PNK (Poly Nucléotide Kinase du phage T4 kinases de New England BioLabs (NEB) (cat. M0236L et cat. M0201L) qui ajoute et enlève respectivement un phosphate en 5' et 3' des ARN. Les ARN sont précipités avec de l'éthanol 95% ammonium acétate et solubilisés dans de l'eau traitée au DEPC. De petites quantités de ces ARN (20-50 ng) sont circularisées et utilisées pour amplifier les extrémités des ARNm. Les ARN circularisés sont extraits au phénol chloroforme alcool isoamyl et précipités avec de l'éthanol 95% ammonium acétate. La transcriptase inverse AMV (Avian Myeloblastosis Virus) (cat. 11495062001) et la Expand High Fidelity de Roche (cat. 11732650001)) sont alors respectivement utilisées pour la RT et la PCR. Pour compléter le 5' de l'ADNc du gène *rnl* nous avons utilisé de l'ARN total dans lequel les poly (A) ont été enlevés.

2.15 Préparation d'une banque ADNc à partir de l'ARN poly (A) et normalisation de la banque

Une banque d'ADNc avait déjà été faite et l'objectif ici était de normaliser la banque de façon à obtenir les transcrits les plus faiblement exprimés absents dans notre librairie. Le kit Creator SMART DNA (CLONTECH cat. 634903) a été utilisé pour préparer la banque ADNc. La transcription inverse a été faite avec les amorces SMART IV et CDS III qui possèdent chacun un site de restriction SfiI et se fixent respectivement sur le 5' et la queue poly (A) des ARNm en utilisant la Power Script Reverse transcriptase. Les produits PCR obtenus par amplification avec les amorces CDSIII et 5' PCR Primer et la Advantage 2 Polymérase sont sélectionnés selon leur taille et purifiés par électroélution. Les fragments sont ensuite traités avec la DSN (Duplex Specific Nuclease EA001 de Innovative Biotechnology) pour normaliser la banque. Une seconde

PCR est alors réalisée avec l'ADNc traité avec la DSN, en présence de SMARTIV et CDSIII, puis les produits sont traités avec la protéinase K et digérés avec l'enzyme de restriction SfiI pour permettre leur ligation dans le vecteur pDNR-LIB. Les produits digérés sont purifiés au phénol chloroform alcool isoamyl.

2.16 Clonage des produits PCR et RT-PCR

Le clonage des ADNc de la banque poly (A) normalisée a été fait dans pDNR-LIB au niveau du site SfiI. Tous les autres produits PCR ou RT-PCR ont été clonés dans le site EcoRV du vecteur pBFL6cat, issu de pBluescript (Stratagene). Comme le vecteur pBFL6cat est déphosphorylé, nous avons phosphorylé les inserts avec la T4 PNK de NEB (cat. M0236L) ou de Roche (cat. 10709557001) avant la ligation dans un rapport 2/1 à 14°C toute la nuit en présence de la T4 RNA ligase de Roche (cat.11449478001) ou de NEB (cat. M0202L).

2.17 Transformation et extraction de l'ADN plasmidique

Les bactéries *E. coli* DH5α ont été transformées par choc thermique avec la moitié de la réaction de ligation et les clones positifs (blancs) sélectionnés sur milieu solide LB/ chloramphénicol (CAM) (10 mg/L) / tétracycline (TET) (5 mg/L) / IPTG (20 mM /L) / X-GAL (40 mg/L). Les colonies blanches sont mises en culture dans des blocs de 96 puits (Qiagen) dans du milieu liquide LB/CAM (5 mg/L) /TET 5 mg/L) à 37°C pendant 20h. L'ADN plasmidique est ensuite extrait avec des colonnes QIAprepR 96 Plate et recueillis dans des plaques de 96 puits avec le kit Mini prep de Qiagen.

2.18 Séquençage des ADN plasmidiques

La réaction de séquençage a été faite au laboratoire en utilisant les amorces universelles M13 forward et M13 reverse qui sont situés sur les vecteurs de clonage et le kit Sanger ABI PRISM Big Dye Terminator version 3.0/3.1 de Perkin Elmer. La lecture

des séquences a été faite à Institut de Recherche en Immunologie et Cancérologie (IRIC) de l'université de Montréal sur séquenceur capillaire ABI 370 Analyser.

2.19 Assemblage et analyse des séquences

Des outils bioinformatiques ont été utilisés pour l'analyse des différentes séquences. Les séquences sont assemblées grâce au programme read2phrap qui utilise Phred/Phrap et analysées par FASTA (Pearson, 2000) et BLAST (Basic Local Alignment Search Tools) (Altschul*, et al.*, 1997), Clustal W (Thompson*, et al.*, 1994) et pour la visualisation nous avons utilisé l'environnement GDE (Smith*, et al.*, 1994).

2.20 Alignement des séquences ADNc et génomiques des gènes mitochondriaux

Pour identifier des sites d'éditions probables, les séquences génomiques et ADNc de chacun des dix gènes annotés et de gènes non identifiés ont été alignées en utilisant Muscle (Edgar, 2004) et ont été visualisées sous environnement GDE.

2.21 Alignement des séquences protéiques déduites de la séquence ADNc des gènes mitochondriaux de *D. papillatum* et d'eucaryotes

Les séquences protéiques des gènes ont été déduites des séquences ADNc et ont été alignées en utilisant Muscle (Edgar, 2004) et visualisées sous environnement GDE.

2.22 Recherche de séquences conservées dans les modules génomiques de *cox1* chez les diplonémides

Nous avons utilisé le site http://weblogo.berkeley.edu/logo.cgi pour obtenir des représentations graphiques de la conservation des séquences dans les régions 5' et 3' des modules génomiques de *cox1* chez *D. ambulator*, *D.* sp. 2, *D. papillatum* et *R. euleeides* (Figures 2, 3, annexe)

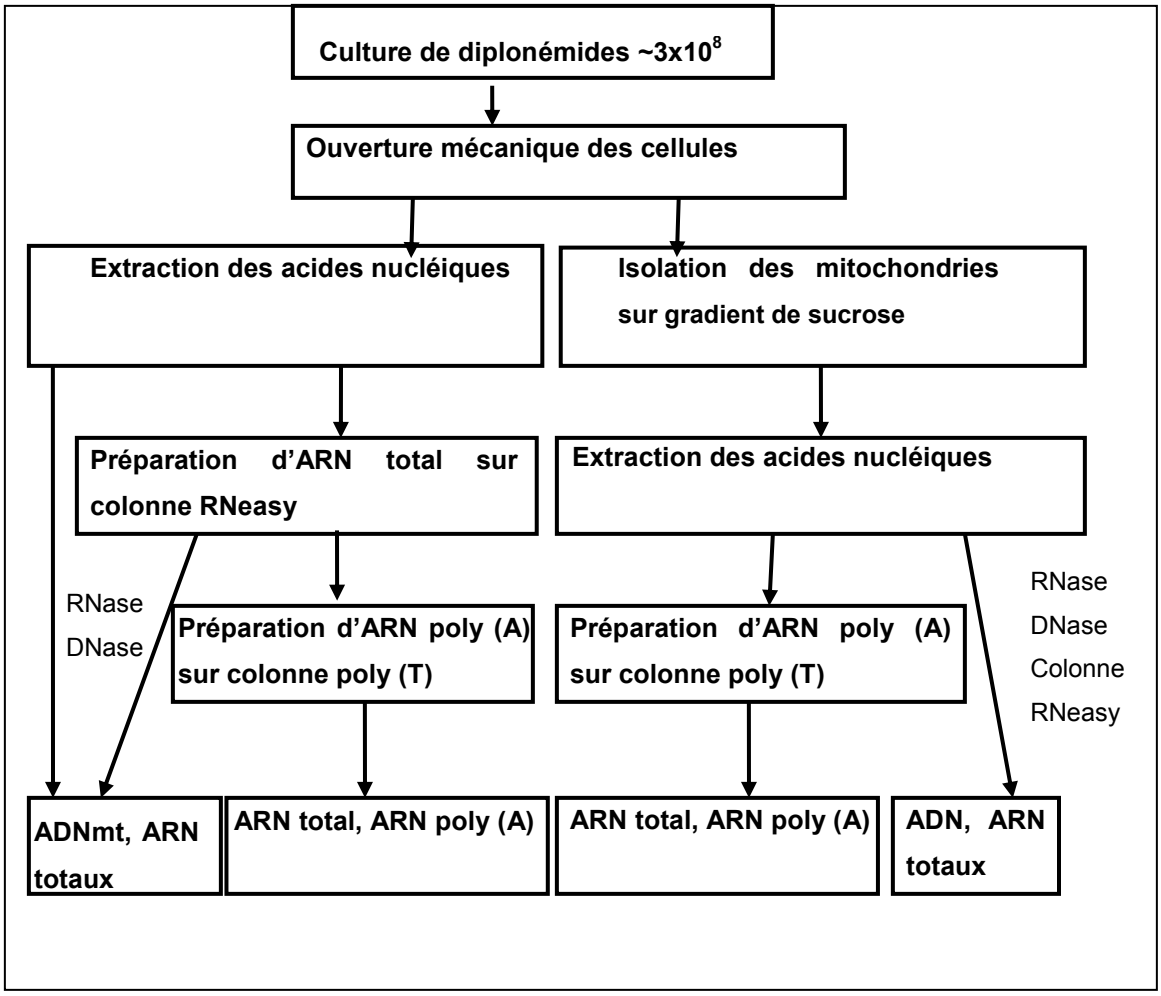


Figure 16. Étapes de préparation de l'ADN total, ARN total et ARN poly (A) mitochondriaux des diplonémides.

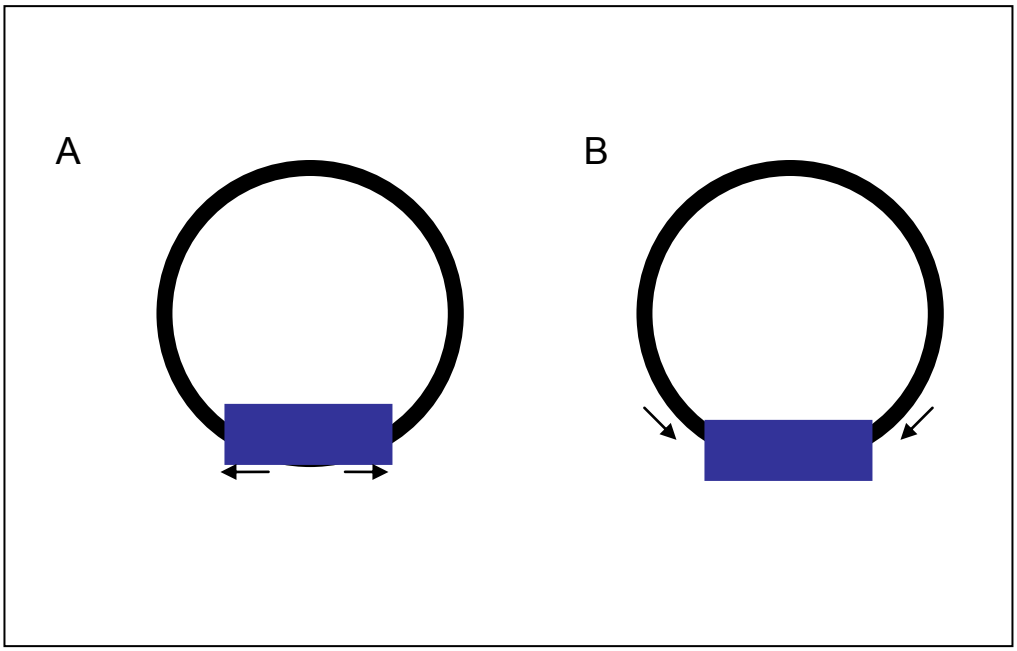


Figure 17. Stratégies d'amplification de chromosomes et modules de *cox1*.
A. amorces divergentes sur le module pour amplifier le chromosome.
B. amorces convergentes dans la région flanquante pour compléter le module.

Figure 18. Stratégies d’amplification de chromosomes portant plusieurs modules de *cox1* de diplonémides. Les amorces spécifiques pour *D. ambulator* (da16, da21) situées sur les modules 3 et 6, pour *D.* sp. 2 (ds12, ds1) situées sur les modules 1 et 4 et pour *R. euleeides* (re27, re15) situées sur les modules 6, 9.

Figure 19. Stratégie d’amplification de longs ARN antisens. Les amorces spécifiques da14 pour *D. ambulator*, ds34 pour *D.* sp. 2 et re27 pour *R. euleeides* ont été utilisées pour la RT.

Tableau IV. Amorces pour amplifier des ARN antisens de gènes mitochondriaux de *D. papillatum* (voir les séquences page 176 Table S9 ou page 230 Table 1).

Gène	Amorce RT	Amorces PCR	Taille de l'ARN antisens attendu (pb)
atp6	dp56	dp56+dp57	197
	dp167	dp167+dp57	324
cob	dp49	dp49+dp50	371
cox2	dp59	dp59+dp62	280
nad5	dp134	dp134+dp55	458
	dp134	dp134+dp173	842
	dp164	dp164+dp173	241
nad7	dp114	dp114+dp97	431
	dp114	dp114+90	337
	dp93	dp93+dp97	427
	dp93	dp93+90	313
	dp95	dp95+dp97	360
	dp95	dp95+90	246
	dp119	dp119+dp97	295
	dp119	dp119+dp90	181
	dp116	dp116+97	381
	dp116	dp116+90	267
	dp118	dp118+97	216

Tableau V. Amorces pour compléter les extrémités des ARNm mitochondriaux de *D. papillatum* (voir les séquences page 176 Table S9 ou page 230 Table1).

Gène	Amorce RT	Amorces PCR
atp6	dp111	dp111+167
cob	dp50	dp50+171
	dp48	dp48+49
	dp48	dp48+171
	dp170	dp170+171
	dp181	dp181+171
nad4	dp166	dp165+166
nad5	dp55	dp55+164
	dp133	dp333+134
	dp173	dp173+172
	dp174	dp174+172
	dp174	dp174+164
nad7	dp90	dp90+91
	dp92	dp92+118
	dp94	dp94+116
	dp97	dp97+96
	dp102	dp102+95
	dp115	dp115+101
	dp119	dp119+97
	dp175	dp175+120
	dp175	dp175+96
rnl	dp71	dp71+179
	dp72	dp72+179
	dp180	dp180+179

Chapitre 3. Résultats

1 Article 1. Evolutionarily conserved *cox1* trans-splicing without cis motifs

1.1 Introduction à l'article

Cet article relève du premier objectif de mon projet et vise à répondre à deux questions principales :

- Le gène mitochondrial *cox1* est-il fragmenté et édité seulement chez *D. papillatum*

ou chez d'autres diplonémides?

- Si la fragmentation du gène *cox1* est conservée chez les diplonémides, y'a t-il des séquences conservées à travers les espèces aux jonctions et dans les régions flanquantes des modules qui dirigent l'épissage en *trans* ou guident l'édition?

Pour investiguer la fragmentation et l'édition de *cox1* chez les diplonémides, nous avons séquencé tous les modules génomiques de *cox1* de trois autres espèces de diplonémides appartenant aux deux genres : *Diplonema ambulator*, *Diplonema* sp.2 et *Rhynchopus euleeides*. Les résultats montrent que *cox1* est fragmenté en neuf modules chez tous les diplonémides étudiés. Les séquences génomiques de chaque espèce ont ensuite été comparées à la séquence ADNc de *cox1* ce qui a permis de conclure qu'il est édité chez les trois diplonémides étudiés exactement comme chez *D. papillatum.*

Pour répondre à la deuxième question qui porte sur des séquences conservées potentielles nous avons testés si des introns discontinus seraient à la base de l'épissage en *trans* chez les diplonémides.

Une analyse *in silico* sur la séquence des modules génomiques de *cox1* avait comme but d'identifier des signatures d'introns de tous types, notamment de groupe I, II, impliqués dans le splicéosome ou ARNt adjacent aux modules de *cox1*. Nous avons ainsi déterminé s'il y a une complémentarité de séquences entre les régions flanquantes

de deux modules voisins et s'il existe des résidus conservés dans les modules au sein d'une espèce et à travers les espèces.

Ces analyses ont exclu la présence d'introns traditionnels ainsi qu'une complémentarité entre les régions flanquantes de deux modules voisins et des résidus conservés. Nous avons donc conclu que l'épissage en *trans* de *cox1* met en jeu des facteurs *trans* plutôt que *cis*.

Pour cet article j'ai réalisé toutes les expériences de séquençage des modules génomiques de *D. ambulator, D.* sp. 2 et *R. euleeides* et j'ai analysé les résultats. J'ai rédigé l'article sous la supervision de ma directrice de recherche. Les analyses *in silico* ont été effectuées par les Drs Turcotte et Burger.

Article 1. Evolutionarily conserved *cox1* trans-splicing without cis motifs
Kiethega GN, Turcotte M and Burger G (2011) *Molecular Biology and Evolution* **28** (9):2425-2428

1.2 Evolutionary Conserved *cox1* Trans-splicing Without cis-Motifs

Georgette N. Kiethega[1], Marcel Turcotte[2] and Gertraud Burger[1,3]

[1]Department of Biochemistry, Université de Montréal, Montreal, Canada
[2]School of Information Technology and Engineering, University of Ottawa, Ottawa, Canada
[3]Robert-Cedergren Centre for Bioinformatics and Genomics, Université de Montréal, Montreal, Canad

***Corresponding author**: Gertraud Burger

Associate editor: Andrew Roger

Keywords: multi-chromosome mtDNA, gene fragmentation, U-insertion RNA editing, diplonemids (Euglenozoa), phylogeny, pattern search

Running title: Fragmented Gene Structure of *cox1* in Diplonemids

1.2.1 Abstract

In the protist *Diplonema papillatum* (Diplonemea, Euglenozoa), mitochondrial genes are systematically fragmented with each non-overlapping piece (module) encoded individually on a distinct circular chromosome. Gene modules are transcribed separately and precursor transcripts are assembled to mature mRNA by a trans-splicing process of yet unknown mechanism. Expression of the *cox1* gene that consists of nine modules, also involves RNA editing by which six uridines are added between Modules 4 and 5. Here we investigate whether the unusual features of *cox1* are shared by all Diplonemea and what the mechanism of trans-splicing might be. We examine three additional species representing both Diplonemea genera, namely *D. papillatum* described before, and *D. ambulator*, *Diplonema sp. 2*, and *Rhynchopus euleeides* and discover that in all Diplonemea the *cox1* gene is discontinuous and split up into nine modules that each reside on a distinct chromosome. Positions of gene breakpoints vary by up to two

nucleotides. Further, all taxa have six non-encoded uridines inserted in *cox1* mRNA at exactly the same position as *D. papillatum*. In silico searches do not detect signatures of introns known to engage in trans-splicing, in particular Group I, Group II, spliceosomal, and transfer RNA introns. Nor did we find statistically significant reverse-complementary motifs between adjacent modules and their flanking regions, or residues conserved within or across species. This provides compelling evidence that trans-splicing in Diplonemea mitochondria does not rely on sequence elements in *cis* but rather proceeds by a mechanism employing matchmaking *trans* factors such as RNAs or proteins.

1.2.2 Introduction

The Diplonemea (or "diplonemid") *Diplonema papillatum* possesses a most unconventional mitochondrial genome (Maslov*, et al.*, 1999, Marande*, et al.*, 2005). Instead of a single chromosome type, *D. papillatum* mtDNA is composed of a hundred or so distinct circular molecules of two sizes (Class A, 6 kbp; Class B, 7 kbp). Intriguingly, these chromosomes do not encode multiple genes, but rather single non-overlapping gene pieces (modules) of ~60-340 bp that are transcribed individually and then assembled to mRNA (Marande & Burger, 2007, Vlcek*, et al.*, 2010). The *cox1* transcript seems to be the only RNA undergoing editing. Here we address two questions. First, are unorthodox genome architecture, gene structure, and gene expression peculiarities of *D. papillatum* or rather shared by all diplonemids? Second, does trans-splicing rely on a known intron-splicing machinery and are any conserved sequence motifs involved?

1.2.3 Results and Discussion

We characterized the *cox1* gene and its transcript from three additional diplonemid species, *D. ambulator*, *Diplonema sp. 2*, and *Rhynchopus euleeides* (Roy*, et al.*, 2007). For methods, see Methods section and for primers used in RT-PCR see supplementary table S6, Supplementary Material. Sequences were deposited in GenBank under acc.

nos. JF698650-80 (supplementary table S5, Supplementary Material). The *cox1* cDNA sequences obtained are contiguous and align perfectly with that from *D. papillatum* (supplementary fig. S1, Supplementary Material). While the translation code is the same in all diplonemids (including UGA=W), codon frequencies and A+T-content differ drastically (supplementary tables S1A-D, Supplementary Material).

Mitochondrial chromosomes of the diplonemids studied here are circular and of two size classes as in *D. papillatum* (Marande & Burger, 2007), but sizes range from 4.5-9 kbp (supplementary table S2, Supplementary Material). For each species, we sequenced the chromosomes that carry the *cox1* portions corresponding to Modules 4, 5 and 9 in *D. papillatum*, and found that these coding regions are each encoded separately by a distinct mitochondrial chromosome as well (supplementary table S3, Supplementary Material). Non-coding chromosome regions share substantial sequence-similar stretches of up to 4 kbp (Supplementary fig. S2, Supplementary Material). Multi-chromosome mtDNAs seem to predominate in all euglenozoan groups: kinetoplastids (Lukes*, et al.*, 2002), euglenids (*Euglena gracilis* (Talen*, et al.*, 1974, Yasuhira & Simpson, 1997, Gray*, et al.*, 2004, Spencer & Gray, 2011) and *Peranema cantuscygni* (Roy*, et al.*, 2007)), and diplonemids described here.

We pinpointed the *cox1* gene modules by sequencing all genomic coding regions plus adjacent noncoding stretches to compare these with the cDNA. Always, *cox1* is split into nine pieces ranging from ~90 to 260 bp as in *D. papillatum* (fig. 1A; for more details, see Results in Supplementary Material). Breakpoint positions occur in highly and weakly conserved coding regions and are astoundingly congruent across diplonemids (supplementary fig. S1, Supplementary Material). Some boundaries are at exactly the same position, others slightly shifted (fig. 1B and C; supplementary fig. S3, Supplementary Material). The *cox1* fragmentation pattern being shared by all examined species allows two extrapolations. First, not only *cox1* but probably all mitochondrial genes are fragmented and trans-spliced in diplonemids as reported for *D. papillatum* (Vlcek*, et al.*, 2010). Second, gene fragmentation and trans-splicing emerged in the common ancestor of diplonemids.

In *D. papillatum*, U-insertion RNA editing occurs between *cox1* Modules 4 and 5. At that position, six nonencoded Us are also encountered in the other three species (fig. 1D). Surprising is the strict conservation of the editing pattern across diplonemids, whereas in kinetoplastids the pattern varies considerably, even within the same genus (Feagin, 1990). (For more details, see Discussion in Supplementary Material).

The Cox1 protein sequence of diplonemids is fairly well conserved within the group, but highly divergent compared to other taxa. Regions most variable between diplonemids coincide with those of general low conservation; further, there is no indication of sequence constraints imposed by module junctions (supplementary fig. S4, Supplementary Material). RNA editing of the diplonemid *cox1* transcripts specifies V-F-S, I-F-S, and L-F-S (fig. 2). Without the added Us, the diplonemid proteins would lack positions otherwise invariably present, although moderately conserved, in Cox1 of other organisms. In the protein's tertiary structure (available for *Bos taurus*), the corresponding tripeptide is located in a loop that interact with a second more C-terminal loop. Interactions involve positively charged residues in Loop 1 and hydrophobic and/or small residues in Loop 2. Remarkably, the situation is inverse in the diplonemid proteins (Marande, 2007). We speculate that the ancestral diplonemid *cox1* gene lost nucleotides at the boundary of Module 4 and/or 5, which was patched secondarily through filling-in bases by a pre-existing uridylyl transferase. Since the resulting amino acids were unable to interact with residues of Loop 2, these latter underwent compensatory substitutions to restore the protein's function-critical tertiary structure.

Known trans-splicing relies on non-contiguous classical introns, notably Group I, Group II, and tRNA (or "archaeal") introns in the case of organelle genes and spliceosomal introns for nuclear genes (Bonen & Vogel, 2001). Group I, Group II, and spliceosomal introns possess distinctive nucleotides and secondary and tertiary structures (Breathnach & Chambon, 1981, Bonen, 1993), which, however, we did not detect for diplonemid *cox1*. Nor could we uncover any nucleotides conserved across all modules or flanking regions of a given species or across all species for a given module. Instead, multiple sequence alignments of homologous regions show a random

distribution of nucleotides (supplementary fig. S5, Supplementary Material). This eliminates the possibility of classical introns mediating trans-splicing in diplonemid mitochondria.

We investigated whether two neighboring modules or their flanking regions can interact with one another via sequence complementarity to form a helix-loop-helix structure typical for tRNA introns (for more details on in silico analyses, see Methods, Supplementary Material). Alternatively, sequence complementarity without this particular 2D structure could point to a new splicing mechanism relying on sequence elements in *cis*. In silico search for sequence complementary assumed a "seed" region of at least six consecutive nucleotides interacting via canonical or wobble base pairing. The dynamic-programming-based algorithm compared for each junction all fixed-length segments from the upstream module with those from the downstream module, and an interaction plot was generated for each junction. Presuming the same trans-splicing mechanism at all junctions, we expected sequence-complementary segments at the same relative location for all junctions. Yet, no such shared position was found. Therefore, we analyzed the minimum offset to circumscribe a common region of complementarity, and this for the original dataset and for simulated datasets of identical dinucleotide composition. Again, the result was negative: there was no significant difference in occurrences of sequence-complementary regions between the original and simulated datasets (supplementary table S4, Supplementary Material). These results strongly suggest that *cis*-elements are not involved in trans-splicing of diplonemid *cox1* and corroborate our hypothesis of match-making trans-factors. Such factors could also ensure, when required, RNA editing, a task being achieved by gRNAs in trypanosome mitochondria (Stuart*, et al.*, 2005). Work is in progress to test in *D. papillatum* if the hypothetical *trans*-acting matchmakers are RNAs. Because preliminary data do not indicate trypanosome-like gRNAs, guiding proteins are a sensible alternative to consider.

Finally, we constructed a phylogenetic tree with Cox1 protein sequences from diplonemids, other euglenozoans, *Naegleria*, and several slow-evolving taxa as

outgroups (for more details on phylogenetic analyses, see Methods, Supplementary Material). Figure 3 confirms that diplonemids are the sister group of kinetoplastids and that euglenids diverge basally to the split of the two former groups, as observed in nuclear phylogenies (Busse & Preisfeld, 2002, Simpson & Roger, 2004, von der Heyden*, et al.*, 2004, Breglia*, et al.*, 2007). However, the topology within diplonemids differs. In most nuclear-gene-based trees, the *Diplonema* genus is monophyletic (with modest statistical support), whereas in the mitochondrial tree *Diplonema* embraces *R. euleeides* (with high support). This indicates that Diplonemea consists of three rather than two genera, notably *Rhynchopus*, the *D. papillatum*- *D. ambulator*-clade, and one represented by *D.* sp. 2 (this topology is shared by one nuclear phylogeny (von der Heyden*, et al.*, 2004)). To reconcile the conflicting molecular and morphology-based classification, it would be worthwhile to re-examine the morphological characters that are traditionally used for distinguishing *Diplonema* and *Rhynchopus*.

1.2.4 Legend of figures

Figure 1. Diplonemid *cox1* module junctions. **A,** module sizes in *D. papillatum*. White boxes indicate modules residing on class-B chromosomes. **B**, a highly conserved and **C**, a most divergent junction (see supplementary fig. S3**,** Supplementary Material**)**. Da, *D. ambulator*; Dp, *D. papillatum*; Ds, *Diplonema*. sp. 2; Re, *R. euleeides*. mod4-genomic, etc., genomic sequence around boundaries of *cox1*-Module 4, etc. Upper case bold, module (coding region); lower case, module-flanking region (not coding); shading, uncertain junction positions. **D**, RNA editing between Modules 4 and 5. Italics, non-encoded nucleotides.

Figure. 2. Multiple sequence alignment of Cox1 proteins. The region shown corresponds to residues 251-276 in *Bos taurus*.

Figure. 3. Phylogeny based on deduced Cox1 protein sequences. The maximum likelihood tree was constructed using the WAG + Γ model and eight discrete rate categories. Statistical support values (from 100 bootstrap replicates) >50 are shown.

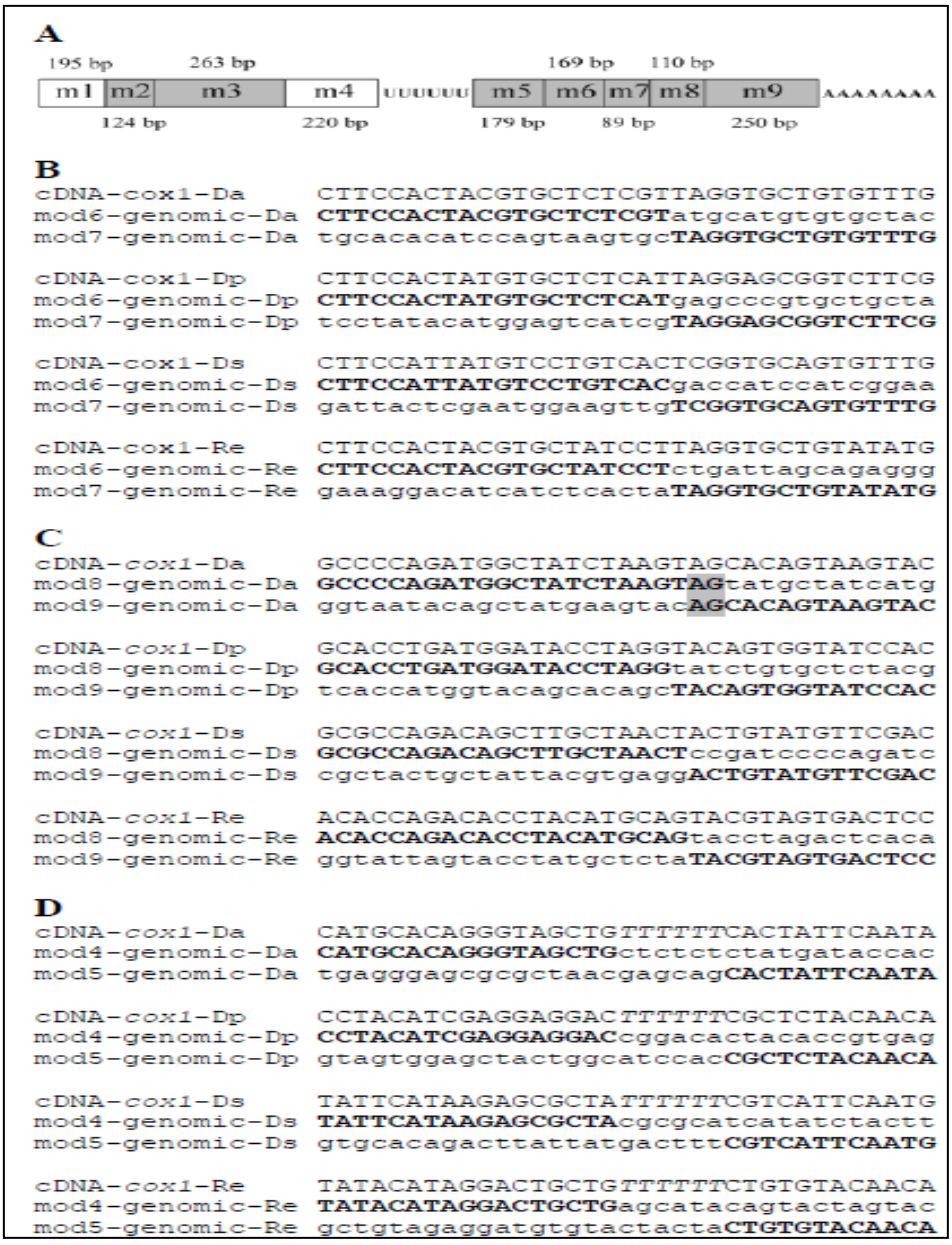


Figure 1. Diplonemid *cox1* module junctions.

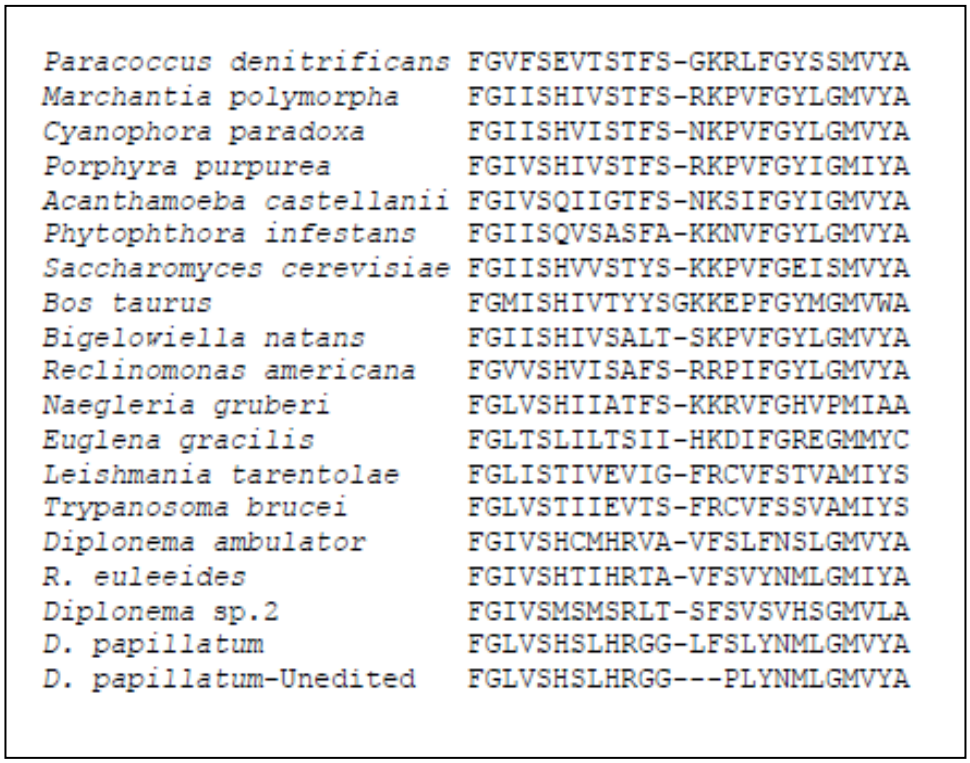


Figure 2: Multiple sequence alignment of Cox1 proteins.

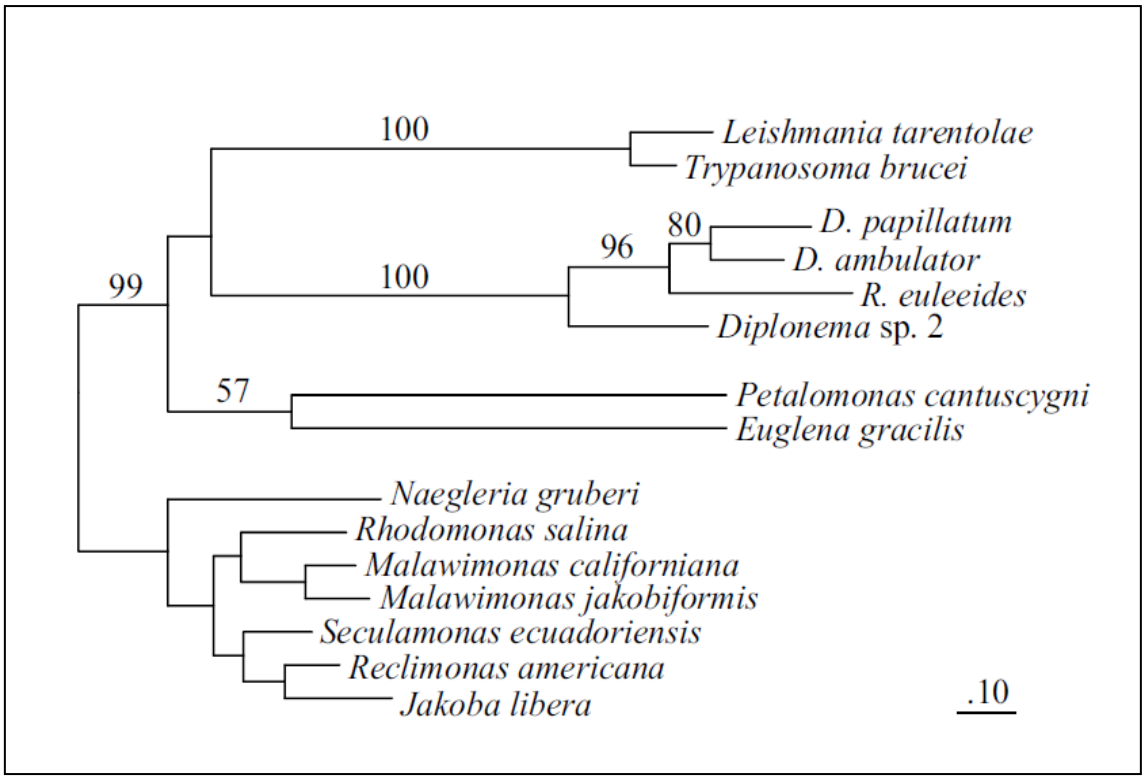


Figure 3. Phylogeny based on deduced Cox1 protein sequences.

1.2.5 Acknowlegdments

We thank S. Teijeiro (Université de Montréal) for generating most cDNA data and providing technical advice and assistance, Yifei Yan (Université de Montréal) for sharing his data on the *D. papillatum cox1* Module 8-Module 9 junctions – data he generated in the context of his Master's thesis -, and Pavel Poliak (University of South Bohemia) for generating data on the *R. euleeides cox1* Module 8-Module 9 junction during an internship in GB's laboratory. B. Franz Lang (Université de Montréal) kindly conducted the phylogenetic analysis and provided helpful comments to the manuscript.

1.2.6 Funding

This work was supported by grants from the Canadian Institute for Health Research (CIHR, grant MOP-79309; GB) and the National Science and Engineering Research Council, Canada (NSERC, grant 250909-2006; MT), and a Ph. D. scholarship from the Programme Canadien de Bourses de la Francophonie (PCBF scholarship; GNK).

1.2.7 References

Bonen, L (1993) Trans-splicing of pre-mRNA in plants, animals, and protists. *FASEB Journal* **7**:40-46.

Bonen L and Vogel J 2001. The ins and outs of group II introns. *Trends in Genetics* **17**:322-331.

Breathnach R, and Chambon P (1981) Organization and expression of eucaryotic split genes coding for proteins. *Annual Review of Biochemistry* **50**:349-383.

Breglia SA, Slamovits CH, and Leander BS (2007) Phylogeny of phagotrophic euglenids (Euglenozoa) as inferred from hsp90 gene sequences. *Journal of Eukaryotic Microbiology* **54**:86-92.

Busse I, and Preisfeld A (2002) Phylogenetic position of *Rhynchopus* sp. and *Diplonema ambulator* as indicated by analyses of euglenozoan small subunit ribosomal DNA. *Gene* **284**:83-91.

Feagin JE (1990) RNA editing in kinetoplastid mitochondria. *Journal of Biology and Chemistry* **265**:19373-19376.

Gray MW, Lang BF and Burger G (2004) Mitochondria of protists. *Annual Review Genetics* **38**:477-524.

Lukeš J, Guilbride DL, Votypka J, Zikova A, Benne R, and Englund PT (2002) Kinetoplast DNA network: evolution of an improbable structure. *Eukaryotic Cell* **1**:495-502.

Marande, W (2007) Structure et expression des gènes mitochondriaux de *Diplonema papillatum*. Montreal (Canada): Biochemistry Department, Université de Montréal. 98p.

Marande W and Burger G (2007) Mitochondrial DNA as a genomic jigsaw puzzle.

Science **318**:415.

Marande W, Lukeš J and Burger G (2005) Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids. *Eukaryotic Cell* **4**:1137-1146.

Maslov DA, Yasuhira S and Simpson L (1999) Phylogenetic affinities of *Diplonema* within the Euglenozoa as inferred from the SSU rRNA gene and partial COI protein sequences. *Protist* **150**:33-42.

Roy J, Faktorova D, Benada O, Lukeš J and Burger G (2007a) Description of *Rhynchopus euleeides* n. sp. (Diplonemea), a free-living marine euglenozoan. *Journal of Eukaryotic Microbiology* **54**:137-145.

Roy J, Faktorova D, Benada O, Lukeš J and Burger G (2007b) Unusual mitochondrial genome structures throughout the Euglenozoa. *Protist* **158**:385-396.

Simpson AG and Roger AJ (2004) Protein phylogenies robustly resolve the deep-level relationships within Euglenozoa. *Molecular Phylogenetic and Evolution* **30**:201-212.

Spencer DF, and Gray MW (2011) Ribosomal RNA genes in *Euglena gracilis* mitochondrial DNA: fragmented genes in a seemingly fragmented genome. *Molecular Genetics and Genomics* **285**:19-31.

Stuart K.D, Schnaufer A, Ernst NL and Panigrahi AK. (2005) Complex management: RNA editing in trypanosomes. *Trends in Biochemical Sciences* **30**:97-105.

Talen JL, Sanders JP, and Flavell RA (1974) Genetic complexity of mitochondrial DNA from *Euglena gracilis*. *Biochimica et Biophysica Acta* **374**:129-135.

Vlcek C, Marande W, Teijeiro S, Lukeš J, and Burger G (2010) Systematically fragmented genes in a multipartite mitochondrial genome. *Nucleic Acids Research* **39**:979-988.

von der Heyden S, Chao EE, Vickerman K and Cavalier-Smith T (2004) Ribosomal RNA phylogeny of bodonid and diplonemid flagellates and the evolution of euglenozoa. *Journal of Eukaryotic Microbiology* **51**:402-416.

Yasuhira S and Simpson L (1997) Phylogenetic affinity of mitochondria of *Euglena gracilis* and kinetoplastids using cytochrome oxidase I and hsp60. *Journal of Molecular Evolution* **44**:341-347.

1.2.8 Supplementary material

1.2.8.1 Table of content

1.2.8.2 Introduction

The most eccentric mitochondrial genome architecture and genes expression has long been attributed to kinetoplastid flagellates (Euglenozoa; for a review see (Foldynova-Trantirkova*, et al.*, 2005). In contrast to conventional mitochondrial genomes consisting of a single chromosome (in multiple copies), kinetoplastid mtDNA is composed of hundreds of distinct minicircles and one type of maxicircle. Maxicircle chromosomes encode typical mitochondrial genes, but in an encrypted fashion. After transcription, certain uridines are deleted and others inserted to reconstitute translatable mRNAs. This RNA editing process is guided by small RNAs (gRNAs) that serve as template for nucleotide additions and removals at specific sites, and these molecules are encoded by minicircle chromosomes. The closest relatives of kinetoplastids are diplonemids, a group of free-living seawater bi-flagellates comprising two genera: *Diplonema* and *Rhynchopus*. One member of the diplonemids, *Diplonema papillatum,* has been studied in detail at the molecular biology level and revealed a quite unconventional mitochondrial genome as well (Maslov*, et al.*, 1999, Marande*, et al.*, 2005, Marande & Burger, 2007). *D. papillatum* mtDNA is composed of a hundred or so distinct circular chromosomes of two size classes that encode single non-overlapping gene pieces (modules) of ~50-300 bp, which are transcribed individually and then assembled together at the RNA level to form a mature mRNA. The *cox1* transcript undergoes U insertion editing exactly at the boundary of two modules (Marande & Burger, 2007).

Here we address two questions. The first is whether the unorthodox genome architecture, gene structure, and gene expression are particularities of *D. papillatum* or rather shared by all diplonemids, and should these features be found in the entire group, in how far they would differ. Therefore, we investigate three additional species, representing both diplonemid genera, by sequencing the genomic regions coding for *cox1* and the corresponding mature transcript. The second question is whether trans-splicing relies on known intron-splicing machineries or whether other conserved

sequence motifs in and around gene modules may be involved. For that, we employed sensitive pattern search algorithms.

1.2.8.3 Results

Chromosome size determination

Chromosome shape and size of the diplonemids studied here was determined by PCR using two proximate but divergent primers, and by partial DNase digestion (linearizing circular molecules). In addition, electron microscopy measurements are available for *R. euleeides* mitochondrial chromosomes. Each diplonemid has circular chromosomes of two size classes, ~4.8 + ~5 kbp in *D. ambulator*, ~5 + ~10 kbp in *Diplonema* sp. 2, ~7 + ~7.5 kbp in *R. euleeides*, compared to 6 + 7 kbp in *D. papillatum* as reported before (supplementary tables S2, S3). In *R. euleeides*, PCR generated in addition to the genuine products also a shorter (5 kbp) and a longer amplicon (12 kbp). We believe that these latter two products are artefacts arising by the highly repetitive sequence of *R. euleeides* mitochondrial chromosomes, as seen by DNA sequencing.

Sequencing of diplonemid mitochondrial chromosomes

For each of the three diplonemids investigated here, we sequenced three mitochondrial chromosomes, notably the ones that include the *cox1* portions corresponding to Modules 4, 5 and 9 in *D. papillatum*. This was achieved by PCR using primers that anneal with the gene modules but point 'outwards', followed by cloning of the resulting products and shotgun sequencing of the clones. Overall A+T content is typical for mtDNA except in *D. papillatum* where it is exceptionally low (supplementary tables S1A-D).

In all species, the sequenced chromosomes comprise a ~90-260 bp *cox1* coding region (gene module) that is flanked by short sequence unique to a given chromosome (~50 bp in *D. papillatum*), and the 'constant' region with high sequence identity within a

chromosome class (supplementary fig. S2). While in *D. papillatum*, we were able to precisely delineate the unique module-flanking regions and distinguish them from the constant region shared by all members of a chromosome class; this was not always possible for the other three diplonemids, where the body of sequence data is much smaller.

Confirmation of fragmented *cox1* gene structure

We tested whether in the here studied taxa a contiguous *cox1* gene copy is present in their mtDNA or nuclear DNA. Yet, PCR did not yield products when using primers spanning a 1000 bp-long region on cDNA and corresponding to Modules 2-6 in *D. papillatum* (data not shown). In order to characterize more precisely the discontinuous *cox1* gene structure of the three diplonemids, we determined the sequence of all genomic coding regions plus adjacent non-coding stretches. For that we designed primers annealing ~100 bp upstream and downstream of module-flanking regions (based on the chromosome sequences), amplified the corresponding stretches by PCR, and cloned and sequenced the products. The sequences are shown in supplementary fig. S3.

The *cox1* cDNA sequence and precise mapping of gene modules

The *cox1* cDNA sequence was determined for all diplonemids and the multiple sequence alignment is shown in supplementary fig. S1. Codon usage differs drastically, with certain triplets absent from the *cox1* coding regions of one species but present in another (supplementary tables S1A-D).

Modules were pinpointed by aligning genomic and cDNA sequence. Gene modules are most certainly not overlapping, although some junctions could be mapped only with a precision of +/- 4 nt. precisely. Two junctions in *D. papillatum* and *R. euleeides* have been determined with additional pre-mRNA data (see Methods). The junctions are indicated by arrows in supplementary fig. S1.

RNA editing

U-insertion RNA editing apparently takes place in all diplonemids, since sequence comparison of *cox1* cDNA and genomic DNA shows essentially the same incongruence (fig. 1D, main text). Only *Diplonema* sp. 2 has three Ts upstream of what we consider the 5' end of the genomic module 5. The exact start position of this module and the number of Us added post-transcriptionally remain to be confirmed by sequencing *cox1* precursor transcripts.

The *cox1* protein sequence

In the Cox1 proteins of *Bos taurus* and *Paracoccus denitrificans* for which tertiary structure data are available, the tripeptide corresponding to those arisen by RNA editing in diplonemids (residues 264-266 in *B. taurus*) is located in a loop (Loop 1) and engages in 3D interaction with a second loop (Loop 2) further toward the C-terminus (residues 328-330 in *B. taurus*). Residues in Loop 1 are positively charged (K and R), while those in Loop 2 contains hydrophobic (L, W, Y, P) and/or small (G) residues. Remarkably, the situation is inverse in the diplonemid proteins: here, small (S) and hydrophobic (L, F, V, I) residues make up the first loop and positively charged amino acids (Arg) the second (Marande, 2007). An evolutionary scenario that may have led to this situation is discussed in the main text.

In silico search of sequence motifs involved in RNA trans-splicing

Pattern searching explored a wide range of values for the main parameters. We considered complementary segments of length 6 or 7; we compared segments with or without Wobble base pairs, and with or without mismatch; and we considered 6, 7 or 8 junctions. For each of the 24 value combinations, we calculated the minimum size of a region comprising sequence-complementary segments. The results for each of the four species are presented in supplementary tables S4A-D. For all but two combinations, it was always possible to find a region of the same size or smaller size in the simulated data. For the two combinations where this is not true, the corresponding values are very

close. But more importantly, these are cases where our simulations were not run to completion, and the number of replicates is relatively low; (for three of the four species, the simulations were completed after 8 days on an IBM p5-595, using 15 out of 64 cores, each core with 1.9 GHz p5 and 128 GB RAM; at that time point, the simulations for the fourth species had been stopped prior to completion). It is highly likely that as small or smaller values would have been found if the simulations were ran for a longer time.

1.2.8.4 Discussion

Sequence determination of *cox1* cDNAs revealed that diplonemid mitochondrial transcripts are poly-adenylated. This is rare but not unheard of for organelle mRNAs. The other eukaryotic groups where mitochondrial transcripts carry poly (A) tails are Metazoa, Apicomplexa, dinoflagellates, and kinetoplastids (Anderson*, et al.*, 1981, Gillespie*, et al.*, 1999, Chaput*, et al.*, 2002, Gagliardi*, et al.*, 2004).

We observe U-insertion RNA editing in *cox1* of all diplonemids and speculated that in the evolutionary past, some nucleotides were lost at the boundary of *cox1* Module 4 or 5 in the diplonemid ancestor. This loss may have been patched secondarily through filling-in bases by a recruited uridylyl transferase. But why nucleotides are added via RNA editing rather than recruiting them from the adjacent modules is puzzling. It is also surprising to observe a strict conservation of the editing pattern across diplonemids, while the diplonemids' sistergroup, the kinetoplastids, are quite variable in this respect even in the same genus (Feagin, 1990). Further, in kinetoplastid mitochondria, the number of editing sites in fluctuates strongly between genes of the same species. For example, in *Trypanosoma brucei,* the *cox3* pre-mRNA is edited at nearly 600 positions with ~560 U-addition and 40 U-deletion sites (as much as 50% of the mRNA sequence is contributed by RNA editing (Feagin, 1990)). In contrast, *cox2* RNA editing involves only three events and exclusively insertions (four Us in total (Benne*, et al.*, 1986)). Limited U-insertion is likely the ancestral form of RNA editing in Euglenozoa and has emerged prior to the divergence of kinetoplastids and diplonemids.

Since RNA editing in diplonemid mitochondria occurs exactly at a module boundary, we speculate that RNA editing and trans-splicing are interlinked. Intriguingly, coupled RNA editing and trans-splicing has also been detected in dinoflagellate mitochondria (Waller & Jackson, 2009). The c*ox3* gene of *Karlodinium micrum* consists of two pieces that are transcribed separately and then catenated, and non-encoded nucleotides, here as are found at the junction. Also in this instance, hallmarks of traditional introns and sequence complementarity around the junction appear to be absent (R. Waller, pers. communication).

1.2.8.5 Methods

For sequences deposited in GenBank, see supplementary table S5.

Strains, culture and DNA extraction

D. ambulator (ATCC 50223) *Diplonema* sp. 2 *(*ATCC 50224) and *R. euleeides* (ATCC 50226) were obtained from the American Type Culture Collection. The organisms were cultivated axenically at room temperature in ATCC medium 1405. Mitochondrial DNA was extracted by TRIzol (Invitrogen), whereby the mtDNA circles remain in the aqueous phase.

DNase I digestion

One µg of mtDNA was digested with 1×10^{-3}, 2×10^{-3}, and 4×10^{-3} units of DNase I (Fermentas) respectively. The reaction takes place at 4º C for 6 min and then stopped by addition of EDTA (50 mM).

PCR and RT-PCR

Supplementary table S6 lists all primer pairs used. The *cox1* chromosomes and genomic modules were amplified by the polymerase chain reaction (PCR) kit of

TAKARA Bio Inc. as recommended by the supplier. PCR amplification included 25 cycles for chromosomes of *D. ambulator* and *Diplonema* sp. 2, and 35 cycles for chromosomes of *R. euleeides*. Reverse transcriptase followed by PCR (RT-PCR) was performed to generate *cox1* cDNAs from polyA-RNA or total RNA as a template. For this experiment, we used an oligo-dT primer (included in the SMART kit, Clontech) for the reverse transcriptase reaction, and PCR was conducted with the same primer plus a *cox1*-specific primer. The latter primer anneals in a highly conserved gene region as inferred from the alignment of the *cox1* cDNA sequence from *D. papillatum* with genomic sequences from other eukaryotes (Vlcek*, et al.*, 2011). Since the resulting cDNA was partial, two or three additional partial cDNAs were generated that together cover the entire transcript. For these steps, reverse transcription was primed with a *cox1*-specific oligonucleotides, and the PCR reaction proceeded with the same primer plus the SMARTII primer that anneals with the 5'end of the synthesized DNA strand (Frohman*, et al.*, 1988). For *Diplonema sp. 2,* we had to perform several RT-PCR and nested RT-PCR experiments to complete the *cox1* cDNA. To confirm the junctions between *cox1* Modules 8 and 9 in *D. papillatum* and *R. euleeides*, we conducted RT-PCR on a template of RNAs circularized with T4 RNA ligase and using divergent primers that span the two modules.

Cloning and DNA sequencing

Mitochondrial chromosomes were amplified quasi entirely by PCR using primers that anneal within a given module but facing outwards. To close the gap ('heart') between the primers, we used PCR with a second pair of primers annealing with module-flanking regions and facing inwards. Quasi-complete chromosomes were subcloned by generating libraries of mechanically random-broken chromosome fragments, while 'hearts' were cloned in their entire length. Cloning involved end-polishing to obtain blunt-ended DNA fragments, size-fractionation of these fragments on agarose gel and cloning into the EcoRV site of vector pBFL6cat (in-house constructed, small derivative of pBluescript, Stratagene). Some heart-amplicons were sequenced

directly (without cloning) using one of the PCR primers. Complementary DNA was cloned into the vector pDNRlib, a component of the SMART kit (see above). To close sequencing gaps that remained after shotgun sequencing of chromosome random libraries, primer walking was performed. The methods have been described in detail elsewhere (Rodriguez-Ezpeleta*, et al.*, 2009). Plasmid DNA was extracted with the QIAprep 96 Turbo Mini prep Kit (Qiagen). For sequencing reactions, we used the Sanger-dye terminator technology (kit ABI PRISM Big Dye terminator version 3.0/3.1, Perkin-Elmer). Reactions were run on an MJ BaseStation or an ABI 3730 capillary sequencer. We assembled readings with Phred, Phrap (-q 0.9) and Consed (Gordon, 2003) using wrapper scripts developed in-house. Fasta-formatted sequence files with integrated annotations (Masterfile) were generated with Cosmea, a program developed in-house.

All in-house developed tools are described at the URL http://megasun.bch.umontreal.ca/ogmp/ogmpid.html, and are available on request.

Basic sequence analysis and phylogeny

We used the in-house developed tools Flip for translation of nucleotide sequence into protein sequence, and Pepper for codon usage analysis and extracting various genetic elements from the Masterfile. Sequence similarity searches were performed locally with Blast (Altschul*, et al.*, 1990) and Fasta (Pearson, 2000), and remotely - against NCBI's databases - with Blast and psiBlast (Altschul*, et al.*, 1997). Multiple sequence alignments were obtained with Muscle (Edgar, 2004). Alignments were visualized and edited with the Genetic Data Environment (GDE) (Smith*, et al.*, 1994). For phylogenetic tree construction, we used RaxML (Stamatakis, 2006) with the WAG + GAMMA model and eight discrete gamma-rate categories. The statistical support of branches was evaluated by 100 bootstrap replicates.

The GenBank accession numbers of the Cox1 protein sequences used for multiple alignments and phylogeny are as follows: *Acanthamoeba castellanii* (AAD11820); *Allomyces macrogynus* (AAC49234); *Amoebidium parasiticum*

(AAN04062); *Bigelowiella natans* (ADV41804); *Bos taurus* (AAZ17133); *Cyanophora paradoxa* (ADW79204); *Euglena gracilis* (CAA69263); *Jakoba libera* (Lang, Burger unpublished); *Malawimonas californiana* (Lang, Burger unpublished); *Malawimonas jakobiformis* (AAG13707); *Monosiga brevicollis* (AAN28355); *Leishmania tarentolae* (P14544); *Marchantia polymorpha* (P26856); *Naegleria gruberi* (AAG17783); *Nephroselmis olivacea* (AAF03191); *Paracoccus denitrificans* (CAA55033); *Petalomonas cantuscygni* (Burger, Roy unpublished); *Phytophthora infestans* (AAW67076); *Porphyra purpurea* (AAD03116); *Saccharomyces cerevisiae* (CAA09824); *Reclinomonas americana* (AAD11923); *Rhodomonas salina* (AAG17762); *Seculamonas ecuadoriensis* (Lang, Burger unpublished); *Trypanosoma brucei* (AAB59223). Unpublished protein sequences can be obtained on request.

Analysis of protein 3D interaction

The 3D-structure of cytochrome *c* oxidase from *Bos taurus* (2OCC, (Tsukihara & Yoshikawa, 1998)) and *Paracoccus denitrificans* (1AR1, (Ostermeier*, et al.*, 1997)), available in the Protein DataBase (PDB, http://www.rcsb.org/pdb/home/home.do), were visually inspected using Jmol (http://www.rcsb.org/pdb/explore/jmol.do). Amino acid pairs are considered to be in contact if the distance between them is smaller than 6 Å (Miyazawa & Jernigan, 1999).

In silico search of sequence complementarity

Bioinformatics analysis of neighbouring modules was conducted using the R statistical computing environment (R Development Core Team, http://www.R-project.com). A dynamic programming algorithm finds the maximum number of base pairs that can be formed between two segments, allowing for Wobble base pairs and mismatches, but no multi-branch loops. For each junction, this algorithm is applied to all fixed-length segments from the upstream module and all the fixed-length segments from the downstream module. This creates two-dimensional interaction maps, which are superimposed to align all the junctions. Finally, the superimposed interaction maps are

analyzed to find minimum size regions comprising complementary segments for 6, 7 or 8 junctions. For the generation of the simulated data, the parameters of a first-order Markov model were obtained from the original data (one model per species).

Multiple sequence alignments and sequence logos

Multiple alignments of nucleotide and protein sequences were obtained with Muscle (Edgar, 2004). To generate graphical representations of nucleotide conservation at *cox1* module boundaries, we used http://weblogo.berkeley.edu.logo.cgi that produces logos from multiple sequence alignments, where the height of a symbol represents its relative frequency at a given position (Schneider & Stephens, 1990, Crooks*, et al.*, 2004).

Table S1A**.** A+T-content and codon usage[1] of *cox1* cDNA from *D. ambulator*.

A+T = 53.5%

F	TTT 18.2	S	TCT 15.9	Y	TAT 31.6	**C**	**TGT 64.3**
F	**TTC 81.8**	S	TCC 6.8	**Y**	**TAC 68.4**	C	TGC 35.7
L	TTA 12.0	S	TCA 15.9	*	TAA 100	W	TGA 42.9
L	TTG 0.0	S	TCG 4.5	*	TAG 0.0	W	TGG 57.1
L	CTT 7.2	**P**	**CCT 56.3**	**H**	**CAT 65.0**	R	CGT 31.3
L	CTC 14.5	P	CCC 12.5	H	CAC 35.0	R	CGC 0.0
L	**CTA 43.4**	P	CCA 25.0	Q	CAA 28.6	R	CGA 6.3
L	CTG 22.9	P	CCG 6.3	**Q**	**CAG 71.4**	R	CGG 0.0
I	ATT 6.7	T	ACT 38.3	N	AAT 60.0	**S**	**AGT 36.4**
I	ATC 43.3	T	ACC 14.9	N	AAC 40.0	S	AGC 20.5
I	ATA 50.0	T	ACA 38.3	K	AAA 0.0	**R**	**AGA 50.0**
M	ATG 100	T	ACG 8.5	**K**	**AAG 100**	R	AGG 12.5
V	GTT 9.6	A	GCT 38.9	**D**	**GAT 73.3**	G	GGT 34.6
V	GTC 7.7	A	GCC 22.2	D	GAC 26.7	G	GGC 19.2
V	**GTA 59.6**	A	GCA 30.6	E	GAA 37.5	G	GGA 21.2
V	GTG 23.1	A	GCG 8.3	**E**	**GAG 62.5**	G	GGG 25.0

[1]For non-stop codons, zero-occurrences are shaded grey and predominantly used codons (>1.5x) are bolded, except for the one-codon 'family' ATG.

Table S1B. A+T-content and codon usage[1] of *cox1* cDNA from *D. papillatum*.

A+T = 44.7%

F	TTT 5.0	S	TCT 4.4	Y	TAT 30.0	C	TGT 36.4
F	**TTC 95.0**	S	TCC 40.0	**Y**	**TAC 70.0**	**C**	**TGC 63.6**
L	TTA 2.4	S	TCA 13.3	*	TAA 0.0	W	TGA 33.3
L	TTG 7.3	S	TCG 4.4	*	TAG 100	**W**	**TGG 66.7**
L	CTT 2.4	**P**	**CCT 46.7**	H	CAT 29.2	R	CGT 31.3
L	CTC 29.3	P	CCC 6.7	**H**	**CAC 70.8**	R	CGC 0.0
L	CTA 24.4	P	CCA 26.7	Q	CAA 0.0	R	CGA 6.3
L	CTG 34.1	P	CCG 20.0	**Q**	**CAG 100**	R	CGG 6.3
I	ATT 0.0	T	ACT 19.4	N	AAT 8.3	S	AGT 8.9
I	ATC 57.1	T	ACC 30.6	**N**	**AAC 91.7**	S	AGC 28.9
I	ATA 42.9	T	ACA 30.6	K	AAA 0.0	R	AGA 12.5
M	ATG 100	T	ACG 19.4	**K**	**AAG 100**	R	AGG 43.8
V	GTT 0.0	A	GCT 20.0	D	GAT 53.3	**G**	**GGT 40.0**
V	GTC 25.0	A	GCC 32.5	D	GAC 46.7	G	GGC 12.0
V	GTA 23.4	A	GCA 40.0	E	GAA 0.0	G	GGA 22.0
V	**GTG 51.6**	A	GCG 7.5	**E**	**GAG 100**	G	GGG 26.0

[1]See footnote of table S1A.

Table S1C. A+T-content and codon usage[1] of *cox1* cDNA from *Diplonema s*p. 2.

A+T = 54.7%

F	TTT	16.7	S TCT	5.9	**Y TAT**	**61.1**	C TGT 53.8
F TTC		**83.3**	S TCC	13.7	Y TAC	38.9	C TGC 46.2
L	TTA	12.5	S TCA	17.6	* TAA	100	**W TGA 70.0**
L	TTG	23.6	S TCG	7.8	* TAG	0.0	W TGG 30.0
L	CTT	16.7	P CCT	18.8	**H CAT**	**75.0**	R CGT 27.3
L	CTC	13.9	P CCC	6.3	H CAC	25.0	R CGC 18.2
L	CTA	26.4	**P CCA**	**62.5**	Q CAA	37.5	R CGA 0.0
L	CTG	6.9	P CCG	12.5	**Q CAG**	**62.5**	R CGG 9.1
I	ATT	32.3	**T ACT**	**43.8**	N AAT	36.4	S AGT 31.4
I	ATC	35.5	T ACC	10.4	N AAC	63.6	S AGC 23.1
I	ATA	32.3	T ACA	25.0	K AAA	0.0	**R AGA 45.5**
M	ATG	100	T ACG	20.8	**K AAG**	**100**	R AGG 0.0
V	GTT	14.0	A GCT	25.0	D GAT	53.8	G GGT 42.5
V	GTC	18.0	A GCC	10.4	D GAC	46.2	G GGC 17.5
V	GTA	38.0	**A GCA**	**50.0**	E GAA	57.1	G GGA 35.0
V	GTG	30.0	A GCG	14.6	E GAG	42.9	G GGG 5.0

[1]See footnote of table S1A.

Table S1D**.** A+T-content and codon usage[1] of *cox1* cDNA from *R. euleeides.*

A+T = 56.2%

F	TTT	22.2	S TCT	17.6	Y TAT	38.5	C TGT 20.0
F TTC		**77.8**	S TCC	13.7	**Y TAC**	**61.5**	**C TGC 80.0**
L	TTA	14.1	S TCA	15.7	* TAA	0.0	**W TGA 72.7**
L	TTG	1.4	S TCG	0.0	* TAG	100	W TGG 27.3
L	CTT	9.9	P CCT	40.0	**H CAT**	**65.2**	**R CGT 41.7**
L	CTC	15.5	P CCC	10.0	H CAC	34.8	R CGC 8.3
L CTA		**45.1**	P CCA	45.0	Q CAA	20.0	R CGA 0.0
L	CTG	14.1	P CCG	5.0	**Q CAG**	**80.0**	R CGG 0.0
I	ATT	15.8	**T ACT**	**55.8**	N AAT	50.0	**S AGT 41.2**
I	ATC	39.5	T ACC	7.0	N AAC	50.0	S AGC 11.8
I	ATA	44.7	T ACA	30.2	K AAA	0.0	R AGA 25.0
M	ATG	100	T ACG	7.0	**K AAG**	**100**	R AGG 25.0
V	GTT	9.8	A GCT	48.8	**D GAT**	**61.5**	**G GGT 56.9**
V	GTC	7.8	A GCC	2.4	D GAC	38.5	G GGC 7.8
V GTA		**52.9**	A GCA	48.8	E GAA	33.3	G GGA 35.3
V	GTG	29.4	A GCG	0.0	**E GAG**	**66.7**	G GGG 0.0

[1]See footnote of table S1A.

Table S2. Size determination of mitochondrial chromosomes from diplonemids.[a]

		D. ambulator	*Diplonema* sp. 2	*R. euleeides*	*D. papillatum*[b]
Electron microscopy		nd	nd	**7.0 + 7.7 kbp**[c]	**6.0 kbp + 7.0 kbp**
DNase digestion of mtDNA		~5 kbp	~5 kbp + 10 kbp	~7 kbp	~7 kbp
PCR of chromosomes carrying a *cox1* module	mod1	5.3 kbp	5.0 kbp	7.0 kbp	nd (B)
	mod2	5.0 kbp	5.0 kbp	8.0 kbp	nd (A)
	mod3	5.3 kbp	5.0 kbp	8.0 kbp	nd (A)
	mod4	5.0 kbp	10 kbp	8.0 kbp	7.0 kbp (B)
	mod5	4.5 kbp	5.0 kbp	8.0 kbp	nd (A)
	mod6	5.0 kbp	5.0 kbp	[5.0 kbp]	nd (A)
	mod7	5.3 kbp	5.0 kbp	8.0 kbp	nd (A)
	mod8	5.3 kbp	5.5 kbp	[12 kbp]	nd (A)
	mod9	5.0 kbp	5.0 kbp	7.0 kbp	6.0 kbp (A)

[a] nd, not determined. Sizes in brackets are most likely PCR artefact due to sequence repeats.

[b] (Marande & Burger, 2007). (A), (B), chromosome class.

[c] (Roy*, et al.*, 2007).

Table S3. Characterization of sequenced mitochondrial chromosomes from diplonemids.[a]

Species	ID, size, and A+T content of chromosomes carrying			Sequence similarity[b] of constant regions			Number of size classes
	Module 4	Module 5	Module 9	Compared chromosomes	Length of similar region	Global identity (local identity)	
D. ambulator	da0827 4.951 kbp A+T=52%	da0773 4.593 kbp A+T=50%	da0562 4.875 kbp A+T=52%	da0827/da0773 da0827/da0562 da0773/da0562	1.03 kbp 3.56 kbp 0.75 kbp	15.3% (71.4%) **63.4%** (89.6%) 11.3% (72.4%)	2
D. sp. 2	ds2016 9.077 kbp A+T=51%	ds1680 5.163 kbp A+T=54%	ds0266 5.184 kbp A+T=54%	ds2016/ds1680 ds1680/ds0266 ds2016/ds0266	0.21 kbp 4.69 kbp 0.30 kbp	3.1% (78.0%) **86.0**% (96.1%) 4.9% (88.9%)	2
R. euleeides	re6478 7.211 kbp A+T=52%	re6105 >6.196 kbp (8.0 kbp)[c] A+T=52%	re5226 ≥7.127 kbp (7.0 kbp)[c] A+T=50%	re6478/re6105 re6478/re5226 re6105/re5226	6.87 kbp 3.88 kbp 2.96 kbp	**80.2%** (80.2%) 66.7% (88.2%) 61.2% (89.7%)	2
D. papillatum	dp3209 (class A)[d] 7.18 kbp A+T=46%	(class A)[d]	dp3207 (class B)[d] 5.86 kbp A+T=45%	dp3207/dp3209 dp3207/dp4001[e] dp4001[e]/dp3209	4.38 kbp 5.54 kbp 4.56 kbp	49.6% (86.0%) **96.8%** (96.8%) 50.3% (75.8%)	2

[a]Note that in all diplonemids, it is Module 4 that is carried by the large-size-class chromosome.
[b]Local identity and length of similar regions (overlap length in constant region of longer chromosome) was obtained by Fasta (Pearson, 2000) comparison. Global identity was calculated as (Fasta identity) x (overlap length of constant region from longer chromosome) / (length of constant region from longer chromosome). Bold numbers indicate the identity of the most similar chromosomes.
[c]Size determined by PCR and electron microscopy; see supplementary table 2.
[d]Ref (Vlcek*, et al.*, 2011).
[e]The A-class chromosome carrying *cox1* Modules 5 of *D. papillatum* has not been fully sequenced. For a comparison of two A-class chromosomes carrying different gene modules, we used here dp4001 that encloses module 6 of *nad7* (GenBank acc. no. HQ288822 (Vlcek*, et al.*, 2011)).

Table S4A. Minimum offset for finding complementary sequences for at least N junctions in *D. ambulator*.[a]

W	GU[b]	N	M	Obs	Rep	Min	Max	Ave	Std
6	T	6	0	**8**	375	3	10	7.6	1.2
6	T	6	1	**1**	375	1	2	1	0.2
6	T	7	0	**11**	375	6	15	10.2	1.4
6	T	7	1	**2**	375	1	2	1.8	0.4
6	T	8	0	**12**	375	8	20	13.9	1.9
6	T	8	1	**2**	375	1	3	2.3	0.5
6	F	6	0	**18**	375	15	43	25.9	4.4
6	F	6	1	**2**	375	1	4	2.7	0.5
6	F	7	0	**19**	375	18	65	34.5	6.0
6	F	7	1	**3**	375	2	5	3.7	0.5
6	F	8	0	**22**	375	23	81	46.9	9.1
6	F	8	1	**4**	375	3	7	5.1	0.7
7	T	6	0	**10**	375	7	20	13.2	2.2
7	T	6	1	**1**	375	1	3	1.9	0.3
7	T	7	0	**13**	375	10	29	17.9	3.0
7	T	7	1	**2**	375	1	3	2.4	0.5
7	T	8	0	**25**	375	14	39	24.4	4.1
7	T	8	1	**4**	375	2	5	3.5	0.6
7	F	6	0	**45**	375	31	159	63.8	19.4
7	F	6	1	**4**	375	2	7	4.8	0.8
7	F	7	0	**55**	353	39	164	85.6	23.4
7	F	7	1	**6**	375	4	9	6.5	0.9
7	F	8	0	**77**	253	45	172	108.9	25.6
7	F	8	1	**7**	375	5	11	8.8	1.1

[a]W, window size; GU, GU pairs; N, number of junctions; M, number of mismatches; Obs, offset observed for the original dataset; Rep, number of randomly generated datasets; Min, minimum offset for the randomly generated datasets; Max, maximum offset for the randomly generated dataset; Ave, average offset for the randomly generated datasets; Std, standard deviation for the randomly generated datasets.
[b]T, true (GU pairs present); F, false (GU pairs absent).

Table S4B**.** Minimum offset for finding complementary sequences for at least N junctions in *D. papillatum*.[a]

W	GU[b]	N	M	Obs	Rep	Min	Max	Ave	Std
6	T	6	0	**5**	375	3	9	6.4	1.1
6	T	6	1	**1**	375	1	1	1.0	0.0
6	T	7	0	**5**	375	5	13	8.7	1.2
6	T	7	1	**2**	375	1	2	1.5	0.5
6	T	8	0	**10**	375	6	16	11.8	1.5
6	T	8	1	**2**	375	1	3	2.0	0.2
6	F	6	0	**23**	375	14	44	25.4	4.7
6	F	6	1	**3**	375	2	3	2.7	0.5
6	F	7	0	**25**	375	19	60	34.0	6.1
6	F	7	1	**4**	375	2	5	3.7	0.5
6	F	8	0	**29**	374	26	85	47.1	9.2
6	F	8	1	**5**	375	3	6	5.0	0.7
7	T	6	0	**12**	375	6	16	10.8	1.8
7	T	6	1	**1**	375	1	2	1.6	0.5
7	T	7	0	**14**	375	8	22	14.5	2.3
7	T	7	1	**2**	375	1	3	2.0	0.3
7	T	8	0	**19**	375	11	35	20.1	3.2
7	T	8	1	**3**	375	2	4	2.9	0.5
7	F	6	0	**47**	374	27	173	64.8	17.5
7	F	6	1	**4**	375	2	6	4.6	0.7
7	F	7	0	**56**	358	38	168	88.4	24.6
7	F	7	1	**6**	375	3	9	6.4	0.9
7	F	8	0	**69**	242	51	172	108.7	23.3
7	F	8	1	**10**	375	5	11	8.8	1.1

[a,b]See footnotes to table S4A.

Table S4C**.** Minimum offset for finding complementary sequences for at least N junctions in *Diplonema* sp. 2.[a]

W	GU[b]	N	M	Obs	Rep	Min	Max	Ave	Std
6	T	6	0	**8**	375	5	14	9.5	1.5
6	T	6	1	**1**	375	1	2	1.4	0.5
6	T	7	0	**10**	375	8	20	12.8	2.0
6	T	7	1	**2**	375	1	3	2	0.2
6	T	8	0	**16**	375	11	26	17.5	2.8
6	T	8	1	**3**	375	2	4	2.8	0.5
6	F	6	0	**28**	375	17	86	40.2	9.4
6	F	6	1	**3**	375	2	5	3.4	0.6
6	F	7	0	**41**	375	21	110	53.6	12.4
6	F	7	1	**4**	375	2	6	4.7	0.7
6	F	8	0	**67**	359	36	172	74.5	19.3
6	F	8	1	**5**	375	4	8	6.4	0.8
7	T	6	0	**16**	375	6	29	17.6	3.3
7	T	6	1	**2**	375	1	3	2.1	0.4
7	T	7	0	**23**	375	13	41	23.9	4.4
7	T	7	1	**2**	375	2	5	3.1	0.5
7	T	8	0	**27**	374	18	65	33.4	7.1
7	T	8	1	**4**	375	2	6	4.4	0.8
7	F	6	0	**80**	274	45	177	107.2	29.1
7	F	6	1	**7**	375	3	9	6.4	1.0
7	F	7	0	**119**	154	45	181	123.8	28.4
7	F	7	1	**9**	375	5	12	8.8	1.2
7	F	8	0	**-**	57	86	180	136.6	25.6
7	F	8	1	**13**	375	6	16	11.8	1.6

[a,b]See footnotes to table S4A.

Table S4D. Minimum offset for finding complementary sequences for at least N junctions in *R. euleeides*.[a]

W	GU[b]	N	M	Obs	Rep	Min	Max	Ave	Std
6	T	6	0	4	**375**	4	10	7.1	1.1
6	T	6	1	1	**375**	1	2	1	0.1
6	T	7	0	6	**375**	5	14	9.7	1.4
6	T	7	1	2	**375**	1	2	1.8	0.4
6	T	8	0	13	**375**	7	20	13.5	2.0
6	T	8	1	2	**375**	1	3	2.1	0.3
6	F	6	0	19	**375**	13	37	22.2	4.2
6	F	6	1	2	**375**	1	3	2.4	0.5
6	F	7	0	23	**375**	17	52	30	5.2
6	F	7	1	3	**375**	2	4	3.3	0.5
6	F	8	0	32	**375**	23	88	41.7	8.8
6	F	8	1	5	**375**	3	6	4.6	0.6
7	T	6	0	11	**375**	5	20	12.3	2.2
7	T	6	1	2	**375**	1	2	1.8	0.4
7	T	7	0	15	**375**	10	26	16.9	2.7
7	T	7	1	2	**375**	2	3	2.3	0.4
7	T	8	0	19	**375**	14	36	23.6	3.9
7	T	8	1	3	**375**	2	5	3.3	0.6
7	F	6	0	26	**375**	21	113	52.7	14.1
7	F	6	1	4	**375**	2	6	4.2	0.7
7	F	7	0	44	**373**	28	159	72.7	20.5
7	F	7	1	6	**375**	4	7	5.8	0.8
7	F	8	0	55	**294**	47	167	96.5	23.5
7	F	8	1	8	**375**	5	10	8	1.0

[a,b]See footnotes to table S4A.

Table S5. Diplonemid *cox1* sequences deposited in GenBank.[a]

Species	*cox1* cDNA	*cox1*-module-containing chromosomes			*cox1* genomic modules plus flanking regions Modules 1-3, 6-8
		Module 4	Module 5	Module 9	
D. ambulator	da0119 (JF698650)[b]	da0827 (JF698653)[b]	da0773 (JF698652)[b]	da0562 (JF698651)[b]	(JF698675-80)[b]
D sp. 2	ds2547 (JF698657)[b]	ds2016 (JF698656)[b]	ds1680 (JF698655)[b]	ds0266 (JF698654)[b]	(JF698658-63)[b]
R. euleeides	re6879 (JF698668)[b]	re6478 (JF698667)[b]	re6105 (JF698666)[b]	re5226 (JF698664,5)[b]	(JF698669-74)[b]
D. papillatum	dp4030riaC21 (EU123538)[c]	B3209 (EU12357)[c]	/	A3207 (EU12356)[c] A3208 (HQ288823)[d]	(HQ288825-33)[c]

[a]Sequence identifier with GenBank accession numbers in parenthesis.
[b]This report.
[c](Marande & Burger, 2007).
[d](Vlcek*, et al.*, 2011).

Table S6. Primers used for PCR, RT-PCR and DNA sequencing.

Primer-ID[a]	Primer sequence[b]	Used for
da1	CAGTACTTCCACTACGTATGCT	PCR for *cox1* m9 (cDNA)
da2	ACTACTAGCTGTAGCACTGCT	PCR for *cox1* m9
da5	TGCTGGTAAAGTACAGGGTCAC	PCR for *cox1* m4
da6	GGTATCGTATCTCACTGCATGC	PCR for *cox1* m4
da7	KAWTGCATGTCACTAGGTASG	PCR for *cox1* m9
da8	TATACGCTGTCCATACATAGC	PCR for *cox1* m9
da10	GCCATGATAGCTATAGGAGT	PCR for *cox1* m5
da11	ACACCATGCCTAGACTATTG	PCR for *cox1* m5
da12	GTTGTACCTGTACTAGCAGG	PCR for *cox1* m4
da13	CCATAGTAGCTACATACGTGG	PCR for *cox1* m4
da14	AGTATGCTTCGGTACGAGTG	PCR for *cox1* m1
da15	GCGGCATTAAGGTAGATACA	PCR for *cox1* m1 (cDNA)
da16	GTCCATGGATGTGTTCATCA	PCR for *cox1* m3
da17	GCTGGGCAATCTATTAATGCC	PCR for *cox1* m2
da18	CACTCATGACACCAGGCAT	PCR for *cox1* m2
da19	ACAGTGAGGTAGACCTATGT	PCR for *cox1* m6
re8	CCTCACTATTAGCTAGTACGA	PCR for *cox1* m6
da20	CATGAGCTACTCATTGGGTA	PCR for *cox1* m7
da21	AGTGGCCAGTACACCAGTA	PCR for *cox1* m7
da22	CCACTTAGATGCATCCCTCA	PCR for *cox1* m8
da23	GCACTGCTAGTTACTGTAGG	PCR for *cox1* m1
da24	CCACTAAGGTACCATCATGCTT	PCR for *cox1* m1
da25	GGCATAGAACTGCAGCTACCTAT	PCR for *cox1* m6
da26	ATACATGGTAGCACACATGC	PCR for *cox1* m6
da27	AGAAAGAGTATAGCGGCAGG	PCR for *cox1* m2
da28	TCATCCTACCTAAGTGTACTCC	PCR for *cox1* m2
da29	GCACACATCCAGTAAGTGCTA	PCR for *cox1* m7
da30	CCATGCAGTAAGTACGAGCAT	PCR for *cox1* m7
da31	GACTAGTAGTCATACTCAGTGG	PCR for *cox1* m3
da32	ACTTGGGTTACCTAGAAGAG	PCR for *cox1* m8
da33	TAGCATGCTCTCATGCAGTG	PCR for *cox1* m3
da34	GCATTATGCTACCACAGCAG	PCR for *cox1* m3
da35	TAGACAGCAGAGCACATGATGG	PCR for *cox1* m8
da36	GGAAGAGTAGTGCTACTGCA	PCR for *cox1* m8
da37	GGCAACACTGTCCTCTTAGA	PCR for *cox1* m7
da38	TGAGGTCAAGTGCTCTGCAT	PCR for *cox1* m7
dp33	CCAGGAGAACACCTTGATGGA	PCR for cDNA
dp82	CCATCAAGGTGTTCTCCTGGA	PCR for cDNA
dp38	GACTACCAGTATACCACAGG	For pre RNA
dp39	GCATCCATGCATCTGGAGG	For pre RNA
dp89	GGTATGCACATCAGTAGTAC	PCR for cDNA
ds5	GGTTGCATCGCAGTATGTT	PCR for cDNA
ds7	GTAGCATGTCTAGTTCGGTCC	PCR for *cox1* m9
ds8	GGATTCGTGACATTCGTGATG	PCR for *cox1* m9
ds9	CTGTATGTTCGACCACTAGTG	PCR for *cox1* m9 (cDNA)
ds10	CTAACTCCGGTACACTCATCTGG	PCR for cDNA

ds11	CTATCGCAATGGACATCATAG	PCR for *cox1* m5 (cDNA)
re14	CGGTTACTTCGTGTGAGCACATCA	PCR for *cox1* m5
ds13	GCGGTGATCCTATGTTGTATCAGC	PCR for *cox1* m4
ds14	CGAACTGTCATAGAAGCTGGTACTCC	PCR for *cox1* m4 (cDNA)
ds15	GCTGATACAACATAGGATCACCGC	PCR for *cox1* m4 (cDNA)
ds16	GGTGTTTGGGCATCCAGAAGTCT	PCR for *cox1* m4 (cDNA)
ds17	GAACGGTGTACACGCTCGTGG	PCR for *cox1* m5
dp33	CCAGGAGAACACCTTGATGGA	PCR for *cox1* m5 (cDNA)
ds18	GGATTACTATTGGTCGTGTGCG	PCR for *cox1* m9 (cDNA)
ds20	GGTTGTTGCAGAACATTACG	PCR for *cox1* m9
ds21	GGTACTATAAGCCGCTACTAG	PCR for *cox1* m4
ds22	CGTATAGCCCAGAGTAACT	PCR for *cox1* m4
ds23	CTACAGTAGATATGATGCGCG	PCR for *cox1* m4
ds24	TGCAGCGTTACTCATAGGA	PCR for *cox1* m4
ds25	ACTAGTAAGTTGCCTAGTCC	PCR for *cox1* m2
ds26	GGCAACTTACTAGTGCCATT	PCR for *cox1* m2
ds27	TGAACCTTCCTCCATCCAAGAT	PCR for *cox1* m3
ds28	ATGGAGGAAGGTTCAGGAA	PCR for *cox1* m3
ds29	CCATGAGCTACAGTGTATGCTTCC	PCR for *cox1* m6
ds30	GCCAAGAAGAATACCAGATGC	PCR for *cox1* m8
ds31	TTCTTCTTGGCAATCCCGCT	PCR for *cox1* m8
ds32	GGTTGCTTACTGGTTGTATG	PCR for *cox1* m7
ds33	GCACTCGTCAATAGTACGAA	PCR for *cox1* m7
ds34	GGGATTCATGTTGTCATGAC	PCR for *cox1* m1
ds35	CCTCTGACTAGTCATGACAA	PCR for *cox1* m1
ds36	ACAGCTCGATGTGGGAAT	PCR for *cox1* m6
ds37	CCTACTCACAACGTGTCAA	PCR for *cox1* m6
ds38	GGGATTTGCGTTAGTAAGTCTC	PCR for *cox1* m2
ds39	GCATCACACCGACTACTAGA	PCR for *cox1* m2
ds40	AAGTGTGCAACAGAACATGC	PCR for *cox1* m8
ds41	CTGGATCTGGATCCGGAT	PCR for *cox1* m8
ds42	ACTCGACAAGTGCTAGCTA	PCR for *cox1* m3
ds43	TCTGGATCTGGATCCGGAT	PCR for *cox1* m3
ds44	TTCGTACTATTGACGAGTGC	PCR for *cox1* m7
ds45	CATACAACCAGTAAGCAACC	PCR for *cox1* m7
ds46	GTACCATATGTCACACTGAGTG	PCR for *cox1* m1
ds47	CCAGCACTTCTCGAGTAATTGG	PCR for *cox1* m7
ds48	CCACAGTAGCAGTAGTGTGCTA	PCR for *cox1* m7
ds49	ATGCAACACATTGCATCACACC	PCR for *cox1* m1
ds50	CCACCGTGTATACCGTTTGAGT	PCR for *cox1* m1
re5	CTCACTATTAGCTAGTACGA	PCR for cDNA
re6	GCTCCATAGCATGGTAGCAC	RT-PCR for pre-mRNA
re7	GTACTACTGATGTGCATACC	RT-PCR for pre-mRNA
re8	CCTCACTATTAGCTAGTACGA	PCR for *cox1* m6 (cDNA)
re10	GTGCTACCATGCTATGGAGCA	PCR for cDNA
re11	GGTGTCCGAATACTCAGAAGAG	PCR for *cox1* m4
re12	CCGGTCTTTGGTATCGTATCAC	PCR for *cox1* m4
re13	GCATTGCATAGATCATGCCAAGC	PCR for *cox1* m5
re14	GGTTACTTCGTGTGAGCACATCA	PCR for *cox1* m5 (cDNA)
re15	CCGGTAGTATGATGATGTACAC	PCR for *cox1* m4
re16	CCGCAGACCCCTATTCCATACC	Primer walking re5226

re17	GGACCTTTTGGGGGTCCCAA	Primer walking re5226
re18	CGTACATGGTATACGTCCTAG	PCR for *cox1* m4
re19	CCATCTTAGCAGGTGCTATCAC	PCR for *cox1* m4
re20	GGTATACACACTGCCATACA	PCR for *cox1* m5
re21	CCATACCACCATTGCATAGTAT	PCR for *cox1* m5
re22	GGTGTTGATGAGTACTAG	PCR for *cox1* m9
re23	GGAGACTATGCACTAGTATACG	PCR for *cox1* m9
re24	GTGTCCTAATGGTCTCCGTA	PCR for *cox1* m5
re25	CCATCAGATCATGAGATCATGC	PCR for *cox1* m5
re26	CTGCTCCTAGATACATGCAT	PCR for *cox1* m1
re27	ACTATCCTTTGGAGCTAGTG	PCR for *cox1* m1
re30	CCAGAAGATAGCAGTAGTACC	PCR for *cox1* m8
re31	TATCTTCTGGAGCATGCATC	PCR for *cox1* m8
re32	ATGATGATAGTCCAGCAGCA	PCR for *cox1* m3
re34	ATAGTGAGGTAGACCTATCC	PCR for *cox1* m6
re35	GGCAATCTGCTGTTACCAGT	PCR for *cox1* m2
re36	GGAGTAGATAGCTGTACTGG	PCR for *cox1* m2
re37	CCTCCTAGTAACAGCACATG	PCR for *cox1* m7
re38	GGAGACATAGGTGCTATGCA	PCR for *cox1* m7
re39	GGAGACTATGCACTAGTATACG	PCR for *cox1* m1
re40	GTAGTTCTCGCATTAGCTGG	PCR for *cox1* m1
re41	GGTTGATATATGCATGCC	PCR for *cox1* m7
re42	GTATACGATAACGTGACC	PCR for *cox1* m7
re43	CTGGTGCTATCAATGTAGTG	PCR for *cox1* m3
re44	TAGCGGTAGTAGTGGTCTAC	RT-PCR for pre-mRNA
re45	GGTGTATGCATACAGTGTATGG	PCR for *cox1* m3
re46	GGAGACTATGCACTAGTATACG	PCR for *cox1* m3
re47	GGTACCATCACATCTACTGATCC	PCR for *cox1* m6
re48	TGCTCAGAGATATGTGGTTCC	PCR for *cox1* m6
re49	GGAGTGCTCATGGATATAGAG	PCR for *cox1* m3

[a]da, *D. ambulator*; ds, *Diplonema* sp. 2; re, *R. euleeides*
[b]K = G or T; W = A or T; S = G or C

1.2.8.6 Legend of supplementary figures

Supplementary figure 1**.** Multiple alignment of *cox1* cDNA sequences from diplonemids. Dp, *D. papillatum;* Da, *D. ambulator*; Ds, *Diplonema* sp. 2; Re, *R. euleeides*. Yellow highlighting indicates start and stop codons. The start of the coding regions was placed tentatively, most closely to that determined in *D. papillatum*. Minor length differences between the species are seen at the 5' and 3'ends of the reading frames. For example in *Diplonema* sp. 2, the start codon (Module 1) is up to 9 nt downstream and the stop codon (Module 9) is up to 12 nt upstream from the corresponding codons in other diplonemids. Red font color indicates nucleotides framing module junctions. If more than two positions are coloured, then the junction is between any adjacent pair of these positions. Arrows indicate junctions. If the precise junction is not known for a species, the position closest to that of the other species is indicated. Purple, nucleotides added by RNA editing. Cons, sequence consensus, across the four diplonemid cDNAs, of nucleotides adjacent to a shared module junction. *, four identical nucleotides; +, three identical nucleotides; ~, two pairs of two identical nucleotides; -, two identical and two different nucleotides.

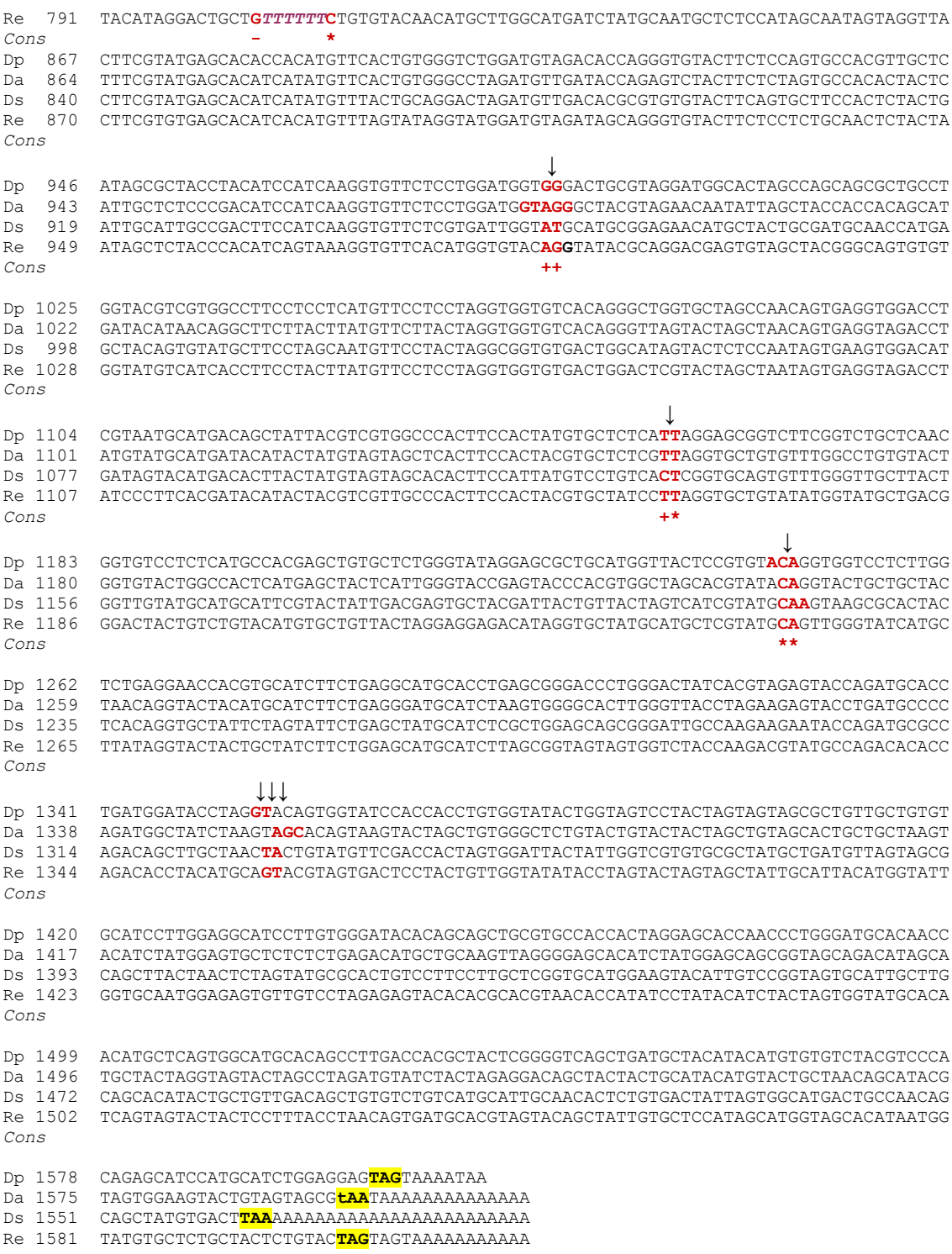
Supplementary figure 2**.** Sequence similarity between chromosomes carrying different modules. Comparison of the three chromosomes that carry *cox1* Modules 4, 5, and 9 from diplonemids. A, *D. ambulator*; B, *D. papillatum* (Vlcek*, et al.*, 2011); C, *Diplonema* sp. 2; D, *R. euleeides*. The module residing on a given chromosome is indicated. The circular chromosome sequences have been linearized for graphical representation. Boxes symbolize gene modules, and bars represent the rest of the chromosome (i. e. unique module-flanking regions and constant regions). Dark shading designates stretches of >90% sequence identity, and light shading designates stretches of ≥ 78% sequence identity based on FASTA comparison. In *D. papillatum*, Modules 5 and 9 reside on an A-class chromosome, and Module 4 resides on a B-class chromosome. The latter chromosome is not fully sequenced, so the missing sequence was completed tentatively with that of another class-A chromosome that has been fully sequenced (dp4001, carrying *nad7* Module 6 (Vlcek*, et al.*, 2011)). In *R. euleeides,* only one of the three chromosome sequences is entirely assembled due to massive repeats. The graphics shown in D is based on a tentative reconstruction guided by the mitochondrial chromosome structure in other diplonemids. Small boxes represent unassembled regions.

Supplementary figure 3. Genomic module sequences plus flanking regions aligned with cDNA of *cox1* from (A) *D. ambulator*, (B) *Diplonema* sp. 2, and (C) *R. euleeides*. For species and abbreviations, see legend of supplementary fig. 1. mod1-genomic, genomic sequence around boundaries of *cox1*-Module 1; same abbreviation for Modules 2-9. Upper case bold face, *cox1* module sequence (coding region); lower case, sequence of module-flanking region (not coding). Uncertain junction positions are shaded.

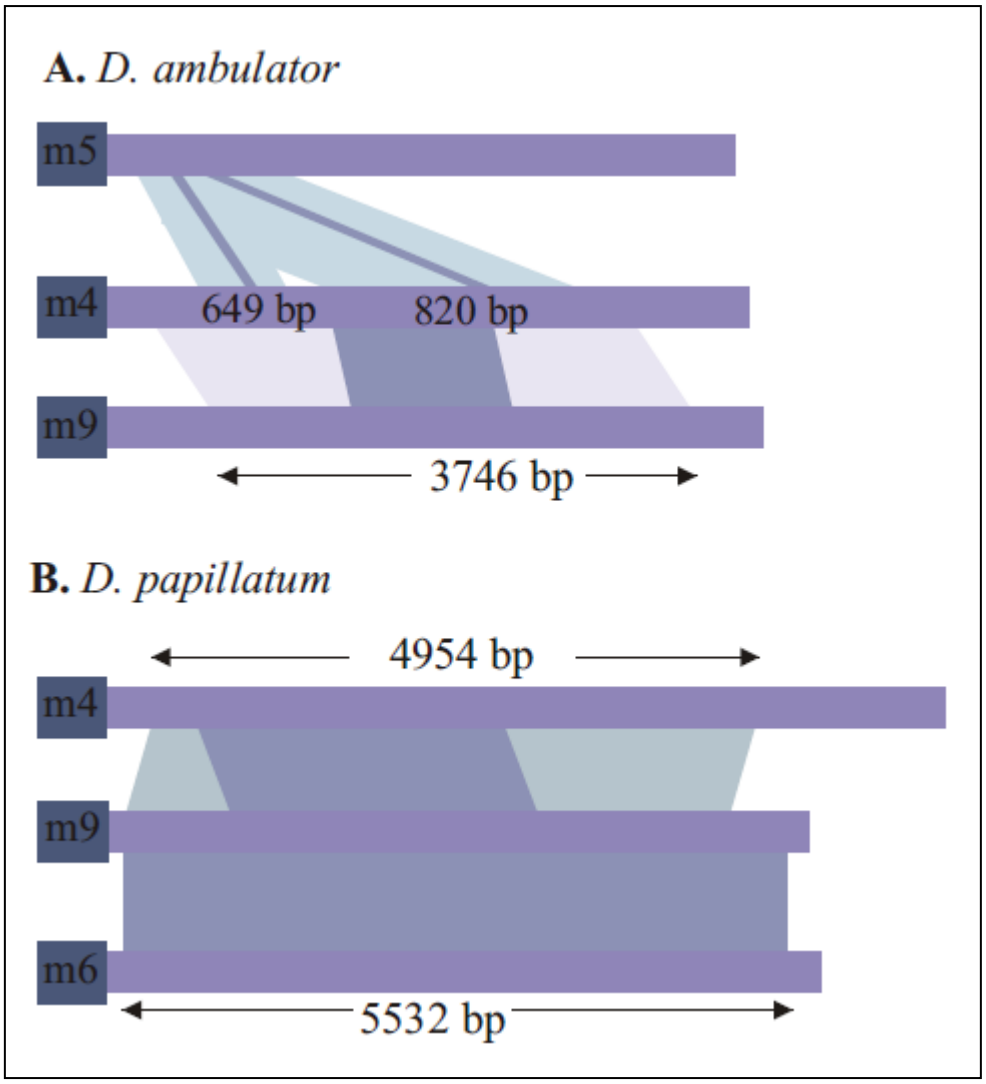
Supplementary figure 4. Multiple sequence alignment of deduced *cox1* proteins from diplonemids and other organisms. Yellow highlighting indicates amino acids corresponding to module junctions. Green highlighting shows amino acids whose complete codon originates from RNA editing (the two flanking codons are partly encoded and partly added post-transcriptionally). Euglenozoan taxa outside diplonemids are shown in blue font colour and diplonemids are indicated in bold face. Taxon abbreviations are: *Rick.prow., Rickettsia prowazekii*; *Homo.sapi., Homo sapiens*; *Amoe.para., Amoebidium parasiticum*; *Bige.nata., Bigelowiella natans*; *Mono.brev., Monosiga brevicollis*; *Acan.cast., Acanthamoeba castellanii*; *Recl.amer., Reclinomonas americana*; *Allo.macr., Allomyces macrogynus*; *Rhod.sali., Rhodomonas salina*; *Phyt.infe., Phytophthora infestans*; *Porp.purp., Porphyra purpurea*; *Cyan.para., Cyanophora paradoxa*; *Marc.poly., Marcantia polymorpha*; *Prot.wick., Prototheca wickerhamii*; *Neph.oliv., Nephroselmis olivacea*; *Naeg.grub., Naegleria gruberi*; *Tryp.bruc., Trypanosoma brucei*; *Leish.tare., Leishmania tarentulae*; *Eugl.grac., Euglena gracilis*; *Peta.cant., Petalomonas cantuscygni*; *Rhyn.eule., Rhynchopus euleeides*; *Dipl.*sp.2, *Diplonema* sp. 2; *Dipl.papi.*, *D. papillatum*; *Dipl.ambu., D. ambulator*. For GenBank acc. nos., see Methods (Supplementary Material).

Supplementary figure 5. Logos of sequence conservation upstream and downstream of genomic *cox1* modules. A, B, consensus across all *cox1* modules of *D. ambulator*; C, D, consensus of one representative genomic *cox1* module across all diplonemid species. The shown Module 2 displays more sequence conservation in the 5' part compared to the 3'part. However, this pattern is not seen for the other *cox1* modules.

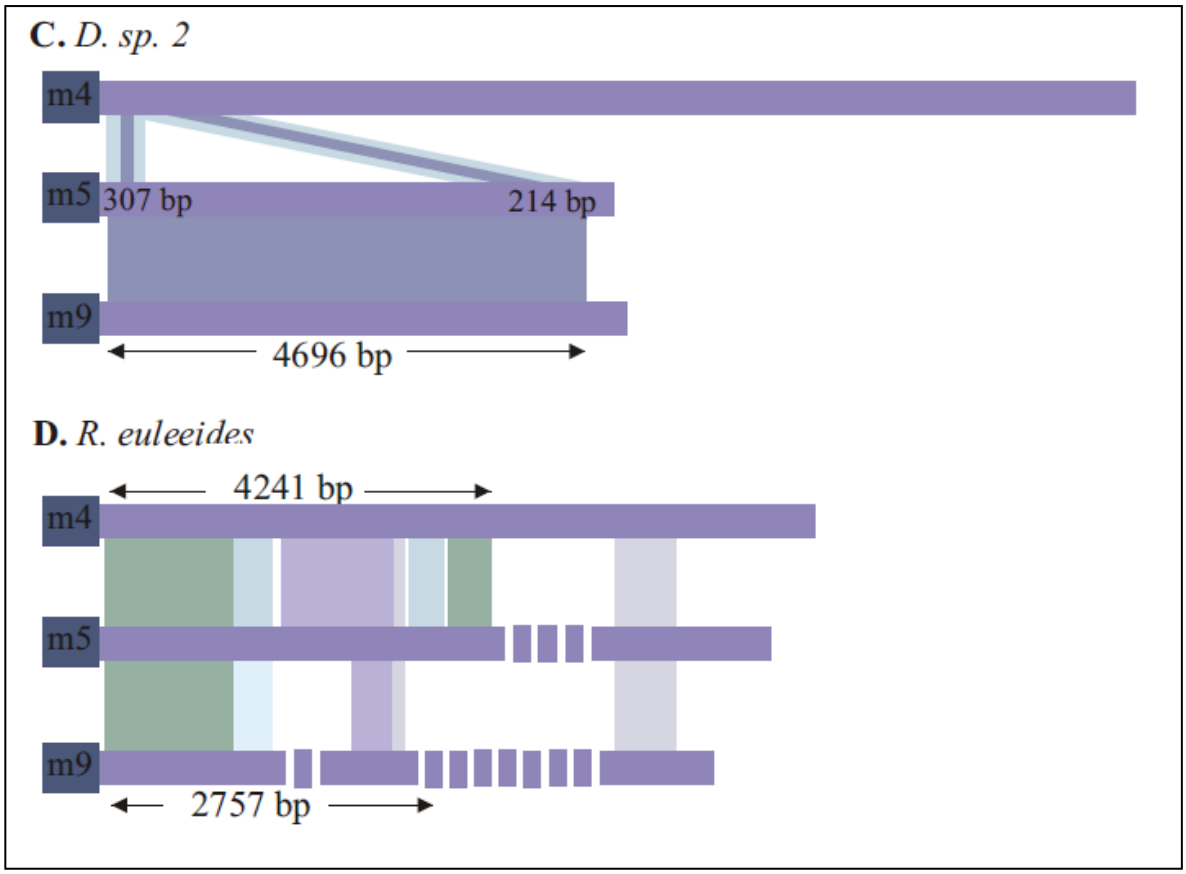
```
Dp     1  ---ATGCACCTAGAGGCTATAGCATCGGTGGTGTGGACTACCAATGCTAAGCTCATAGGGTGCGTGTACCTCAGCTGGT
Da     1  ------ATGTACAGCTTAGTAGCTACAATACTGTGGACTACGAATGCTAAGCTAATAGGGTGTATCTACCTTAATGCCG
Ds     1  ---------ATGAACACCACAGTAGCACTCATGATGACAACGAATGCTAAGTTAGTAGGAACAATCTACTTAGCACTCA
Re     1  ATAACCTACTATTAGTGTAGTACCAGCTATTGCATGAACTACTAATGCTAAGGTAGTGGGATGCATGTATCTAGGAGCAG
Cons

Dp    77  CCATAGCATTCGGGGTCTCAGGACTCCTCATGAGCTGGATCATGCGTGCTGAGCTATGTGGGCTATCCGAGCAGGTGCT
Da    74  CAGTATGCTTCGGTACGAGTGGGCTACTGCTATCATGGGTGATGCGTGGTGAACTAGGAGGTCTGAGTGAGCAGCTGCT
Ds    50  GTGTGACATATGGTACTATGGGATTCATGTTGTCATGACTAGTCAGAGGAGAGTTGTGTGGTCTAGGAGAACAACTGTT
Re    80  CACTATCCTTTGGAGCTAGCGGTATGCTCTTATCATGAGTACTACGTGGAGAGATAGGTGGTCTAGGTGAACAGCTACT
Cons

                                                        ↓
Dp   156  GTTCGGAGACCACCAGCTGTACAACGTCCTCACTACGACGCATGGCATCCTCATGCTCTTCTACTTCATCATGCCAGGG
Da   153  CTTTGGTGACCACCAACTATACAACGTACTCACTACCAGTCATGCCATGCTGATGATCTTCTTCTTCATCATGCCTGGT
Ds   129  GTTCGGTGATCATCAGTTGTACAACGTCCTAACAACATCACATGCAATGCTTATGATATTCTTCTTCATCATGCCAGGC
Re   159  GTTCGGAGACCATCATCTATACAACGTACTCATCACTAGTCATGCAATGCTGATGCTCTTCTTCTTCATCATGCCAGCA
Cons                                                    -*

Dp   235  GTGATGAGCGGTCTGGGTAACCTGCTGGTGCCCATCCAGCTCGGTGTCCCTGAGCTCATGTTCCCTAAGGTGAACAACG
Da   232  GTCATGAGTGGGCTGGGCAATCTATTAATGCCTATACATCTCTGCGTTCCCGAGATGGCGTTCCCTAAGGTGAATAATC
Ds   208  ACGATGGCAGGACTAGGCAACTTACTAGTGCCATTCCAGATGAGTGTACCGGAGTTAGTATTCCCAAAGATTAATAACA
Re   238  GCTATGAGTGGACTAGGCAATCTGCTGTTACCAGTACAGCTATCTACTCCTGAGATGATGTATCCTAAGGTTAACAACC
Cons

           ↓↓
Dp   314  TCGGGACATGGCTGCTAGTGGATGGCTACCTACTGCTAGTAGGATCCTCCTGAGTGGATGAGGGCGCAGGGACAGCATG
Da   311  TGGGAGCATGACTACTAGTAGATGGCTACTTACTGATAGTGGGGTCATCATGGATAGATGAAGGAGCAGGCACAGCATG
Ds   287  TCGGTATATGATTTTTAGTATGTGGTCTACTTTTGATTACGGGTTCATCTTGGATGGAGGAAGGTTCAGGAACGGCCTG
Re   317  TCGGAGCATGATTACTACTTAATGGATACCTACTAATCATAGGATCCTCATGAGTAGATGAAGGAGTAGGCACAGCATG
Cons       ~+

Dp   393  AACAGTGTACCCTCCGCTATCCATGACAGCTAGTCATGGTGGTGTGAGCGTAGACACCTTCATCGTGTCCCTACACGCA
Da   390  GACCATCTACCCACCACTGAGTATGACTACTAGTCATGGAGGGCTGTCCATGGATGTGTTCATCATATCATTACATGCT
Ds   366  AACCGTCTATCCACCACTAGCGCTCACTGCAAGTCATAGCGGACTTGCTGTAGATACGTTCATTATCGCATTGCACATG
Re   396  AACAGTATACCCTCCACTATCTACTGGTAGTACACATGGTGGTACTAGTATGGAGGTGTTCATAGTATCACTGCATGCT
Cons

Dp   472  GCAGGACTGTCATCCCTGACGGGTGCCATCAACCTGATGGTGACCTGCTGCTATGCCAGGAGGACACACAGCTGCGTCC
Da   469  GTAGGAGTATCATCGCTTACGGGTGCCATCAACATCATGGTAACAGGATGTTATGCTAGACGTACTCACACAGCGCTCA
Ds   445  GCCGGTGCAAGCTCCCTTACAGGAAGCATCAACCTTATATGTACAATCGCCTATGCCCGCCGTTCACTCATGGCGATGC
Re   475  GCTGGACTATCATCATTAACTGGTGCTATCAATGTAGTGAGTACATGCTACTATACTAAGAGAGCATCTAGTACACTAC
Cons

                                              ↓↓
Dp   551  TGCAGTCCTCATTGTATCCATGGAGCGTGGCAATCACAGGTGCACTGCTAGTAGGTATCATACCTGTACTAGCTGGTGC
Da   548  TGCAGACATCACTATACCCATGATCTATAGGCATCACAGGAGCTCTTCTAGTGGGTGTTGTACCTGTACTAGCAGGGGC
Ds   524  TGCAGTCATCACTTTATCCCTGATCCATTACAATCACTGCAGCGTTACTCATAGGAGTTGTGCCTGTGCTAGCAGGTGC
Re   554  TACATGCATCTCTATATCCATGAGCACTCCTTATTACTGCAGCTCTACTCATAGGAGTGATACCCATCTTAGCAGGTGC
Cons

Dp   630  CATCACCATGCTGCTCACTGATAGGAACACTGGTACGGTGTTCTATGACGTGGTAGCAGGTGGTGATCCGGTGATGTAC
Da   627  CATCACTATGCTGCTCACTGATAGATGTACCTGTACTACCTTCTATGATGTGCTCTCTGGTGGTGACCCTGTACTTTAC
Ds   603  TATCACGATGCTACTCACTGATAGAAGTTGGAGTACCAGCTTCTATGACAGTTCGGCAGGCGGTGATCCTATGTTGTAT
Re   633  TATCACTATGCTGTTATCAGATAGAGGATGCAGTACTAGCTATTATGATGTAGTAGCTGGTGGTGATCCTATCATGTAC
Cons

Dp   709  GAGCACCTCTTCTGGGTCTTCGGGCACCCAGAGGTCTACGTGATCATACTCCCGGTGTTCGGGCTGGTATCCCATTCCC
Da   706  CAGCATCTATTCTGGGTATTCGGTCACCCTGAAGTATACATCATAATACTACCTGTCTTTGGTATCGTATCTCACTGCA
Ds   682  CAGCACTTATTCTGGGTGTTTGGGCATCCAGAAGTCTATATCATCATACTTCCAGTATTCGGTATAGTCAGCCACGTTA
Re   712  CAACACCTCTTCTGAGTATTCGGACACCCAGAAGTGTACATCATCATACTACCGGTCTTTGGTATCGTATCACATACTA
Cons

                    ↓      ↓
Dp   788  TACATCGAGGAGGACCTTTTTTCGCTCTACAACATGCTGGGCATGGTGTACGCCATGATAGCCATAGCCGTGGTCGGGTA
Da   785  TGCACAGGGTAGCTGTTTTTTTCACTATTCAATAGTCTAGGCATGGTGTATGCCATGATAGCTATAGGAGTAGTGGGCTA
Ds   761  TTCATAAGAGCGCTATTTTTTCGTCATTCAATGTTCTTGGCATGGTGTATGCTATGATGTCCATTGCGATAGTTGGTTA
```



Supplementary figure 1.



Supplementary figure 2 A-B.



Supplementary figure 2. C-D.

D. ambulator

cDNA-*cox1*-Da …ACAACGTACTCACTACCAGTCATGCCATGCTGATG…
mod1-genomic-Da …**ACAACGTACTCACTACCAGT**atacgtatgcataat…
mod2-genomic-Da …agagtatagcggcagggtca**CATGCCATGCTGATG**…

cDNA-*cox1*-Da …TAAGGTGAATAATCTGGGAGCATGACTACTAGTAG…
mod2-genomic-Da …**TAAGGTGAATAATCTGGGAG**tacacttaggtagga…
mod3-genomic-Da …atgctaccacagcaggtaag**CATGACTACTAGTAG**…

cDNA-*cox1*-Da …ACTATACCCATGATCTATAGGCATCACAGGAGCTC…
mod3-genomic-Da …**ACTATACCCATGATCTATAGGC**tatgctactatac…
mod4-genomic-Da …atgaagtagctataaatact**GCATCACAGGAGCTC**…

cDNA-cox1-Da …CATGCACAGGGTAGCTG*TTTTTT*CACTATTCAATA…
mod4-genomic-Da …**CATGCACAGGGTAGCTG**ctctctctatgataccac…
mod5-genomic-Da …tgagggagcgcgctaacgagcag**CACTATTCAATA**…

cDNA-*cox1*-Da …CAAGGTGTTCTCCTGGATGGTAGGGCTACGTAGAA…
mod5-genomic-Da …**CAAGGTGTTCTCCTGGATGGTA**gcgcaagggaata…
mod6-genomic-Da …catagaactgcagctaccta**TAGGGCTACGTAGAA**…

cDNA-*cox1*-Da …CTTCCACTACGTGCTCTCGTTAGGTGCTGTGTTTG…
mod6-genomic-Da …**CTTCCACTACGTGCTCTCGT**atgcatgtgtgctac…
mod7-genomic-Da …tgcacacatccagtaagtgc**TAGGTGCTGTGTTTG**…

cDNA-*cox1*-Da …CACGTGGCTAGCACGTATACAGGTACTGCTGCTAC…
mod7-genomic-Da …**CACGTGGCTAGCACGTATAC**tcatgatgctcgtac…
mod8-genomic-Da …atacgccctgctagagcatg**AGGTACTGCTGCTAC**…

cDNA-*cox1*-Da …GCCCCAGATGGCTATCTAAGTAGCACAGTAAGTAC…
mod8-genomic-Da …**GCCCCAGATGGCTATCTAAGTAG**tatgctatcatg…
mod9-genomic-Da …ggtaatacagctatgaagtac**AGCACAGTAAGTAC**…

Supplementary figure 3A.

Diplonema sp. 2

cDNA-*cox1*-Ds …ACAACGTCCTAACAACATCACATGCAATGCTTATG…
mod1-genomic-Ds …**ACAACGTCCTAACAACATCACA**gactagcactaac…
mod2-genomic-Ds …ggaagtactattcataagat**CATGCAATGCTTATG**…

cDNA-cox1-Ds …AAAGATTAATAACATCGGTATATGATTTTTAGTAT…
mod2-genomic-Ds …**AAAGATTAATAACATCGGTA**cactcaacggatcac…
mod3-genomic-Ds …gcgacaacccaaaaatcacc**TATGATTTTTAGTAT**…

cDNA-cox1-Ds …TTTATCCCTGATCCATTACAATCACTGCAGCGTTA…
mod3-genomic-Ds …**TTTATCCCTGATCCATTACA**tgatcccgagatccc…
mod4-genomic-Ds …tatgaccagatgcactgatc**ATCACTGCAGCGTTA**…

cDNA-*cox1*-Ds …TATTCATAAGAGCGCTA*TTTTTT*CGTCATTCAATG…
mod4-genomic-Ds …**TATTCATAAGAGCGCTA**cgcgcatcatatctactt…
mod5-genomic-Ds …gtgcacagacttattatgacttt**CGTCATTCAATG**…

cDNA-cox1-Ds …AGGTGTTCTCGTGATTGGTATGCATGCGGAGAACA…
mod5-genomic-Ds …**AGGTGTTCTCGTGATTGGTA**ctagtgcgtagcaat…
mod6-genomic-Ds …agatgaactcccaggaccac**TGCATGCGGAGAACA**…

cDNA-cox1-Ds …CTTCCATTATGTCCTGTCACTCGGTGCAGTGTTTG…
mod6-genomic-Ds …**CTTCCATTATGTCCTGTCAC**gaccatccatcggaa…
mod7-genomic-Ds …gattactcgaatggaagttg**TCGGTGCAGTGTTTG**…

cDNA-cox1-Ds …GTTACTAGTCATCGTATGCAAAGTAAGCGCACTAC…
mod7-genomic-Ds …**GTTACTAGTCATCGTATGCA**tactaacacacacac…
mod8-genomic-Ds …cccaagagctcccaaagct**AAGTAAGCGCACTACT**…

cDNA-*cox1*-Ds …GCGCCAGACAGCTTGCTAACTACTGTATGTTCGAC…
mod8-genomic-Ds …**GCGCCAGACAGCTTGCTAACT**ccgatccccagatc…
mod9-genomic-Ds …cgctactgctattacgtgagg**ACTGTATGTTCGAC**…

Supplementary figure 3B.

R. euleeides

cDNA-cox1-Re …ACAACGTACTCATCACTAGTCATGCAATGCTGATG…
mod1-genomic-Re …**ACAACGTACTCATCACTAGT**agcatgcagtacact…
mod2-genomic-Re …aggtacagtgtactagtact**CATGCAATGCTGATG**…

cDNA-cox1-Re …CTAAGGTTAACAACCTCGGAGCATGATTACTACTT…
mod2-genomic-Re …**CTAAGGTTAACAACCTCGGA**gaccactagtagtaa…
mod3-genomic-Re …aacattaccctattactagt**GCATGATTACTACTT**…

cDNA-cox1-Re …TATATCCATGAGCACTCCTTATTACTGCAGCTCTA…
mod3-genomic-Re …**TATATCCATGAGCACTCCTTA**ctgcatagtatgca…
mod4-genomic-Re …gacctttaacattaccctac**ATTACTGCAGCTCTA**…

cDNA-*cox1*-Re …TATACATAGGACTGCTG*TTTTTT*CTGTGTACAACA…
mod4-genomic-Re …**TATACATAGGACTGCTG**agcatacagtactagtac…
mod5-genomic-Re …gctgtagaggatgtgtactacta**CTGTGTACAACA**…

cDNA-cox1-Re …AGGTGTTCACATGGTGTACAGGTATACGCAGGACG…
mod5-genomic-Re …**AGGTGTTCACATGGTGTACA**gtacctatgctcata…
mod6-genomic-Re …catgtcctctcgtcttagaa**GGTATACGCAGGACG**…

cDNA-*cox1*-Re …CTTCCACTACGTGCTATCCTTAGGTGCTGTATATG…
mod6-genomic-Re …**CTTCCACTACGTGCTATCCT**ctgattagcagaggg…
mod7-genomic-Re …gaaaggacatcatctcacta**TAGGTGCTGTATATG**…

cDNA-*cox1*-Re …TGCTATGCATGCTCGTATGCAGTTGGGTATCATGC…
mod7-genomic-Re …**TGCTATGCATGCTCGTATGC**tcagaccatatgtac…
mod8-genomic-Re …actactaatagcatacaact**AGTTGGGTATCATGC**…

cDNA-*cox1*-Re …ACACCAGACACCTACATGCAGTACGTAGTGACTCC…
mod8-genomic-Re …**ACACCAGACACCTACATGCAG**tacctagactcaca…
mod9-genomic-Re …ggtattagtacctatgctcta**TACGTAGTGACTCC**…

Supplementary figure 3C.

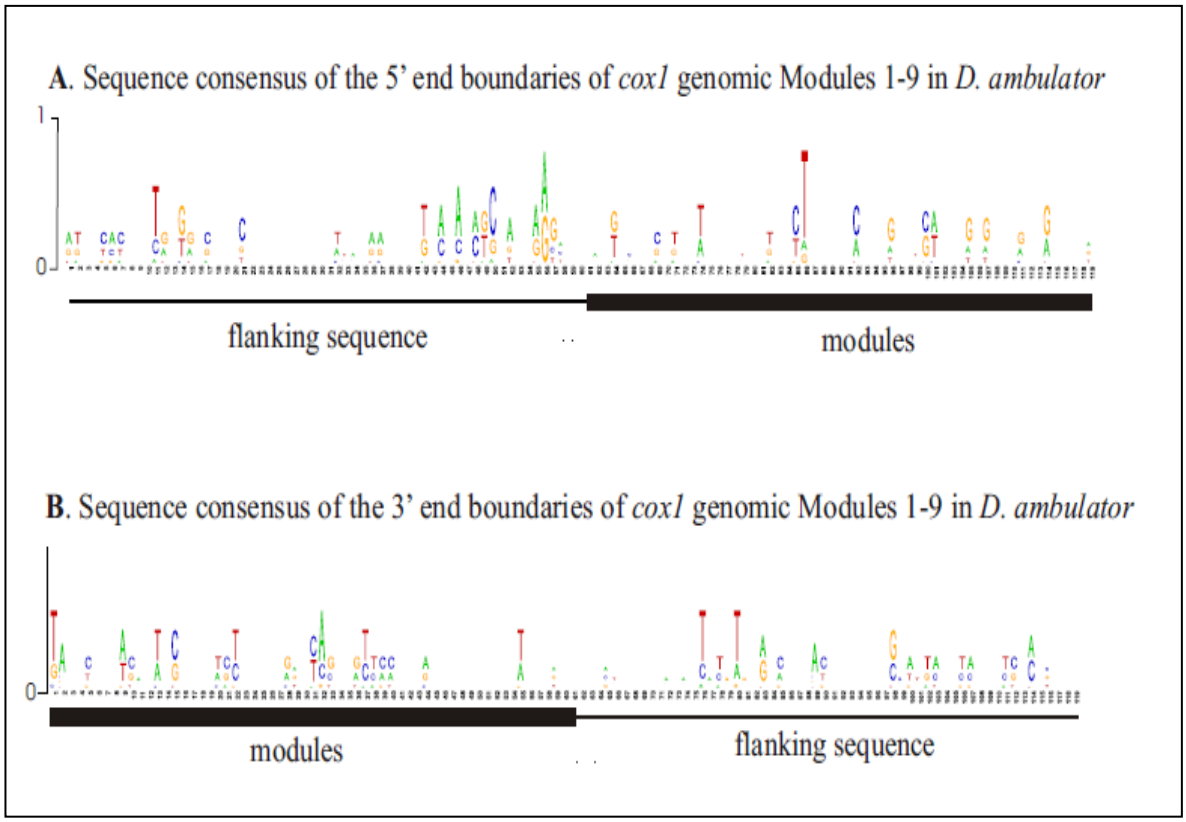
```
                1         11         21         31         41         51         61         71         80
                |          |          |          |          |          |          |          |          |
Rick.prow.      1                MEFDQ VLLSRIQFAF TISFHIVFPT FTIGLASFLA VIEGLWLKTK -NPIY----- QEIYKFWVK-
Homo.sapi.      1                  MFAD RWLFSTNHKD IGTLYLLFGA WAGVLGTALS LLIRAELGQP GN--L--LGN DHIYNVIVTA
Amoe.para.      1               MASWAQ KWLFSTNHKD IGMMYIIAGA FGGLLGTTFS VLIRIELGVP GT--V--LGD NHMYNVIITA
Bige.nata.      1  MSLGLSGFSQ STVQTSPSIE RWLYSTDHKD IGTMYFIYGA FAGVLGTVMS IYMRMELSYP SVQIL--AEN NQLYNVLVTA
Mono.brev.      1                MSWLT RWVFSTNHKD IGVLYFIFGS FSGFLGTAMS VIIRMELSIP GSPFL--AGD SHLYNVIVTA
Acan.cast.      1      MINRLL NNLTSFFTDN RWLFSTNHKD IGTLYLIFGG FSGIIGTIFS MIIRLELAAP GSQIL--SGN SQLYNVIITA
Recl.amer.      1               MANSFVK RWFFSTNHKD IGALYIMFGT FAGITATTIS VVMRLELGLP GNQIL--QGN HQLYNVLITA
Allo.macr.      1              MFQRNTVY RWLFSTNAKD IGTLYLVFSI FAGMIGTAFS VLIRFELAGP GVQYL--YGD HQLYNVIITA
Rhod.sali.      1                 MAFIN RWLFSTNHKD IGVLYLVFAI FSGVVGTALS ILIRAELSGP GVQVL--GGN HQLYNVIVTG
Phyt.infe.      1           M NFQNINKWST RWLFSTNHKD IGTLYLIFSA FAGVVGTTFS LLIRMELAQP GNQIF--MGN HQLYNVVVTA
Porp.purp.      1             MQKSLNNWIF RWIYSTNHKD IGTLYLIFGA FSGVLGACAS ILIRMELAQP GNQLL--LGN HQVYNVLVTE
Cyan.para.      1                 MDAFIH RWLYSTNHKD IGTLYLVFGA FSGLLGTAFS FLIRLELANP GNQIL--AGN HQLYNVIVTA
Marc.poly.      1                MNNFAQ RWLFSTNHKD IGTLYLIFGA IAGVMGTCFS VLIRMELAQP GNQIL--GGN HQLYNVLITA
Prot.wick.      1                   MVT RWLYSTNHKD IGTMYLIFGA FSGVLGTVFS LLIRMELAQP GNQIL--NGN HQLYNVIITA
Neph.oliv.      1                MSNFVQ RWLFSTNHKD IGTLYLIFGA FSGVLGTAFS LIIRMELAQP GNQIL--AGN HQLYNVIITA
Naeg.grub.      1                MLNFCK SWIFTTNHKR IGILYLFFGV FNGFLAVLLS MLMRLELAFP GDQIL--FGE YHFYNMITTV
Tryp.bruc.      1                 MFFLC LVCLSVSHKM IGICYLLVAI LCGFIGYIYS LFIRLELSLI GCGVL--FGD YQFYNVLITS
Leis.tare.      1                 MFWLC LVCLSVSHKM IGLCYLLVAI LSGFVGYVYS LFIRLELSLI GCGIL--FGD YQFYNVLITS
Eugl.grac.      1              MINNIMHMIN KYTLTTSHKI IGILYGYMGY IAGILGYIIS MLIRMELNTQ GLAIVRKVKE VTIYNNWITI
Rhyn.eule.      1                MPTISVVP AIAWTTNAKV VGCMYLGAAL SFGASGMLLS WVLRGEIGGL GEQLL--FGD HHLYNVLITS
Dipl.sp.2       1                   MNTTV ALMMTTNAKL VGTIYLALSV TYGTMGFMLS WLVRGELCGL GEQLL--FGD HQLYNVLTTS
Dipl.papi.      1                MHLEAIA SVVWTTNAKL IGCVYLSWSI AFGVSGLLMS WIMRAELCGL SEQVL--FGD HQLYNVLTTT
Dipl.ambu.      1                  MYSLVA TILWTTNAKL IGCIYLNAAV CFGTSGLLLS WVMRGELGGL SEQLL--FGD HQLYNVLTTS

                81         91        101        111        121        131        141        151      160
                |          |          |          |          |          |          |          |        |
Rick.prow.     59  -------IFA VTFGMGVVSG VVMSYQFGTN WSNFSDKVGN V----LGPLL GFEVFT-AFF LESSFLGIML F-------GF
Homo.sapi.     61  HAFVMIFFMV MPIMIGGFGN WLVPLMIGAP DMAFP-RMNN MSFWLLPPSL LLLLAS-AMV EAGAGTGWTV YPPLAGNYSH
Amoe.para.     63  HAFLMIFFMV MPVLVGGFGN WLLPILIGAP DMAFP-RLNN ISLWLLPPAL ILLVSS-ALV EQGAGTGWRY YPPLSGLEAH
Bige.nata.     79  HAFLMIFFMV MPVLMGGFGN WFMPILIGAP DMAFP-RLNN LSLWLLTPSL FLLLLS-SLV ETGVGTGWTV YPPLSSIQYH
Mono.brev.     64  HAFLMIFFLV MPFLMGGFGN FFVPLMIGAP DMSFP-RMNN ISFWLLPPSL ILLVAS-SLV EGGAGTGWTV YPPLSSVEFH
Acan.cast.     75  HAFVMIFFFV MPVMIGGFGN WFVPLMIGAP DMAFP-RLNN ISFWLLPPSL FLLLCS-SLV EFGAGTGWTV YPPLSSIVAH
Recl.amer.     66  HGLLMLFMVV MPILLGGFGN FFVPLLIGAP DMAFP-RLNN ISFWLLPPAL LLLFFS-ALV EVGAGTGWTA YPPLSGIQSH
Allo.macr.     67  HAFIMIFFLV MPAMLGGFGN YFVPIMIGAP DMAFP-RLNN ISFWLLPPSL ILLVGS-AFV EQGAGTGWTV YPPLSSIGFH
Rhod.sali.     64  HAFIMIFFMV MPALIGGYGN FLVPIMIGAV DMAFP-RMNN VSFWLLPPAL LLLISS-TLT EGGAGTGWTV YPPLSSVEGH
Phyt.infe.     70  HAFIMVFFLV MPALIGGFGN WFIPLMIGAP DMAFP-RMNN ISFWLLPPSL LLLVSS-AIV ESGAGTGWTV YPPLSSVQAH
Porp.purp.     69  HAFLMIFFMV MPVLIGGFGN WFVPIMIGAP DMAFP-RLNN ISFWLLPPSL CLLLGS-AMV EVGAGTGWTL YPPLSSIQSH
Cyan.para.     65  HAFIMVFFMV MPVLIGGFGN WFVPIMIGAP DMAFP-RLNN ISFWVLPPSL LLLTVS-ALV ETGAGTGWTV YPPLAAIQGH
Marc.poly.     65  HAFLMIFFMV MPAMIGGFGN WFVPILIGSP DMAFP-RLNN ISFWLLPPSL LLLLSS-ALV EVGCGSGWTV YPPLSGITSH
Prot.wick.     62  HAFLMIFFML MPALMGGFGN WFLPILIGAP DMAFP-RLNN ISFWLLPPSL LLLVSS-ALV EVGAGTGWTV YPPLASIASH
Neph.oliv.     65  HAFLMIFFMV MPVLIGGFGN WFVPIMIGAP DMAFP-RLNN ISFWLLPPSL LLLLSS-ALV EVGAGTGWTV YPPLSSIQYH
Naeg.grub.     65  HGVLMLFVVV MPILFGGFGN YFVPILIGAP DMSFP-RLNN FSFWLLPGAI LLAVLA-TYS EGGPGTGWTV YPPLSSLQSH
Tryp.bruc.     64  HGLIMVFAFI MPITMGGFTN YFAPVMVGFP DMVFP-RINN MSFWMFIGGF GCLVSG-FLT EEGMGVGWTL YPTLICIDFH
Leis.tare.     64  HGLIMVFAFI MPVMMGGLVN YFIPVMAGFP DMVFP-RLNN MSFWMYLAGF GCVVNG-FLT EEGMGVGWTL YPTLICIDFH
Eugl.grac.     71  HGLIMLFVFI MPVGIGFYGN YLIPMLIGTS ELSMP-RMNG ISFWMLIVGV VIFVISNVLM SKPISSGWTL YPPLSTRDAD
Rhyn.eule.     67  HAMLMLFFFI MPAAMSGLGN LLLPVQLSTP EMMYP-KVNN LGAWLLLNGY LLIIGS-SWV DEGVGTAWTV YPPLSTGSTH
Dipl.sp.2      57  HAMLMIFFFI MPGTMAGLGN LLVPFQMSVP ELVFP-KINN IGIWFLVCGL LLITGS-SWM EEGSGTAWTV YPPLALTASH
Dipl.papi.     66  HGILMLFYFI MPGVMSGLGN LLVPIQLGVP ELMFP-KVNN VGTWLLVDGY LLLVGS-SWV DEGAGTAWTV YPPLSMTASH
Dipl.ambu.     68  HAMLMIFFFI MPGVMSGLGN LLMPIHLCVP EMAFP-KVNN LGAWLLVDGY LLIVGS-SWI DEGAGTAWTI YPPLSMTTSH
```

```
              161        171        181        191        201        211        221        231       240
                |          |          |          |          |          |          |          |         |
Rick.prow.120  N-KVTKKVHF ISTLIVAIGT IISAFWILAA SSWMHTPAGF ELRDEGFFYP L---NWLEII FNPSFPYRFF HMITASYLTT
Homo.sapi.139  P-GASVDLTI FSLHLAGVSS ILGAINFITT IINMKPPAMT QYQTPLFVWS VLITAVLLLL SLPVLAAGIT MLLTDRNLNT
Amoe.para.141  S-GGSVDLAI FSLHLAGVSS LLGAINFITT TINMRTPKMG MHELPLFVWA IFITAFLLLL SLPVLAGGIT MLLTDRNLNT
Bige.nata.157  P-GASVDLAI FSLHIAGVSS IAGSINFITT VINMRAPGMY MHRMPLFAWA VFITSWLLVL SLPVLAGAIT MLLTDRNLNT
Mono.brev.142  S-GGSVDLAI FSLHLAGVSS LLGASNFITT ILNMRAPGMT MHKLPLFVWA VFITAILLLL SLPVLAAGIT MLLTDRNFNT
Acan.cast.153  S-GGSVDLAI FSLHLAGISS LLGAINFITT IFNMRVPGLS MHKLPLFVWS VLITAFLLLF SLPVLAGAIT MLLTDRNFNT
Recl.amer.144  S-GASVDLAI FSLHLSGASS VLASINFITT IFNMRAPGMT MHRMPLFVWS ILVTSFLLVF ALPVLAGGIT MLLTDRNFNT
Allo.macr.145  S-GGSVDLAI FSLHLAGISS MLGSINFITT ILNMRAPGMT MHKLPLFVWS ILITAILLLL SLPVLAGAIT MLLTDRNLNT
Rhod.sali.142  P-SAAIDLGI FSLHVAGASS ILGAINFITT IFNMRCPGMT FHRLPLFVWA VLITAFLLLL SLPVFAGAIT MLLTDRNFNT
Phyt.infe.148  S-GPSVDLAI FSLHLSGISS LLGAINFIST IYNMRAPGLS FHRLPLFVWS ILITAFLLLL TLPVLAGAIT MLLTDRNLNT
Porp.purp.147  S-GGAVDLAI FSLHLSGASS VLGAINFITT IFNMRNPGQS MYRIPLFVWS ILITAFLLLL AVPVLAGAIT MLLTDRNFNT
Cyan.para.143  S-SASVDLAI FSLHLSGAGS ILGAVNFIST IFNMRALGLK MHQLPLFVWA ILITAFLLLL SLPVFAGALT MLLTDRNFNT
Marc.poly.143  S-GGSVDLAI FSLHLSGVSS ILGSINFITT IFNMRAPGLT MHRLPLFVWS VLVTAFLLLL SLPVLAGAIT MLLTDRNFNT
Prot.wick.140  S-GGSVDLAI FSLHLAGVSS ILGAINFICT VFNMRAPGMS MHRLPLFVWA VFITAWLLLL CLPVLAGGIT MLLTDRNFNT
Neph.oliv.143  S-GGSVDLAI FSLHVSGASS ILGAINFITT IFNMRGPGMT MHRLPLFVWA VLITAFLLLL SLPVFAGAIT MLLTDRNFNT
Naeg.grub.143  S-GASVDLMI FSFHLVGIGS IVAAINFICT IFYYKNEAMF NKDLPLFVWS VAVTSFLVIV AIPVLAAAIT LLLFDRNFNT
Tryp.bruc.142  S-SLACDFII FSVHFLGISS ILNSINVVGT IFCCRRKYFS FLIWTLFIWG ALLTSILLII TLPVLAGGVT LLLCDRNFNT
Leis.tare.142  S-SLACDFVM FAVHLLGISS ILNSINLLGT LFCCRRKFFS FLSWSLFIWA ALITAILLII TLPVLAGGVT LILCDRNFNT
Eugl.grac.150  NIGVNIDLSL LVVHVLGISS TIGSVNYITT NKYNRHVGLT FMNINIYNFS IIVTSILLIG SLPILGVAIT GLLLDRNINS
Rhyn.eule.145  G-GTSMEVFI VSLHAAGLSS LTGAINVVST CYYTKRASST LLHASLYPWA LLITAALLIG VIPILAGAIT MLLSDRGCST
Dipl.sp.2 135  S-GLAVDTFI IALHMAGASS LTGSINLICT IAYARRSLMA MLQSSLYPWS ITITAALLIG VVPVLAGAIT MLLTDRSWST
Dipl.papi.144  G-GVSVDTFI VSLHAAGLSS LTGAINLMVT CCYARRTHSC VLQSSLYPWS VAITGALLVG IIPVLAGAIT MLLTDRNTGT
Dipl.ambu.146  G-GLSMDVFI ISLHAVGVSS LTGAINIMVT GCYARRTHTA LMQTSLYPWS IGITGALLVG VVPVLAGAIT MLLTDRCTCT

              241        251        261        271        281        291        301        311       320
                |          |          |          |          |          |          |          |         |
Rick.prow.196  SF---VIGGV ASFYLLNNRY KKHAKIMLFM AVLMALIVSP IQIF--IGDL HGLNTLKYQP VKVSAIEGI- WNTEKGASFN
Homo.sapi.218  TFFDPAGGGD PILYQHLFWF FGHPEVYILI LPGFGMISHI VTYYSGKKEP FGYMGMVWAM MSIGFLGFIV W----AHHMF
Amoe.para.220  TFYDPAGGGD PVLYQHLFWF FGHPEVYILI IPAFGILSHV VQHYS-HKSI FGYLGMVYAM LSIGVLGFIV W----AHHMF
Bige.nata.236  AFFDPAGGGD PVLYQHLFWF FGHPEVYILI LPGFGIISHI VSALT-SKPV FGYLGMVYAM ISIGFLGFIV W----AHHMY
Mono.brev.221  SFFDPAGGGD PILYQHLFWF FGHPEVYILI IPGFGIVSHI VSTFS-DKPV FGYLGMVYAM LSIGLLGFIV W----AHHMY
Acan.cast.232  SFFDPSGGGD PILYQHLFWF FGHPEVYILI LPAFGIVSQI IGTFS-NKSI FGYIGMVYAM LSIAVLGFIV W----AHHMY
Recl.amer.223  TFFDPAGGGD PVLFQHLFWF FGHPEVYILV IPGFGVVSHV ISAFS-RRPI FGYLGMVYAM SSIGVLGFIV W----AHHMY
Allo.macr.224  TFYDPAGGGD PVLYQHLFWF FGHPEVYIII IPGFGIISQV ISTFS-RKPI FGYLGMVYAM ASIGILGFIV W----SHHMY
Rhod.sali.221  TFFNPAGGGD PVLYQHLFWF FGHPEVYILI LPGFGIISQI ISTFS-RKPV FGYVGMIYAM LSIGLLGFIV W----AHHMY
Phyt.infe.227  SFYDPSGGGD PVLYQHLFWF FGHPEVYVLI LPAFGIISQV SASFA-KKNV FGYLGMVYAM LSIGLLGSIV W----AHHMF
Porp.purp.226  TFFDPSGGGD PVLYQHLFWF FGHPEVYILI LPGFGIVSHI VSTFS-RKPV FGYIGMIYAM LSIGILGFIV W----AHHMY
Cyan.para.222  TFFDPAGGGD PILYQHLFWF FGHPEVYILI IPGFGIISHV ISTFS-NKPV FGYLGMVYAM LSIGILGFIV W----AHHMY
Marc.poly.222  TFFDPAGGGD PILYQHLFWF FGHPEVYILI LPGFGIISHI VSTFS-RKPV FGYLGMVYAM ISIGVLGFIV W----AHHMF
Prot.wick.219  SFFDPAGGGD PILYQHLFWF FGHPEVYILI IPGFGIISHV IATFS-KKPI FGYLGMVYAM CSIGILGFIV W----AHHMY
Neph.oliv.222  TFFDPAGGGD PILYQHLFWF FGHPEVYILI IPAFGIVSHV ISTFS-KKPV FGYLGMVYAM MSIGILGFIV W----AHHMY
Naeg.grub.222  SFYDPVGGGD VVLYQHLFWF FGHPEVYILI LPGFGLVSHI IATFS-KKRV FGHVPMIAAM LMIGLIGFIV W----AHHMY
Tryp.bruc.221  SFYDVVGGGD LVLFQHLFWF FGHPEVYIII LPVFGLVSTI IEVTS-FRCV FSSVAMIYSM LLISVLGMFV W----AHHMF
Leis.tare.221  SFYDVVGGGD LILFQHIFWF FGHPEVYIIL LPVFGLISTI VEVIG-FRCV FSTVAMIYSM ILIAILGMFV W----AHHMF
Eugl.grac.230  TIYDVI--GD PVLYQHLFWF FGHPEVYVII LPVFGLTSLI LTSII-HKDI FGREGMMYCI ISIGVVGYFV W----AHHMF
Peta.cant.  1             ..LWERCASE FGHPEVYVII LPAFGIVSMS MSRLT-SFSV SVHSGMVLAI LAISMVGFFV W----AHHMF
Rhyn.eule.224  SYYDVVAGGD PIMYQHLFWV FGHPEVYIII LPVFGIVSHT IHRTA-VFSV YNMLGMIYAM LSIAIVGYFV W----AHHMF
Dipl.sp.2 214  SFYDSSAGGD PMLYQHLFWV FGHPEVYIII LPVFGIVSHV IHKSA-IFSS FNVLGMVYAM MSIAIVGYFV W----AHHMF
Dipl.papi.223  VFYDVVAGGD PVMYEHLFWV FGHPEVYVII LPVFGLVSHS LHRGG-LFSL YNMLGMVYAM IAIAVVGYFV W----AHHMF
Dipl.ambu.225  TFYDVLSGGD PVLYQHLFWV FGHPEVYIII LPVFGIVSHC MHRVA-VFSL FNSLGMVYAM IAIGVVGYFV W----AHHMF
```

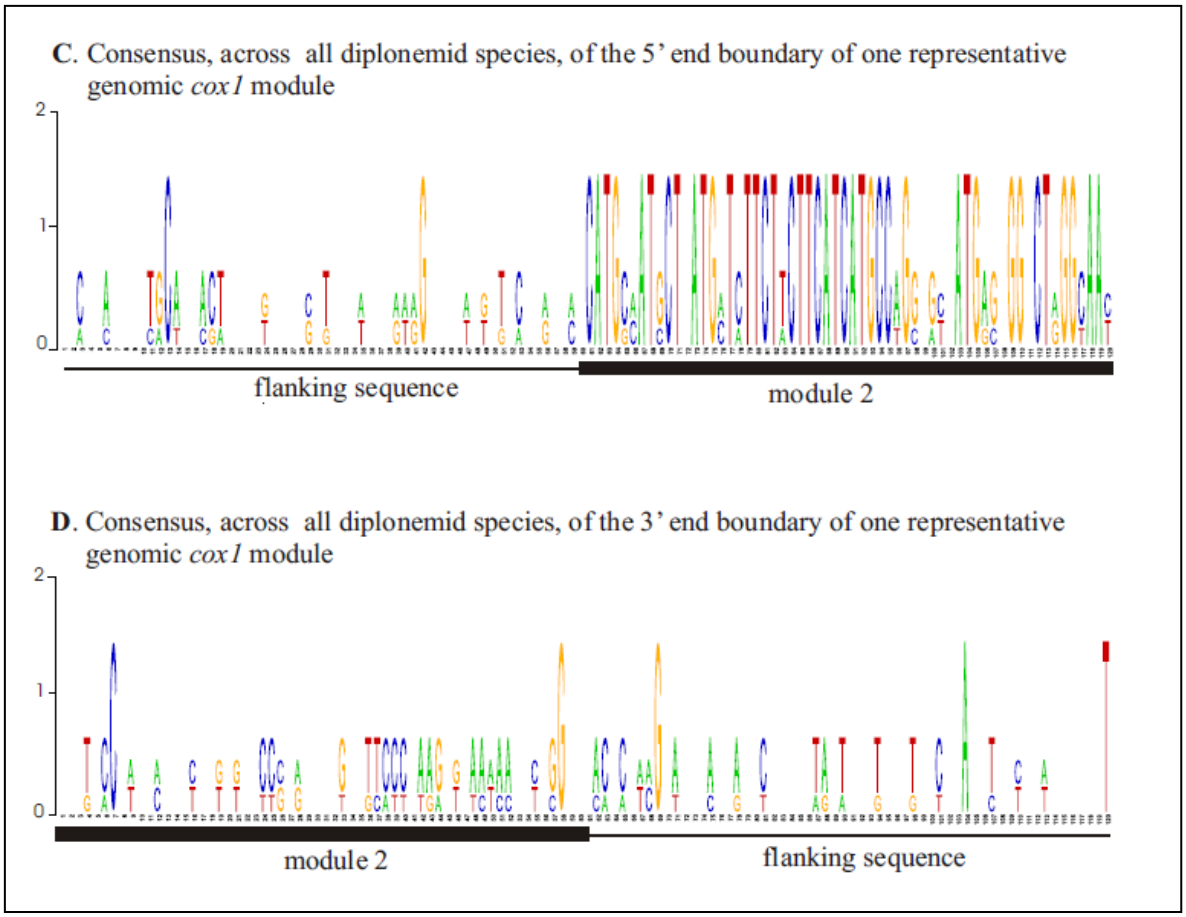
```
             321        331        341        351        361        371        381        391       400
               |          |          |          |          |          |          |          |         |
Rick.prow.270  LIGLPDKEEE KTKYAIEIPY ASSLILTHSL DGEVKGLKEW TKEER-PPV AVVFFSFRIM LGIGCLMVFT GIAGLYLY--
Homo.sapi.294  TVGM----DV DTR-----AY FTSATMIIAI PTGVK-VFSW LATLH-GSN MKWSAAVLWA LGFIFLFTVG GLTGIVLANS
Amoe.para.295  TVGM----DV DSR-----AY FTAATMIIAV PTGIK-IFSW LATLF-GGS IRLTSAMVFG IGFLFLFTIG GLTGIALANG
Bige.nata.311  TVGM----DV DVR-----AY FTAATMIIAV PTGIK-IFSW LATMW-GGQ VYLTTPLLFA LGFIFLFTVG GVTGVALANA
Mono.brev.296  TVGM----DV DTR-----AY FTASTMIIAV PTGIK-IFSW LGTMY-GGS IRLKVPMYWA LGFIFLFTLG GITGVMLANG
Acan.cast.307  TVGL----DV DTR-----AY FTAATMMIAV PTGIK-IFSW IATLW-GGQ IVRKTPLLFV IGFLILFTLG GLTGIVLSNA
Recl.amer.298  TVGM----DV DTR-----AY FTAATMVIAI PTGIK-IFSW LATMW-GGS IELKAPMLFA VGFVFLFTFG GLTGVVLSNS
Allo.macr.299  TVGL----DV DTR-----AY FTAATMIIAV PTGIK-IFSW LATLY-GGN ILYRTPAYFA LGFLFLFTIG GVTGVMLANA
Rhod.sali.296  TVGM----DV DTR-----AY FTAATMIIAV PTGIK-IFSW IATMW-GGS IYYRTPMIFA VGFIFLFTIG GLTGIVLSNA
Phyt.infe.302  TVGL----DV DTR-----AY FSAATMIIAV PTGIK-IFSW LATLW-GGS LKFETPLLFV LGFILLFVMG GVTGVAMSNS
Porp.purp.301  TVGL----DV DTR-----AY FTAATMIIAV PTGIK-IFSW VATMW-EGS IFLKTPMLFA IGFIFLFTIG GLTGIILANS
Cyan.para.297  TVGL----DV DTR-----AY FTAATMIIAV PTGIK-IFSW IATMW-GGS IVLHTPMLFA VGFILLFTIG GLTGIVLSNS
Marc.poly.297  TVGL----DV DTR-----AY FTAATMIIAV PTGIK-IFSW IATMW-GGS IQYKTPMLFA VGFIFLFTVG GLTGIVLANS
Prot.wick.294  VVGL----DI DTR-----AY FTAATMIIAV PTGIK-IFSW VATMW-GGS IELRTPMLFA VGFLFLFTVG GLTGVVLANS
Neph.oliv.297  TVGL----DV DTR-----AY FTAATMIIAV PTGIK-IFSW IATMW-GGS IEFKTPMLFA VGFIFLFTIG GFTGIILSNS
Naeg.grub.297  TSGI----DT STK-----AY FTAATMVIAI PTGIK-VFNW IATMW-GGS IWMYTPMYFA AGFVVLFTIG GITGIILSNA
Tryp.bruc.296  VVGM----DV DSR-----AY FGSITVLIGL PTCIK-LFNW IYSFL-FTD MCICFEIYFI YMFILMFLAG GLTGLFLSNV
Leis.tare.296  VVGM----DV DSR-----AY FGGVSILIGL PTCVK-LFNW IYSFL-YTD MIITFEVYFV IMFIFMFLIG AVTGLFLSNV
Eugl.grac.303  TVGL----DI DSR-----SY FSIATSIISI PTSVK-MFSY INTWASGRG FRGNNSSWSF FSFLICFCFG GFTGLLLSSG
Peta.cant. 64  VAGI----DD DSR-----VY FSTATMVIAV PTAAK-IFTW IACLV-SLQ CSV-IDMVII LMFLICFTAG GFTGIALSNC
Rhyn.eule.299  SIGM----DV DSR-----VY FSSATLLIAL PTSVK-VFTW CTGIR-RTS VA-TGSVWYV ITFLLMFLLG GVTGLVLANS
Dipl.sp.2 289  TAGL----DV DTR-----VY FSASTLLIAL PTSIK-VFSW LVCMR--RT CYCDATMSYS VCFLAMFLLG GVTGIVLSNS
Dipl.papi.298  TVGL----DV DTR-----VY FSSATLLIAL PTSIK-VFSW MVGLR--RM ALASSAAWYV VAFLLMFLLG GVTGLVLANS
Dipl.ambu.300  TVGL----DV DTR-----VY FSSATLLIAL PTSIK-VFSW MVGLR--RT ILATTTAWYI TGFLLMFLLG GVTGLVLANS

             401        411        421        431        441        451        461        471       480
               |          |          |          |          |          |          |          |         |
Rick.prow.346  -LNKRLYTTY WFQ-YWYILM SPSGFIAVLA GWLVTEVGRQ PYIVYNILKT VDTVSPLLG- ---------- ----------
Homo.sapi.362  SLDIVLHDTY YVVAHFHYVL SMGAVFAIMG GFIHWFPLFS GYTLDQTYAK IHFTIMFIGV NLTFFPQHFL GLSGMPRRYS
Amoe.para.363  GLNIALHDTY YVVAHFHYVL SMGAVFGVFM GIYHWCGVMT GSTISERLGY LHFGLMFLGV NITFFVQHFL GLAGMPRRIP
Bige.nata.379  GLDIAFHDTY YVVAHFHYVL SMGAVFAMFA GFYFWAPKAT GMQYNELLGK LHFWVFFVGV NVTFFPMHFL GLAGMPRRIP
Mono.brev.364  GLDIALHDTY YVVAHFHYVL SMGAVFALIG GVYYWIGKVT GYAYPETWGK IHFWLMFIGV NLTFFPQHFL GLAGFPRRYN
Acan.cast.375  GLDIMLHDTY YVVAHFHYVL SMGAVFAFFA GFYYWFWKIS GYTYNEMYGN VHFWLMFIGV NLTFFPMHFV GLAGMPRRIP
Recl.amer.366  GLDIALHDTY YVVAHFHYVL SMGAMFAIYA AFYYWFGKIT GYQYPEKLAQ VQFWTTFIGV NLTFFPMHFL GLSGMPRRIP
Allo.macr.367  SLDVALHDTY YVVAHFHYVL SMGAVFALFA GFYYWIGKIT GKQYNEFWGQ VHFWTMFIGV NVTFFPMHFL GLNGMPRRIP
Rhod.sali.364  GLDIAFHDTY YVVAHFHYVL SMGAVFAVFA GYYYWIGKIT GFQYPENLGV IHFWCTFVGV NLTFFPQHFL GLAGMPRRIP
Phyt.infe.370  GLDIALHDTY YIVGHFHYVL SMGAVFGIFT GFYFWIGKIS GRRYPEILGQ IHFWLFFIGV NVTFFPMHFL GLAGMPRRIP
Porp.purp.369  GLDISLHDTY YVVAHFHYVL SMGAVFAIFA GFYYWFEKIS GFQYSEILGQ IHFWGTFIGV NLTFFPMHFL GLAGMPRRIP
Cyan.para.365  GLDIAFHDTY YVVGHFHYVL SMGAVFAMFA GFYYWIGKIS GLKYPELLGK IHFFSTFIGV NMTFFPMHFL GLAGMPRRIP
Marc.poly.365  GVDIALHDTY YVVAHFHYVL SMGAVFALFA GFYYWIGKIT GLQYPETLGQ IHFWITFFGV NLTFFPMHFL GLAGMPRRIP
Prot.wick.362  GLDVAFHDTY YVVAHFHYVL SMGAVFALFS GFYYWIGKIT GLQYPETLGQ IHFWLMFLGV NITFFPMHFL GLAGMPRRIP
Neph.oliv.365  GLDIALHDTY YVVGHFHYVL SMGAVFGMFA GFYYWIGKIT GLQYPETLGQ IHFWLFFIGV NVTFFPMHFL GLAGMPRRIP
Naeg.grub.365  GIDTSIHDTH YVVGHFHYVL SMGAVFAIYG GFYYWWGKMT GLSYSESLGQ AHFWMTFIGV NFTFFPMHFL GSGGMPRRIP
Tryp.bruc.364  GIDILMHDTY FVVAHFHYVL SLGAVVGVFG GFFHFLMKWI PIELHTFWLF FFISTLWFGS NMVFFPLHSL GMFAFPRRIS
Leis.tare.364  GIDIMLHDTY FVVGHFHYVL SLGAVVGFFT GFIHFLAKWL PIELYLFWMF YFISTLFIGS NMLFFPMHSL GMYAFPRRIS
Eugl.grac.372  SLDIMLHDTY FVVGHFHTVL SLAATFGLLI AHYFFLPIIF SYSIFESFSF YHTFLLLVGA LLVFYPMHLA GLSGMARRVP
Peta.cant.131  SMDVLYHDTY YVVGHFHFVW ...
Rhyn.eule.366  EVDLSLHDTY YVVAHFHYVL SLGAVYGMLT GLLSVHVLLL GGDIGAMHAR MQLGIMLIGT TAIFWSMHLS GSSGLPRRMP
Dipl.sp.2 356  EVDMIVHDTY YVVAHFHYVL SLGAVFGLLT GCMHAFVLLT SATITVTSHR MQVSALLTGA ILVFWAMHLA GAAGLPRRIP
Dipl.papi.365  EVDLVMHDSY YVVAHFHYVL SLGAVFGLLN GVLSCHELCS GYRSAAWLLR VQVVLLVWGT TCIFWGMHLS GTLGLSRRVP
Dipl.ambu.367  EVDLCMHDTY YVVAHFHYVL SLGAVFGLCT GVLATHELLI GYRVPTWLAR IQVLLLLTGT TCIFWGMHLS GALGLPRRVP
```

```
             481        491        501        511        521        531        541        551       560
               |          |          |          |          |          |          |          |         |
Rick.prow.403  ----KYVFIS LIAFVVVYLI IFGVGIY-YI IYLIKKG--- ---------- ---------- ---------- ----------
Homo.sapi.442  DYPDAYTTWN ILSSVGSFIS LTAVMLMIFM IWEAFAS--- KRKVLMVEEP S--------- ---------- ----------
Amoe.para.443  DYPDAYISWN VISSYGSIVS LLGTLVFLYL LYSSQAK--- ---------- ---------- ---------- ----------
Bige.nata.459  DYPDAYAGWN HVASVGSFIS FLSVALFFYI VYDMFIG--- ---------- ---------- ---------- ----------
Mono.brev.444  DFPDAYAEWN LLSSFGSLIS VVAVIVFMYV IYRTLTD--- GVVVG--NNY W--------- ---------- ----------
Acan.cast.455  DYPDNYYYWN ILSSFGSIIS SVSVIVFFYL IYLAFNNNNT PKLIKLVHSI F--------- ---------- ----------
Recl.amer.446  DYPDAFSGWN AVSSYGSLVT TFSIILWFYI VYRTLTD--- GVKCG--NDP W--------- ---------- ----------
Allo.macr.447  DYPDAFTQWN VISSFGSIIS IVSTIVFLYG LYLTLSQ--- PAVSLA-NNY WHVPSF---- ---------- ----------
Rhod.sali.444  DYPDSYAEWN MLSSFGSYFS VFAIFIFFVL IYETLTN--- MEQCE--VNP W--------- ---------- ----------
Phyt.infe.450  DFPDAMSGWN AVSSFGSYIS FFSALFFFYI VYVTLVH--- GKKIE--N
Porp.purp.449  DYPDSYAGWN TIASYGSYVA LFSTLFFFYL VFNTLVT--- PRKVPARNNP W--------- ---------- ----------
Cyan.para.445  DYPDAFAGWN AVASYGSYIS VVSAIFFFYV VYKTLTS--- GEPCE--NNP W--------- ---------- ----------
Marc.poly.445  DYPDAYAGWN AFSSFGSYVS VVGIFCFFVV VFLTLTS--- ENKCA--PSP W--------- ---------- ----------
Prot.wick.442  DYPDCYAGWN AVASYGSYLS ITAVLFFFYV VYKTLTS--- NEVCP--RNP W--------- ---------- ----------
Neph.oliv.445  DYPDAYAGWN AIASFGSYLS ILGALFFFYV VYATLTG--- NEKVG--NNP W--------- ---------- ----------
Naeg.grub.445  DYPDMYQYHN TLASFGAFIS FFSLLFFFYV IYCSFTD--- QVKCP--RNP WIFIDYNDLI DRMIGVIYYV EKYGAFGKDE
Tryp.bruc.444  DYPISFLFWS AFTLYGMLLL TF-LVIFCCC LFNVILF--- WDYCLFFINL F--------- ---------- ----------
Leis.tare.444  DYPVSFLFWS SFMLYGMLL- LASLILFLCA LFCVFLF--- WDYCLFFVSL F--------- ---------- ----------
Eugl.grac.452  EYADIFIPFM TVGFHGTFLL IFSTLTFIRS YFQFLSH--- ---------- ---------- ---------- ----------
Rhyn.eule.446  DTPDTYMQYV VTPTVGIYLV LVAIALHGIG AMESVVL--- ---------- ---------- ---------- ----------
Dipl.sp.2 436  DAPDSLLTTV CSTTSGLLLV VCAMLMLVAQ LTNSSMRTVL PCSVH----- ---------- ---------- ----------
Dipl.papi.445  DAPDGYLGTV VSTTCGILVV LLVVALLLCA SLEASLWDTQ QLRATTRST- ---------- ---------- ----------
Dipl.ambu.447  DAPDGYLSST VSTSCGLCTV LLAVALLLST SMECSLW--- DMLQVRGAHL W--------- ---------- ----------

             561        571        581        591        601
               |          |          |          |          |
Rick.prow.435  ---------- ---------- --IEAIDNNE TYVEHWLSNR L*
Homo.sapi.490  ---------- ---------- ---------- -MNLEWLYGC PPPYHTFEEP VYMKS*
Amoe.para.480  ---------- ---------- ------QNHQ GAISELAAST SSFFQSRKEY IYLV*
Bige.nata.496  ---------- ---------- ---GFEEKRF VARDAWSKTR ITKASLLKLH *
Mono.brev.490  ---------- ------RSKE LFEKEGDIPT IHSLEWAETS PPHFHCYNEL PYLVASRTHT S*
Acan.cast.506  ---------- ---------A PYINTLSKNL LTFASIKSTS DSSFFKFSKF FIFF*
Recl.amer.492  ---------- ---------G LAVGEPGKEH FATLEWTLTS PPLSHTFEEV PYIKETIKK*
Allo.macr.499  ---------- ------FSST HSLYGDTTQT SSSLEWVLPS PPAFHAFNHL PVQS*
Rhod.sali.490  ---------- ---------K FSNEDVKNDF EYTLEWLVGS PPAFHTFNEV PLIKETVVSK IN*
Porp.purp.497  ---------- ---------- --NFEDSKIG STTLEWEISS PPAYHTFNEI PLVRETEISL KIN*
Cyan.para.491  ---------- ---------V FDEKKGNNKQ SHNIEWVLSS PVQTHLFEEL PIIPVKVKKK KN*
Marc.poly.491  ---------- ---------- -----AVEQN STTLEWMVPS PPAFHTFEEL PAIKESI*
Prot.wick.488  ---------- ---------- ----ETTPGV SPTLEWMLPS PPAFHTFEEI QV*
Neph.oliv.491  ---------- ---------- ---ADKTRDY ASTLEWMVGS PPAFHTFHQI PTIKETSA*
Naeg.grub.520  PRTISPDFIE KWISFNYVTN RYIISAESIK TTTLEWTLTS PPYLHTFVVP PKLFTTGSHY FEYRWNAILN KKRKFIPLYL*
Tryp.bruc.491  ---------- ---------- ---TYSLSIF FYFYTWVPVC MAIYLLVIDF AHIILDYLLI ILCFCFVFYI FFWQAFLLFF*
Leis.tare.491  ---------- ---------- ---VFSLYCF FYFSTWLPCV MVLYLLLVDF AHIVLDYLFL ILCFCFVFFI FFWQSLFLFF*
Eugl.grac.489  ---------- ---------- ---------- ---------- ------INHS NYL*
Rhyn.eule.483  ---------- -------EST HARNTISYTS TSGMHISSTT PLPNSDARST AIVLHSMVAH NGMCSATLY*
Dipl.sp.2 481  ---------- ---------- ---GSTLSGS ALLAAHTAVD SCVCHALQHS VTISGMTANS SYVT*
Dipl.papi.494  ---------- ---------N PGMHNHMLSG MHSLDHATRG QLMLHTCVST SHRASMHLEE *
Dipl.ambu.495  ---------- ---------S SGSRHSMLLG STSLDVSTRG QLLLHTCTAN SIRSGSTVVA *
```

Supplementary Figure 4.



Supplementary figure 5 A-B.



Supplementary figure 5 C-D.

1.2.8.7 References

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

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, and Lipman. DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**:3389-3402.

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, and Young IG (1981) Sequence and organization of the human mitochondrial genome. *Nature* **290**:457-465.

Benne R, Van den Burg J, Brakenhoff JP, Sloof P, Van Boom JH, and Tromp MC (1986) Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* **46**:819-826.

Chaput H, Wang Y, and Morse D (2002) Polyadenylated transcripts containing random gene fragments are expressed in dinoflagellate mitochondria. *Protist* **153**:111-122.

Crooks GE, Hon G, Chandonia JM, and Brenner SE 2004. WebLogo: a sequence logo generator. *Genome Research* **14**:1188-1190.

Edgar RC (2004a) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**:113.

Edgar RC (2004b) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**:1792-1797.

Feagin J. E (1990) RNA editing in kinetoplastid mitochondria. *Journal of Biological Chemistry* **265**:19373-19376.

Foldynova-Trantirkova S, Paris Z, Sturm NR, Campbell DA, and Lukeš J (2005) The *Trypanosoma brucei* La protein is a candidate poly(U) shield that impacts spliced leader RNA maturation and tRNA intron removal. *International Journal for Parasitology* **35**:359-366.

Frohman MA, Dush M K, and Martin GR (1988) Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide

primer. *Proceedings National Academy of Sciences of Unites States of America* **85**:8998-9002.

Gagliardi D, Stepien PP, Temperley RJ, Lightowlers RN, and Chrzanowska-Lightowlers ZM (2004) Messenger RNA stability in mitochondria: different means to an end. Trends in Genetics **20**:260-267.

Gillespie DE, Salazar NA, Rehkopf DH, and Feagin JE (1999) The fragmented mitochondrial ribosomal RNAs of *Plasmodium falciparum* have short A tails. *Nucleic Acids Research* **27**:2416-2422.

Gordon, D (2003) Viewing and editing assembled sequences using Consed. Curr Protoc *Bioinformatics* **Chapter 11**:Unit 11.12.

Marande, W (2007) Structure et expression des gènes mitochondriaux de *Diplonema papillatum*. Montreal (Canada): Biochemistry Department. Université de Montréal. 98p.

Marande Wand Burger G (2007) Mitochondrial DNA as a genomic jigsaw puzzle. *Science* **318**:415.

Marande W, Lukeš J and Burger G (2005) Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids. *Eukaryotic Cell* **4**:1137-1146.

Maslov DA, Yasuhira S, and Simpson L (1999) Phylogenetic affinities of *Diplonema* within the *Euglenozoa* as inferred from the SSU rRNA gene and partial COI protein sequences. *Protist* **150**:33-42.

Miyazawa S and Jernigan RL (1999) Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins* **34**:49-68.

Ostermeier C, Harrenga A, Ermler U and Michel H (1997) Structure at 2.7 A resolution of the *Paracoccus denitrificans* two-subunit cytochrome c oxidase complexed with an antibody FV fragment. *Proceedings National Academy of Sciences of Unites States of America* **94**:10547-10553.

Pearson WR (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods in Molecular Biology* **132**:185-219.

Rodriguez-Ezpeleta N, Teijeiro S, Forget L, Burger G and Lang BF (2009) 3. Generation of cDNA libraries: Protists and Fungi. *in* J. Parkinson, ed. *Methods in*

Molecular Biology: Expressed Sequence Tags (ESTs). Humana Press, Totowa, NJ.
Roy J, Faktorova D, Lukeš J, and Burger G (2007) Unusual mitochondrial genome structures throughout the Euglenozoa. *Protist* **158**:385-396.
Schneider TD and Stephens RM (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research* **18**:6097-6100.
Smith SW, Overbeek R, Woese CR, Gilbert W and Gillevet PM (1994) The genetic data environment an expandable GUI for multiple sequence analysis. *Computater Applications in Biosciences* **10**:671-675.
Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**:2688-2690.
Tsukihara T and Yoshikawa S (1998) Crystal structural studies of a membrane protein complex, cytochrome c oxidase from bovine heart. *Acta Crystallography Acta* **54**: 895-904.
Vlcek C, Marande W, Teijeiro S, Lukeš J, and Burger G (2011) Systematically fragmented genes in a multipartite mitochondrial genome. *Nucleic Acids Research* **39**:979-988.
Waller RF and Jackson CJ (2009) Dinoflagellate mitochondrial genomes: stretching the rules of molecular biology. *Bioessays* **31**:237-245.

2 Résultats non publiés sur la fragmentation et l'épissage en *trans* du gène *cox1*

Les résultats présentés ici font partie d'expériences réalisées dans le but de mettre en évidence la fragmentation et l'épissage en *trans* du gène *cox1* chez les diplonémides. Ils n'ont pas été inclus dans l'article 1 à cause de la limite d'espace dictée par la revue.

2.1 La fragmentation de *cox1* chez *D. ambulator*, *D.* sp. 2 et *R. euleeides*

Les résultats du Southern Blot (Figure 20) avec une sonde ADN se fixant sur le dernier module permettent de conclure à l'existence de chromosomes portant le module 9 du gène *cox1* de taille variant de 3-10 kb chez ces espèces. Le nombre de modules ainsi que leurs tailles ont été déterminés par PCR et séquençage des chromosomes portant les différents modules de *cox1*.

2.2 Confirmation de la fragmentation de *cox1* chez *D. ambulator*, *D.* sp. 2 et *R. euleeides*.

Les PCR réalisés à partir de l'ADN total, avec des amorces couvrant les modules 1-4 chez *D. ambulator*, 6-9 chez *D.* sp. 2 et 3-6 chez *R. euleeides* n'ont pas permis d'amplifier un produit correspondant à la taille de plusieurs modules. La fragmentation de *cox1* chez les diplonémides a ensuite été confirmée par le séquençage.

2.3 Les chromosomes de *cox1* chez les diplonémides

Pour décrire la fragmentation de *cox1* chez les diplonémides, nous avons amplifié tous les chromosomes portant les 9 modules de *cox1* chez les trois nouvelles espèces et séquencé au total 9 chromosomes et 18 modules génomiques de *cox1*. Les chromosomes de *cox1* de deux classes de tailles de 4.5-10 kb portent des modules de 89-263 pb (Tableau VI).

Les chromosomes portant les modules 5 et 9 de *cox1* chez *R. euleeides* sont encore incomplets malgré le nombre élevé de séquences disponibles pour l'assemblage. Cela s'explique par la présence de nombreuses séquences répétées qui empêchent l'assemblage en un seul contig.

2.4 L'épissage en *trans* de *cox1* chez les diplonémides

Des expériences de Northern Blot réalisées avec l'ARN total, le poly (A) et le module terminal comme sonde chez les trois espèces, montrent des transcrits de tailles différentes (Figure 21). Cela démontre l'existence de transcrits intermédiaires résultant de la liaison de modules de *cox1*.

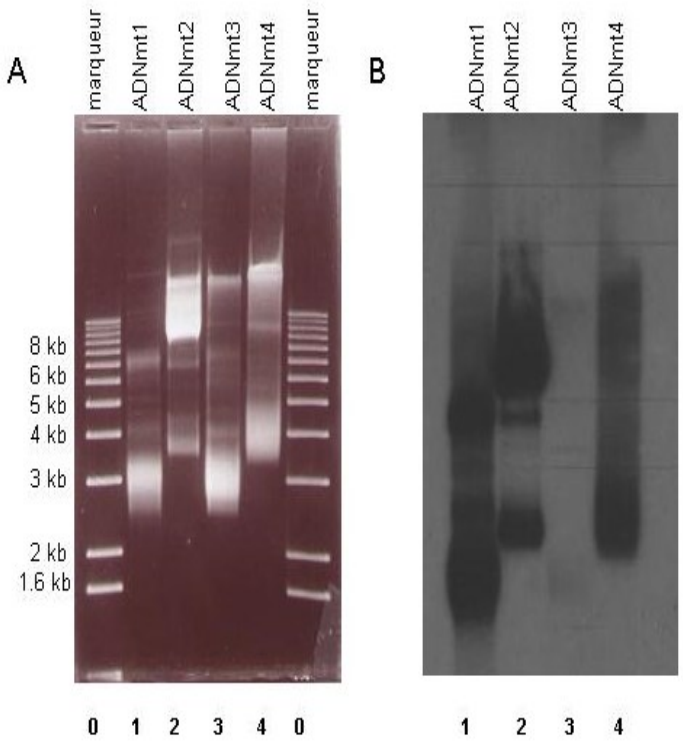


Figure 20. Analyse de l'ADNmt de diplonémides par Southern Blot. A. ADNmt de diplonémides séparés sur gel d'agarose 0.8%. Les puits 0-4 contiennent respectivement le marqueur 1 kb+, de l'ADN total de *D. ambulator*, *D. papillatum*, *D.* sp. 2 et *R. euleeides.*

B. Résultat de l'hybridation des ADNmt avec le module terminal de *cox1*. On obtient des bandes prédominantes de 2, 2.5, 3, 5 kb pour *D. ambulator*, 2.5, 4, 6, 7 kb pour *D. papillatum*, 2.5, 5, 7, 8, 10 kb pour *R. euleeides* et des bandes plus discretes à 1.6, 2, 3. 4 kb pour *D.* sp. 2 correspondants à des chromosomes ou fragments de chromosomes portant le module 9 de *cox1*.

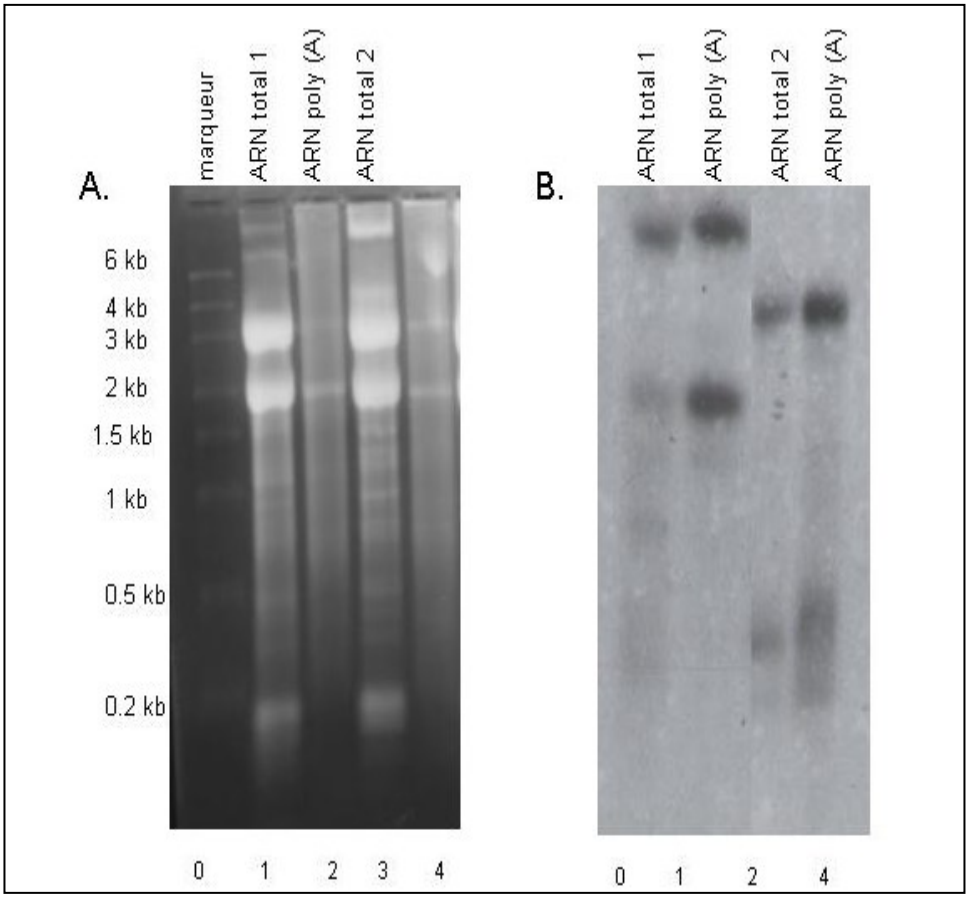


Figure 21. Analyse des transcrits mitochondriaux de *cox1* par Northern Blot. A. ARN total et ARN poly (A) séparés sur gel 1.5% agarose dénaturant (37 % formaldéhyde). Les puits 0, 1, 2, 3, 4 contiennent respectivement le marqueur ARN, l'ARN total, l'ARN poly(A) de *D. ambulator*, l'ARN total et l'ARN poly (A) de *R. euleeides*. B. Détection de transcrits intermédiaires après hybridation des ARN avec le module 9 de *cox1*. Transcrits de *D. ambulator* (0. 2- 6 kb) et *R. euleeides* (0.2-4 kb) détectés après Northern Blot.

Tableau VI. Les modules de *cox1* et les classes de taille[a] des chromosomes chez les diplonémides.

Modules de cox1	*D. ambulator*	*D. sp. 2*	*R. euleeides*	*D. papillatum*
Module 1	192 pb (B)	168 pb (A)	189 pb (A)	195 pb (B)
Module 2	124 pb (B)	122 pb (A)	124 pb (B)	124 pb (A)
Module 3	262 pb (B)	263 pb (A)	263 pb (B)	263 pb (A)
Module 4	222 pb (B)	220 pb (B)	220 pb (B)	220 pb (B)
Module 5	180 pb (A)	179 pb (A)	179 pb (B)	179 pb (A)
Module 6	168 pb (B)	169 pb (A)	169 pb (A)	169 pb (A)
Module 7	90 pb (B)	91 pb (A)	90 pb (B)	89 pb (A)
Module 8	112 pb (B)	109 pb (A)	111 pb (B)	110 pb (A)
Module 9	245 pb (B)	237 pb (A)	246 pb (A)	250 pb (A)

[a]A, B sont les classes distinctes de taille des chromosomes. Chez *D. ambulator* (4.5 et 5 kb), *D.* sp 2 (5 et 10 kb), *R. euleeides* (7 et 8 kb) et *D. papillatum* (6 et 7 kb).

3 Article 2. RNA-level unscrambling of fragmented genes in *Diplonema* mitochondria

3.1 Introduction à l'article

Dans cet article nous avons rapporté entre autre les résultats de mes recherches dont l'objectif était de répondre aux questions suivantes :

- Ya t-il d'autres transcrits mitochondriaux édités à part celui de *cox1*?
- Si oui l'édition par insertion d'uridines est-elle la seule forme d'édition dans la mitochondrie de *D. papillatum*?
- Est-ce que l'édition s'effectue toujours entre deux modules?

Dans ce contexte, nous avons analysé tous les transcrits mitochondriaux caractérisés chez *D. papillatum* en comparant les séquences génomiques et ADNc disponibles de neuf gènes mitochondriaux (*cox2*, *cox3*, *atp6*, *cob*, *rnl*, *nad1*, *nad4*, *nad5* et nad7 et *X1,* un gène non identifié).

Nous avons découvert un deuxième évènement d'édition qui consiste en l'ajout de 3 uridines au bout 3' du module terminal du transcrit du gène *cob*. Cette édition ne s'effectue pas sur un module interne comme dans le cas de *cox1* mais plutôt sur un module terminal.

L'article aborde également d'autres aspects à savoir la transcription des gènes mitochondriaux, la maturation des transcrits et le mécanisme de l'édition de *cox1* chez *D. papillatum*. Outre les tentatives d'identification expérimentales d'ARN guides impliqués dans l'épissage en *trans* et l'édition de *cox1,* l'identification *in silico* d'ARN guides y est également rapportée.

Les expériences de séquençage des différents transcrits mitochondriaux ont été réalisées par plusieurs personnes : William Marande, un ancien étudiant au doctorat du laboratoire, Yifei Yan, étudiant de maitrise à ce moment- là et moi-même dans le cadre de ma thèse.

Yifei Yan a effectué les expériences visant l'identification expérimentale des ARN guides. J'ai réalisé les expériences pour rechercher des gènes édités et pour tester la longueur des ARN antisens (primer extension).

La prédiction *in silico* d'ARN guides impliqués dans l'épissage en *trans* et l'édition a été réalisée par le Dr Turcotte, professeur en informatique à l'université d'Ottawa. J'ai analysé les différentes données expérimentales avec le Dr Burger et Yifei Yan. Tous les auteurs ont contribué à l'article.

Article 2. RNA-level unscrambling of fragmented genes in *Diplonema* mitochondria

Kiethega GN, Yan Y, Turcotte M and Burger G (2013) *RNA Biology* **10**:2, 1-13

3.2 RNA-level unscrambling of fragmented genes in *Diplonema* mitochondria

Georgette N. Kiethega[1#], Yifei Yan[1#], Marcel Turcotte[2], Gertraud Burger[1,3]*

[1]Department of Biochemistry, Université de Montréal, Montreal, Canada

[2]School of Electrical Engineering and Computer Science, University of Ottawa, Ottawa, Canada

[3]Robert-Cedergren Centre for Bioinformatics and Genomics, Université de Montréal, Montreal, Canada

***Corresponding author**: Gertraud Burger

#Both authors contributed equally to the work.

KEYWORDS: trans-splicing, U-insertion RNA editing; *Diplonema papillatum*; Euglenozoa; mitochondria

RUNNING TITLE: RNA-level unscrambling

3.2.1 Abstract

We previously reported a unique genome with systematically fragmented genes and gene pieces dispersed across numerous circular chromosomes, occurring in mitochondria of diplonemids. Genes are split into up to twelve short fragments (modules), which are separately transcribed and joined in a way that differs from known trans-splicing. Further, *cox1* mRNA includes six non-encoded uridines indicating RNA editing. In the absence of recognizable cis-elements, we postulated that trans-splicing and RNA editing are directed by trans-acting molecules. Here we provide insight into

the post-transcriptional processes by investigating transcription, RNA processing, trans-splicing, and RNA editing in *cox1* and at a newly discovered site in *cob*. We show that module precursor transcripts are up to several thousand nt long and processed accurately at their 5' and 3' termini to yield the short coding-only regions. Processing at 5' and 3' ends occurs independently, and a processed terminus engages in trans-splicing even if the module's other terminus is yet unprocessed. Moreover, only cognate module transcripts join, though without directionality. In contrast, module transcripts requiring RNA editing only trans-splice when editing is completed. Finally, experimental and computational analyses suggest the existence of RNA trans-factors with the potential for guiding both trans-splicing and RNA editing.

3.2.2 Introduction

The arguably most eccentric genome architecture and gene structure is found in the mitochondrion of diplonemids (Euglenozoa), a group of free-living unicellular flagellates with phagotrophic or osmotrophic mode of nutrition. Diplonemids are the sistergroup of the notorious kinetoplastids whose members are responsible for serious diseases in humans. The third group within Euglenozoa are the euglenids, which emerged prior to the split of diplonemids and kinetoplastids (Simpson & Roger, 2004).

The first molecular study of diplonemid mtDNA was published in 1999 by L. Simpson's group (reviewed in Maslov*, et al.*, 1999) and indicated that the mitochondrial genome of *Diplonema papillatum* consists of a complex array of small covalently-closed DNA molecules. Today, we know that this mtDNA is composed of hundreds of 6 and 7 kbp long circular molecules termed A-class and B-class chromosomes, respectively (Marande*, et al.*, 2005). Most of the chromosomes' sequence is non-coding and quasi identical among members of the same class. Only a small region, the 'cassette', is unique to each chromosome. The cassette encloses a gene fragment of ~70-350 bp (Marande & Burger, 2007) that is bounded by on average 50-nt long non-coding sequence on each side (Fig. 1A). All genes in mtDNA of *Diplonema* appear to be fragmented; as of now, not a single contiguous gene has been found in this genome. The

cox1 gene for example is broken up into nine pieces and therefore requires nine different chromosomes to specify its coding region. To generate the contiguous *cox1* mRNA, gene pieces are transcribed individually and then assembled by trans-splicing, as is evidenced by transcript intermediates readily visible in Northern hybridization experiments (Marande & Burger, 2007).

The unorthodox genome organization and gene structure of *D. papillatum*, and diplonemids in general (Kiethega*, et al.*, 2011), is contrasted by a rather ordinary set of mitochondrial genes, which much resembles the gene complement in kinetoplastid mitochondria. Genes identified so far in *D. papillatum* mtDNA encode components of the respiratory chain and oxidative phosphorylation, i.e., ATP synthase subunit 6, apocytochrome b, cytochrome oxidase subunits 1-3, NADH dehydrogenase subunits 1, 4, 5, 7, and 8, as well as the mitochondrial large subunit ribosomal RNA (rRNA) (Vlcek*, et al.*, 2011). The gene for the small subunit rRNA is believed to be also present on *D. papillatum* mtDNA - as is the case for all other eukaryotes - but has remained undetected, most probably because its sequence is highly divergent. All recognized mitochondrial genes of *D. papillatum* are trans-spliced.

A well-known mechanism of mitochondrial trans-splicing involves discontinuous Group I or Group II introns, where cognate exons are brought into close proximity through intermolecular pairing that forms a distinctive intron RNA secondary structure (Bonen, 1993, Moreira*, et al.*, 2012). However, even the most sensitive in silico search failed to detect intron-typical sequence patterns, conserved residues, or sequence-complementary motifs at module boundaries (Kiethega*, et al.*, 2011). In the absence of recognizable cis-elements, we postulated trans-active factors that would guide both, trans-splicing and RNA editing (Marande & Burger, 2007, Vlcek*, et al.*, 2011). Here we analyze in a comprehensive fashion post-transcriptional processes in *Diplonema* mitochondria, and identify, by experimental and in silico methods, trans-factors with the potential to direct and control the various RNA maturation steps. Experimental studies of *Diplonema* mitochondrial transcripts have been extremely challenging due to the difficulties to obtain sufficient quantities of cell material, mitochondria, and RNA.

3.2.3 Results

Primary mitochondrial transcripts

Transcription of mitochondrial gene modules in *Diplonema* is thought to start, as in animal mitochondria (Falkenberg*, et al.*, 2007), at the replication origin. Previously, we mapped the origin tentatively by in silico methods to the shared constant region of chromosomes (Fig. 1A, (Vlcek*, et al.*, 2011)). Aiming at a more precise experimental determination of transcription start sites, we now performed *in vitro* RNA capping experiments, since in most mitochondria (with few notable exceptions, e.g., *Neurospora crassa* (Kennell & Lambowitz, 1989)), primary transcripts have a triphosphate 5' end that can be labeled with α-^{32}P-GTP and guanylyl transferase. Nuclear mRNAs, in contrast, are naturally capped during transcription and therefore are not labeled, except cytosolic 5S rRNA (or a portion of it).

In *Diplonema* mitochondria, we expected four major groups of primary gene module transcripts, one each for A-class and B-class chromosomes and one each for orientation "(+)" and "(-)" (Fig. 1B). All these primary transcripts should be above 1.2 kb in size (the minimum distance between modules and shared constant region). This lower-bound size estimate is corroborated by module precursors that were identified in cDNA libraries and by RT-PCR experiments (see below and Fig. 1C). Yet, capping of *Diplonema* total RNA yielded only two, relatively small labeled bands (0.12 kb and 0.3 kb; Fig. 2A). Although the quantity of the labeled material was insufficient to perform RNA sequencing or to use it as a hybridization probe, we still can make specific inferences on the nature of these RNA species. The 0.12-kb molecule is almost certainly cytosolic 5S rRNA based on its size and high abundance in ethidium bromide staining (Fig. 2B; (Sturm*, et al.*, 2001)). In contrast, the 0.3-kb band is apparently of mitochondrial origin, because of its low concentration (not visible by staining) and high capping efficiency; it most likely represents the equivalent of human mitochondrial 7S RNA that primes mtDNA replication (Lee & Clayton, 1998). Human 7S is a stable RNA whose synthesis is sponsored by the promoter for transcription of L-strand encoded

genes. Both the human mitochondrial L- and H2-strand polycistronic transcripts can have nearly full- genome length, but are rather short-lived (Bonawitz*, et al.*, 2006). The same appears to apply to *Diplonema* mitochondria, with primary transcripts processed too rapidly to be detected by the method applied.

End-processing of gene module transcripts

In *Diplonema*, the RNAs transcribed from individual chromosomes undergo multiple maturation steps, which we investigated by three experimental procedures. First, since mitochondrial mRNAs of *Diplonema* are polyadenylated, we constructed classical full-length cDNA libraries by priming the reverse transcription of the 1st DNA strand with an anchored oligo-dT primer that anneals with the proximal region of the poly(A) tail; the 2nd DNA-strand synthesis was primed with an oligonucleotide binding to all cDNA 3' ends (see Methods). Second, double-stranded cDNA was produced as above, but transcripts were PCR-amplified using various combinations of gene-specific primers. A third type of experiment involved RNA circularization followed by RT-PCR with diverse pairs of 'divergent' gene-specific primers (Figure. 3A). These experiments detected, in addition to (mature) mRNAs, two kinds of incomplete transcripts, one containing multiple modules ('oligo-module transcripts') and the other including a single module ('mono-module transcripts'). We also encountered several transcripts containing exclusively module-flanking regions, with one terminus corresponding exactly to the nucleotide adjacent to a module. Such large chunks of flanking regions appear to be liberated by precise endonucleolytic cleavage. Supplementary Tables S1 and S2 compile detailed information on transcripts that include modules or exclusively flanking regions, respectively. These tables also list the corresponding clones that are referred to in the following paragraphs.

Among mono-module transcripts, we found fully and partially processed modules, the different types of which are depicted in Fig. 1D-F. Fully processed modules consist exclusively of coding region (e.g., *cox1*-m5 clone dp7341), except for the 5'-terminal ('first') module of genes, which retains a 5' untranslated leader (5'UTR) that is typically

~25 nt long (e.g., *cob*-m1, clone dp5996). Partially processed mono-module transcripts include flanking regions of non-coding sequence that are up to ~1150 nt long reaching far into the chromosome's constant region (see Fig. 1A, C and Supplementary Table S1). Flanking regions may border the upstream, the downstream, or both sides of the module (e.g., *nad4*-m7).

As mentioned above, gene modules are encoded on either A- or B-type chromosomes and in either orientation with respect to the constant region (referred to as A(+), A(-), B(+), and B(-) (see Fig. 1B and Supplementary Table S1, column 2). Interestingly, immature module transcripts whose adjacent regions reach into the constant regions of chromosomes have the potential to pair with one another, notably A(+) with A(-) precursors, and B(+) with B(-) precursors (Fig. 4A). In fact, intermolecular hybridizations are quite likely to occur, given the relatively high steady-state concentration of precursors in mitochondria. Such pairing would evidently not align modules in the correct order for trans-splicing, but might allow to 'herd' the hundred or so distinct module transcripts for further processing by a dedicated machinery.

Special cases are transcripts of 3'-terminal ('last') modules. These occur with region (e.g., *cox1*-m9, clone dp5927). We detected only a few last-module transcripts that are 3'-processed but not poly-adenylated, suggesting that the two steps are tightly coordinated.

A tentative estimation of the relative abundance of processing intermediates indicates processed and unprocessed ends, as well as with a poly(A)-tail attached directly 3'-adjacent to the coding that the steady-state concentration of mono-modules retaining 5'- and 3'-adjacent regions is generally higher than that of partially and fully end-processed transcripts (see Figs. 1D-F; Supplementary Table S3).

Trans-splicing of gene module transcripts

While single-module transcripts described above provide insight into their end processing, transcripts containing several modules inform us about how trans-splicing

proceeds. The observed oligo-module transcripts include various numbers of modules, covering virtually any interval of the mRNA (e.g., *cox1*-Modules 3 to 6; *cox1*-Modules 6 to 9; Supplementary Table S4). Further, modules are all arranged in correct order, thus representing putative intermediates of the trans-splicing process.

We analyzed in how far trans-splicing depends on module end-processing and, in the case of terminal modules, polyadenylation. Several oligo-modules were found to retain a 5'- or a 3'-flanking region (Supplementary Tables S1, S4). Moreover, oligo-modules including a last module may be poly-adenylated or not (e.g., *cox1*-m9, clones dp5977 and dp0655). Obviously, partially processed modules can readily engage in trans-splicing, and polyadenylation of the last module seems not required for joining with its upstream partner. These data taken together allowed inferring the assembly line by which transcripts are built in *Diplonema* mitochondria (Fig. 4B-D).

Non-encoded nucleotides in mitochondrial transcripts

In general, genomic and cDNA sequences of *Diplonema* mitochondria are congruent - yet, with a few notable exceptions. Most conspicuous is the occurrence of six non-encoded Us in the *cox1* transcript exactly between Modules 4 and 5 (Marande & Burger, 2007), and this editing event is evolutionarily conserved across diplonemids (Kiethega*, et al.*, 2011). Here we examine how exactly these extra Us are added. They may be inserted after ligation of the two modules, or alternatively, prior to ligation attached either to the 3' end of Module 4 or to the 5' end of Module 5 (Supplementary Fig. S1). The insertion-scenario implies the existence of an *cob*-mRNA sequence plus the first position of the stop codon, which is completed to UAA via polyadenylation. We predict that the edited terminal *cob* Module 6 also carries a 3'-phosphate group just like the edited Module 4 of *cox1*, and that U appendage to the *cob* module is a prerequisite for polyadenylation.

We also observed differences between mitochondrial genomic and transcriptomic sequences that do not involve Us. Differences pertain to the termination codon, which is generated only post-transcriptionally by polyadenylation as experimentally confirmed

for *atp6* (Fig. 5B), but most likely also applying to all other genes. The only terminal module enclosing an encoded stop codon is that of *cox1*, but curiously, polyadenylation creates an additional stop codon (Fig. 5C). Post-transcriptionally generated stop codons have first been reported for human mitochondria, where, in contrast to *Diplonema*, it coincides with an extreme reduction of both intergenic regions and the overall genome size (Anderson*, et al.*, 1981).

Experimental detection of postulated RNAs guiding post-transcriptional processes

Earlier we showed by rigorous in silico analyses that trans-splicing in *Diplonema* mitochondria is most certainly not directed by cis-elements, i.e., sequence motifs located in modules or their flanking regions (Kiethega*, et al.*, 2011). Therefore, we posited trans-acting matchmaking factors, which could be RNA, protein or DNA molecules. Here we describe a set of experiments that test for the existence of RNAs that may guide module trans-splicing as well as RNA editing. We refer to these hypothetical molecules as post-transcriptional processes-guiding RNAs (ppRNAs).

First we searched for gRNA-like molecules known from kinetoplastid mitochondria to direct RNA editing. These RNAs are characterized by 50-70 nt length, high abundance, a 5'-triphosphate and a 3'-poly(U) tract (Blum & Simpson, 1990). Yet, electrophoretic separation of RNAs, capping experiments (see above) and in vitro incorporation of radiolabeled uridine (not shown) did not reveal RNA species in *Diplonema* akin to kinetoplastid gRNAs. Since in *Diplonema*, the posited ppRNAs may be present at only low concentration, the more sensitive RT-PCR methodology was employed on total *Diplonema* RNA. We inquired for molecules that are antisense to mRNA and cover several or even all module junctions of a gene at once. One experiment aimed at anti-sense transcripts including Modules 5 to 9 of *cox1* (spanning four junctions), and antisense transcripts of three additional genes spanning two to four junctions were tested as well. However, no amplicons were detected (results not shown), refuting the hypothesis of long anti-sense RNAs directing trans-splicing of multiple modules simultaneously.

A second series of experiments tested for the presence of ppRNAs that cover only a single module junction. These RT-PCR experiments used a primer pair that targets the central region of hypothetical ppRNAs (Fig. 3B). Here we obtained amplicons of the expected size and sequence for all five of the examined *cox1* junctions, M2/M3, M3/M4, M4/M5, M5/M6, and M8/M9 (Fig. 6); Table 1 ("Central') compiles the results of the individual experiments. However, the exact transcript sequence that served as template for the RT product cannot be inferred from these experiments. One reason is that the primer pairs were designed to anneal a few nucleotides adjacent to module junctions (to increase the chance of detecting the postulated molecules), with the consequence that the resulting amplicons include only two to eight 'novel' nucleotides (Fig. 6, red labels), while most of the sequence originates from the primers. Second, the primer-derived sequence of amplicons may not fully correspond to the sequence of the targeted RNA. This is because primers were designed with the assumption that the hypothetical ppRNA is an exact reverse complement of pre-mRNA, but the pre-mRNA:ppRNA duplex region may contain G:U pairs. In addition, the primers may extend beyond the 5' and 3' termini of the hypothetical ppRNA.

In an attempt to characterize the distal regions and the length of the detected RNAs, we conducted 'divergent' RT-PCR on circularized RNAs (Fig. 3A), expecting to discover sequences that match presumptive mitochondrial coding regions (in cassettes). Yet, no significant hit with unassigned cassettes was found, and amplicon clones differ in sequence among each other, suggesting that these particular RT-PCR products are mostly spurious (Table 1, 'Distal'). Nevertheless, a single candidate (dp8189) was found to match junction M5/M6 and covering ~60 nt of both modules in antisense direction (see Supplementary Table S6 for a detailed analysis and sequence). The reason why divergent RT-PCR yielded such sparse results could be due to the short length of the target RNA together with mismatches between primer and target RNA. Finally, the length of potential ppRNAs was also investigated by primer extension (run-off reverse transcription) aiming at the antisense RNA covering the *cox1* junctions M4/M5, which appears to be the highest expressed ppRNA candidate (see Table 1). However, no signal

was detected, most likely because of the lower sensitivity of this method compared to RT-PCR.

Given the limited sequence information obtained for ppRNA candidates, it is not possible to determine unambiguously their coding regions on the (nuclear or mitochondrial) genome. Therefore, we mapped the candidates' genomic positions by silico analyses, as described in the following section.

In silico detection of postulated trans-factors guiding post-transcriptional processes

Potential trans-acting elements directing *cox1* trans-splicing and RNA editing were searched computationally in the available genome and transcriptome sequences. The in silico analyses were designed in a way to test many more scenarios and in a more rigorous way than would be feasible experimentally. For example, we permitted guiding factors not only to be (nucleus and mitochondrion-encoded) RNA but also DNA molecules. (Note that RNA and DNA uptake by this organelle has been well documented (Koulintchenko*, et al.*, 2003); reviewed in (Salinas*, et al.*, 2008)). Moreover, we allowed guiding factors to bind to as few as six contiguous nucleotides ('anchor', parameter a=6, see Fig. 3C) in each of the neighboring modules, and three out of six pairs in the module/guide duplex region may be G:U.

The first and simplest guide model we searched for required the binding sites of the trans-factor to be directly adjacent to the module junction, and the trans-factor to be collinear with the sequence across the junction (distance d_1, d_2=0; no loop or bridge, L=0, see Fig. 3C). Further, each guide was scrutinized for the potential to mis-assemble, i.e., to direct joining of non-cognate modules (e.g., Modules 2 and 4). If so, these guides were removed. Finally, a data set must include guides for at least six out of eight *cox1* junctions (to account for incomplete genome sequences), otherwise, candidates are not reported. For this model, we detected 34 candidates of distinct sequence in the mitochondrial genome data, nearly four times as many in nuclear ESTs, and close to 3,000 candidates in nuclear genome data (Table 2**,** column 3). The guide candidates detected in mtDNA match all *cox1* junctions except M1/M2 and M3/M4 (Supplementary

Table S7). Surprisingly, only one candidate is predicted to reside on one of the six fully sequenced chromosomes (out of an estimated >100 mitochondrial chromosomes residing in mtDNA). This finding refutes our earlier hypothesis that ppRNAs are encoded in the constant region of all chromosomes.

More complex guide models were tested as well. We allowed that the guide-binding sites in modules may occur at various distances from the corresponding junction (d_1, d_2>0) and that the two module-binding sites in guides are at various distances from one another (L>0). We also considered that the two binding sites on the guiding molecule may be arranged in permutated order (Topologies 1 and 2; Fig. 3C, D). Although permutated structural RNAs are not without precedent (e.g., (Keiler*, et al.*, 2000)), they have been widely ignored in computational searches (but see (Soma*, et al.*, 2007)). The combination of the above parameters yields nearly 720,000 distinct guide classes, i.e., sets of guides that share identical parameters and conformation (Supplementary Table S8). For this extended guide model, candidates were detected in all data sets and in very large numbers (Table 2, columns 4-7; for statistics of search results, see Supplementary Table S8). To summarize, the *Diplonema* mitochondrial genome (as well as its nuclear genome) indeed contain candidates for trans-factors that direct trans-splicing and RNA editing in mitochondria.

3.2.4 Discussion

End processing and trans-splicing of gene modules proceed in parallel

We investigated in a comprehensive fashion the post-transcriptional processes in *Diplonema* mitochondria, processes that are involved in the generation of full-length transcripts from multiple, separately transcribed gene pieces (modules). Analysis of the immature transcript population (mono-module and oligo-module transcripts) provided insight into the transcription, transcript-end processing, and trans-splicing of gene modules, allowing reconstruction of the full post-transcriptional processing pathway in *Diplonema* mitochondria. Apparently, transcription of individual modules initiates and

terminates in the shared constant regions of chromosomes (see Fig. 1A) as evidenced by mono-module transcripts including long flanking regions (see Fig. 1C). These non-coding, flanking regions are subsequently removed from mono-module transcripts by precise endonucleolytic cleavage, since we detected sizeable transcripts starting or ending directly at a module boundary and containing exclusively flanking region. Still, adjacent regions included in module precursors are tremendously variable in length (see e.g., *cox1*-m2 with 22 to 943 nt-long 5' extensions; Supplementary Table S1, Fig. 1C), which might reflect that additional, exonucleolytic trimming is at work, but prematurely-stalling reverse transcriptase reactions in RT-PCR experiments might contribute to this phenomenon as well. Most important, the results show that 5'- and 3'-end processing of mono-module precursor transcripts occurs independently from one another, and that 5'-end processing of terminal modules is independent of the polyadenylation step, which apparently takes place rapidly after 3'-end processing.

Mono-modules are able to engage in trans-splicing as soon as one of the two termini is end-processed (Fig. 4B-D), since many oligo-module transcripts include 5' or 3'-flanking regions. In addition, trans-splicing seems to start with any pair of cognate modules, without imposing any directionality (such as transcript elongation from 3' to 5'), which is indicated by the simultaneous finding of oligo-modules covering either a 5'-terminal, a central, or a 3'-terminal portion of the mature transcript. Taken together, module-end processing and trans-splicing in *Diplonema* mitochondria is highly parallelized – not serial. Further noteworthy is that trans-splicing in this system is highly accurate, as mis-joined modules, across genes or within genes but out of order, have not been observed.

Candidates identified for trans-factors that guide post-transcriptional processes

Earlier we showed that trans-splicing in *Diplonema* mitochondria is most likely not guided by sequence elements in cis (5), and therefore, we postulated the existence of trans-acting factors (termed post-transcriptional processes-guiding RNAs (ppRNAs). Here we were able to confirm experimentally that low-abundant anti-sense RNAs indeed

exist that cover a single module junction (Fig. 6), thus having the potential to direct trans-splicing, as well as RNA-editing. Although it was only possible to determine a small sequence portion of the presumed ppRNAs (see Results), data suggest that these molecules are rather short. The limited sequence information also precluded to map unambiguously the presumed guides to the nuclear or mitochondrial genome. It should be noted, however, that ppRNAs may not be directly encoded by the genome after all, but alternatively, reverse-transcribed from mRNAs and then transmitted epigenetically to daughter cells, as in the case of RNA-mediated genome rearrangements in ciliates (Nowacki*, et al.*, 2008)).

Instead of mapping the presumed ppRNAs to genome sequence, we computationally predicted guide candidates for *cox1* in the genome sequences of *Diplonema*. This analysis allowed testing numerous different structures and conformations (see Fig. 3C, D) and to exclude solutions that lead to module misjoining. As expected, the large nuclear genome has the potential to encode numerous trans-acting factors, but the most important result is that in the available mtDNA sequence (which represents an estimated 50% of the entire mitochondrial genome), we located 34 distinct ppRNA candidates for six out of eight *cox1* junctions (Table 2). This finding corroborates the view that the mitochondrion itself encodes its guides for trans-splicing and RNA editing of mitochondrial genes in *Diplonema.*

RNA editing at two sites by U-appendage

Uridine (U)-based mitochondrial RNA editing is known from plants and kinetoplastids, where it involves nucleotide modifications and insertion/deletions, respectively (reviewed in Gray, 2003). In contrast, RNA editing in *Diplonema* mitochondria relies on U addition. More precisely, editing at the first site, in *cox1*, proceeds by appending six Us to Module 4 of *cox1*, prior to trans-splicing to Module 5 (Fig. S1). The finding that Us are only found attached to Module 4, but not with Module 5 makes it highly unlikely that the non-encoded Us represent an overlooked mini-module, because modules have no preference for either of their two neighbors to join

with the first. Note that we never encountered intermediates with an incomplete or excessive number of Us, indicating that U attachment is rapid and highly precise, and tightly coordinated with module joining. Since *cox1* Modules 4 and 5 apparently do not trans-splice prior to U addition, RNA editing is a crucial prerequisite for the biosynthesis of the *cox1* mRNA as a whole. The second site of RNA editing in *Diplonema* mitochondria, reported here for the first time, is located at the 3'end of the terminal *cob* module and involves appendage of three Us (Fig. 5A).

Non-encoded Us in mitochondrial transcripts were also observed in a close relative of dinoflagellates (Slamovits*, et al.*, 2007), where two mRNAs carry a U tract at their 5'end. It remains unknown whether the extra nucleotides originate indeed from RNA editing or rather from sloppy transcription. Finally, RNA editing involving terminal homo-oligomer addition has been discovered in a single mitochondrial gene (*cox3*) of select dinoflagellates. Their *cox3* gene is bipartite and the transcripts of both fragments are oligo-adenylated as are all other mRNAs in these organisms. Interestingly, five nucleotides of the A-tail from the upstream *cox3* fragment are retained in mRNA (Jackson*, et al.*, 2007).

Parallels of mitochondrial post-transcriptional processes in *Diplonema* and kinetoplastids

We showed above that processing of module-precursor RNAs in *Diplonema* mitochondria involves base-precise endonucleolytic cleavage, suggesting that not only trans-splicing and RNA editing, but also end-processing may be guided by the postulated trans-factors. Interestingly, end-processing in *Diplonema* mitochondria is formally equivalent to the first step of RNA editing in kinetoplastids. There, pre-mRNA is cleaved within a sequence stretch to which a gRNA is bound, the cut site being immediately adjacent to the anchor region. The reaction is performed by endoribonucleolytic enzymes that are part of the editosome (reviewed in Stuart*, et al.*, 2005).

It came as a surprise that the RNA-editing intermediate *cox1* Module4-UUUUUU (m4-6xU) carries a 3'-nucleoside monophosphate (3'NMP). This may reflect a mechanistic similarity between U-'appendage' editing in *Diplonema* and U-insertion editing in kinetoplastids. In *Trypanosoma brucei*, in vitro assays with mitochondrial extracts have shown that TUTase occasionally appends more Us at the pre-mRNA cleavage site than specified by the corresponding gRNA, and it was suggested that excess nucleotides are trimmed by an exonuclease (3' => 5' exoUase; Byrne*, et al.*, 1996). Functional characterization of mitochondrial proteins in trypanosomes revealed three enzymes implicated in U-insertion/deletion RNA editing, TbMP42, TbMP99, and TbMP100 (Brecht*, et al.*, 2005, Kang*, et al.*, 2005). TbMP42 displays in vitro exoUase activity leaving 3'- monophosphate ends, which, as in *Diplonema*, cannot be ligated with the 5' terminus of the downstream pre-mRNA cleavage fragment. In turn, TbMP99 and TbMP100 exhibit 3'-specific nucleotidyl phosphatase activity converting the 3'NMP to a 3' hydroxyl group, which permits pre-mRNA re-ligation (Niemann*, et al.*, 2009). This dephosphorylation step was proposed to serve as quality control in trypanosome RNA editing: only when the number of inserted nucleotides permits full pairing with the gRNA, the 3'-monophosphate would be removed from the terminal U, and re-sealing of pre-mRNA would proceed (Niemann*, et al.*, 2009).

Kinetoplastid editosomes include virtually all catalytic activities required for post-transcriptional processes in *Diplonema* mitochondria, notably an endoribonuclease for module end processing, a TUTase for U addition, an exoUase leaving the 3'NMP, a 3'-nucleotidyl phosphatase that 'repairs' 3' termini generated by this exonuclease, and finally, RNA ligase for module joining. This raises the question whether in *Diplonema* mitochondria these activities are exerted by homologs of the kinetoplastid enzymes and whether the enzymes are also organized in a multi-functional protein complex. If so, this might allow us to trace the basic eukaryotic machinery from which the kinetoplastids' editosome may have evolved. However, it is equally possible that the modes of transcript maturation in *Diplonema* and kinetoplastid mitochondria are fundamentally different. For example, module ligation in *Diplonema* mitochondria might not be an

enzymatic, but rather a yet undescribed RNA-catalyzed reaction. To address this question, it will be required to characterize the catalytic entities that carry out the post-transcriptional processes in *Diplonema* mitochondria.

3.2.5 Conclusion and outlook

Cis- and trans-splicing of traditional introns (spliceosomal, Group I, Group II, or archaeal/tRNA introns) intimately links removal of flanking sequences and exon joining, so that splicing intermediates are generally difficult to examine. In contrast, trans-splicing in *Diplonema* mitochondria takes place in clearly separated steps with readily detectable intermediates, and therefore permitted straightforward investigation of the underlying processes as reported here.

This study detected candidates for trans-factors that direct trans-splicing and RNA editing in *Diplonema* mitochondria. The logical next step is to validate the predicted function of these candidates. For example, we expect engineered DNA or RNA guides to mediate trans-splicing of non-cognate module transcripts, and to direct appendage of an arbitrary number of Us to a gene module transcript not uridylated in vivo. Another approach, recently initiated in our laboratory, is to identify and isolate mitochondrial protein complexes of *Diplonema* that display trans-splicing and RNA editing activity, and investigate which of the complex constituents act as guide. This approach is suited to detect guides that are not only RNA or DNA, but also protein molecules.

A more general question bears on the biological role of such complicated post-transcriptional processes. We believe that they offer an effective handle for regulation at various levels: polyadenylation generates stop codons acting on the effectiveness of mRNA translation; gene module-end processing controls the steady-state levels of the building blocks from which mRNAs are made; and finally, RNA editing events act as checkpoints during transcript maturation.

3.2.6 Material and methods

Sequences deposited in public-domain databases

About 17 kbp newly determined sequences were deposited in GenBank under the accession numbers JQ302962, JQ314396, and JQ302963. These entries contain the sequences of the entire chromosome A4005 that carries Module 8 of *nad7*, the entire chromosome A3216 that carries the 3'terminal module of the yet unidentified gene X2, and the 3'terminal piece of the gene specifying mitochondrial large subunit (LSU) rRNA, plus adjacent regions.

Strain, culture and extraction of mtDNA and RNA

Diplonema papillatum (ATCC 50162) was obtained from the American Type Culture Collection. The organism was cultivated axenically at ~20ºC in artificial seawater enriched with 1% fetal horse serum (Wisent) and 0.1% Bactotryptone. Mitochondrial DNA was extracted from an organelle-enriched fraction isolated by differential and sucrose gradient centrifugation (Lang & Burger, 2007). RNA was extracted either from the mitochondria-enriched fraction or from total cell lysate by a home-made Trizol substitute (see (Rodriguez-Ezpeleta*, et al.*, 2009)). Residual DNA was removed from RNA preparations either by RNeasy (Qiagen) column purification or by digestion with RNase-free DNase I (Roche) followed by phenol-chloroform extraction.

RNA Capping

DNase-treated RNA was labelled with α-[^{32}P]-GTP in the presence of capping enzyme (ScriptCap, Epicenter Biotechnologies), followed by phenol-chloroform extraction and electrophoretic separation in denaturating polyacrylamide gels of various concentrations (5%, 12%, 16%, and 4%-10% and 10%-20% gradients).

Primer extension

Run-off transcription experiments aimed at determining the length of potential antisense guiding RNAs. We followed the protocol devised by Promega (http://www.promega.com/resources/protocols/technical-bulletins/0/primer-extension-system-amv-reverse-transcriptase-protocol/). Briefly, a primer that was labelled at its 5'end with γ-[^{32}P] -ATP was annealed with 4, 50, or 200 µg of DNAse-treated poly(A), mitochondria-enriched, or total RNA, respectively, and then incubated with AMV reverse transcriptase (Roche) in the presence of 1 mM dNTPs and 40 µM pyrophosphate. For positive controls we used the primers dp145, dp153 and 138 that anneal with *cox1* Modules 1 and 5, and the terminal module of *rnl*, respectively. These controls gave rise to predominant products of 220, 180, and 340 nt arising from reverse transcription of the corresponding processed modules, together with minor band representing precursors and trans-splicing intermediates (for primer sequences, see Supplemental Table S9). Negative controls left out either RT or RNA. Oligonucleotide dp207 was used for primer extension of the hypothetical antisense RNA that is complementary to the M4/ M5 junction of *cox1*. The samples were separated on an 8% poly-acrylamide gel (19:1) containing 7 M urea, resolving a size range from 20 to 1000 nt. As size markers served the φX174 (Hinf) marker (Promega), the low-range RNA ladder, and the 1-kbp plus DNA ladder (Fermentas) that we also end-labelled with γ-[^{32}P] -ATP. After migration, the gel was exposed on an X-ray film to visualize the bands.

RNA circularization

DNase-treated RNA was incubated with tobacco acid phosphatase (TAP, Epicenter) and T4 polynucleotide kinase (PNK, New England Biolabs). We used both the unmodified enzyme M0236L that possesses 3'-phosphatase activity, and the engineered form M0201L without this activity. RNA was diluted to 20 ng/µL and circularized using T4 RNA ligase (Roche).

RT-PCR

The first strand (cDNA) was generated with Powerscript reverse transcriptase of the Creator Smart cDNA library construction kit (Clonetech) or avian myeloblastosis virus (AMV) reverse transcriptase (Roche). PCR was performed with the Takara PCR kit (Bio Inc.), typically for 35 cycles. Generally, two gene-specific primers were used (Fig. 3A, B), but for certain RT-PCR experiments, PCR amplification was conducted with only one gene-specific primer (for first-strand synthesis) plus the Smart IV primer that anneals with the overhanging G residues at the 5'-end extension of the first-strand DNA (Rodriguez-Ezpeleta*, et al.*, 2009). Primer sequences are given in Supplemental Table S9. For all RT-PCR experiments aiming at the detection of RNAs that mediate trans-splicing and RNA editing, a negative control was performed where no template RNA was added. Other negative control experiments involved the use of a primer combination that is not expected to yield an amplicon, notably RT-PCR where one of the two primers had an insertion or mismatches at its 3'-end compared to the target sequence, and another control where both primers would bind to the same strand. Controls without template did not yield a (visible) product, whereas controls with inappropriate primers produced amplicons of very low amounts. Sequencing showed that the latter RT-PCR products were artefactual, originating from unspecific priming of the sense-strand.

Cloning and sequencing of amplicons

Amplicon termini were rendered blunt with T7 DNA polymerase and the Klenow fragment of DNA polymerase I (New England Labs), agarose gel-purified, phosphorylated with T4 PNK (New England Labs) and ligated into the vector pBFL6cat, which is an in-house constructed, small pBlueScript derivative. cDNA libraries were cloned into pDNR-LIB (Clonetech). After transformation into *E. coli* DH5α, plasmid DNA was extracted using the Qiagen 96-well mini-prep kit. Sequencing reactions were performed with the BigDye Terminator v3.1 Cycle Sequencing Kit from Applied Biosystems and sequenced on an ABI 370 Analyzer.

Clustering of reads and sequence assembly

Reads obtained in experiments aiming at the detection of trans-splicing guide RNAs contained large numbers of identical sequences and therefore were clustered, prior to analysis, with the tool CD-Hit (Huang, *et al.*, 2010). Default parameters were used except for the c and g parameters that were set to 0.9 (90% identity) and 1 (a sequence is clustered into the most similar cluster), respectively. Representative reads obtained by clustering were assembled by phred/phrap (Ewing & Green, 1998) using highly stringent parameters (-minmatch 300, -maxgap 10 –repeat_stringency 0.99 –shatter_greedy –q 95 –penalty -9 –minscore 200). Contigs were inspected using consed (Gordon, 2003) to identify potential misassembly. The consensus sequence in Masterfile format was generated using the in-house tool cosmea. Bioinformatics tools and Masterfile grammar are described at http://megasun.bch.umontreal.ca/ogmp/ogmpid.html. Software is available on request.

Analysis of amplicon sequences

For each contig in a Masterfile, primers were located by BLAST searches, and inserts by either BLAST or FASTA searches (Altschul, *et al.*, 1990, Pearson, 2000) against available mitochondrial (~250 kbp) and nuclear (~8 Mbp) genome, and EST (20 kbp) sequences of *D. papillatum*. Using the in-house MotSearch program, we searched regular expressions corresponding to hypothetical guiding RNAs whose sequence is complementary to module junctions, by allowing in addition to canonical base pairs also G:T pairs.

Computational detection of guiding RNAs or DNAs

Guiding RNAs and DNAs were searched in all available mitochondrial genomic DNA, mitochondrial cDNA, nuclear EST and nuclear genome sequences. The mitochondrial genomic sequences (~350 kbp in total) represent approximately 50% of the genome and consist of six completely sequenced chromosomes and two collections

of incomplete chromosome contigs. The mitochondrial chromosomes carry the modules *cox1*-m9, *cox1*-m9 (second copy), *cox1*-m4, geneX2-m(k), *nad7*-m6, and *nad7*-m8. Chromosome names (sizes and NCBI acc. numbers in parentheses) are: dp3207L.all (5,856 nt; EU123536); dp3208L.all (5,661 nt; HQ288823); dp3209bT.all (7,182 nt; EU123537); dp3216-X2-L.all (; 5,763 nt); dp4001.all (*nad7*-m6; 5,794 nt; HQ288824); dp4005.all (5,763 nt; JQ302962). The two overlapping collections of incomplete chromosome contigs are dpapimt.all (219,754 nt, sequenced inhouse) and prag-mt-readings-phred-assembled_2010apr19.all (377,456 nt, sequenced by the Institute of Molecular Genetics, Academy of Sciences of the Czech Republic in Prague). The mitochondrial cDNAs (dpapemt.all_stringent; 68,404 nt) include the GenBank entries HQ288819-22, EU123538, and JQ302963. Nuclear EST sequences (4,952 nt; generated by us previously) were downloaded from our publicly accessible TBestDB (http://tbestdb.bcm.umontreal.ca). Nuclear genome data (Vlcek*, et al.*, 2011) included 79,784 contigs of 93,641,047 nt total length and probably represent 50-75% of the entire genome.

In these data sets, we searched guiding trans-factors with the following parameter combinations: match length of guiding RNA in Module i and Module i+1, a=6; distances of match from module boundary d1, d2=0,1,..83; bridge length L=0,1,..50; Topology 1 and 2. For an illustrative representation of parameters, see Fig. 3C, D. The combination of all parameters yields 719,712 guide groups. The underlying assumption is that all guide molecules adhere to one and the same group. We searched for all these groups unless the number of detected candidates was excessively large (>100,000) for the first ten groups tested, so that a similarly large number of candidates could be expected for all other groups. The detected guide candidates were filtered in order to eliminate those that would guide module mis-assembly (e.g., joining of Module 1 with Module 3). The groups were screened for their capacity to guide the joining of at least 6 out of the 8 *cox1* junctions, otherwise the group was discarded. We calculated the number of guides located at different positions in the query sequence and, in addition, the number of guides with distinct sequence. Note that even guide groups of L=0 can

have several members of different sequence, since G:U pairs are permitted at any position in the duplex regions. The algorithm is described in the Supplementary Information.

The search in the nuclear data set was executed on a Sun SPARC Enterprise M9000 server with 64 quad-core 2.52 Ghz Sparc64 VII processors capable of Chip Multi Threading with 2 hardware threads, and 2 TByte memory. Execution time for searching Topology 1 (384 guide groups) and Topology 2 (306 guide groups) was 54.20 h and 7.90 h, respectively. Smaller data sets were analyzed on an iMac 2.66 GHz quad-core Intel Core i5.

Table 1. Experimental detection of RNAs potentially guiding *cox1* trans-splicing and RNA editing.

Targeted ppRNA region[a]	*cox1*-junction	Primers used	Clone series from separate experiments[b]	Obtained sequences (nr. of distinct types / nr. of clones of a given type)[c]
Central	Module 2 / 3	dp88 (RT) + dp80	dp7901-96	
			dp8373-96	**ppRNA candidates (1/32)**
			dp8401-24	Spurious sequences (14/21)
			dp9237-72	
	Module 3 / 4	dp146 (RT) + dp147	dp8425-72	**ppRNA candidates (1/1)**
			dp9273-84	Spurious sequences (2/3)
	Module 4 / 5	dp129 (RT) + dp109	dp6401-96	
			dp6501-96	**ppRNA candidates (1/67)**
			dp6601-96	Spurious sequences (37/59)
		dp129(RT) + Smart	dp7101-96	
			dp6901-96	
	Module 5 / 6	dp150 (RT) + dp151	dp8473-96	**ppRNA candidates (1/5)**
			dp9285-96	Spurious sequences (1/1)
	Module 8 / 9	dp154 (RT) + dp41	dp8649-96	**ppRNA candidates (1/1)**
			dp9337-96	Spurious sequences (16/61)
Distal	Module 2 / 3	dp138 (RT) + dp139	dp8001-24,37-48	Spurious sequences (7/30)
	Module 3 / 4	dp148 (RT) + dp149	dp8149-72	Spurious sequences (2/24)
	Module 5 / 6	dp152 (RT) + dp153	dp8173-96	**ppRNA candidates (1/1)**
				Spurious sequences (6/14)
	Module 7 / 8	dp84 (RT) + dp41	dp7801-96	Spurious sequences (34/73)
		dp141 (RT) + dp140	dp8025-36,49-96	

[a]Determination of central ppRNA regions by 'convergent' RT-PCR, and of distal regions by 'divergent' RT-PCR on circularized RNA (see Fig. 3A, B).
[b]Each clone series (dp7901-96, dp8373-96, etc.) was obtained from a separate RT-PCR reaction. Two different RNA preparations were used. In the experiments targeting the junction of Module 4/5, both preparations were used (Preparation #1 for dp64xx, dp65xx, dp66xx, and dp69xx, and Preparation #2 for dp71xx). The number and nature of the resulting amplicons was very similar. All other experiments were conducted with Preparation #2.
[c]Sequences designated as of the same type are ≥99% identical in the stretch between the primers; those designated ppRNA candidates are 100% identical in this stretch. Additional but irrelevant differences between reads occur at their start and end, where the primers may be present fully or only partially, and adjacent vector sequences may be included or not. Sequences considered as originating from ppRNA candidates must

suffice the following criteria. In experiments targeting the central region of the hypothetical ppRNA, candidates must carry both primers in correct orientation, further, the primer used for RT must prime the antisense transcript, and the sequence between these primers must match that of the corresponding module junction in mRNA (U:G pairs allowed, but not insertions/deletions). In the case of experiments targeting the distal region of the hypothetical ppRNA, candidates must carry both primers in correct orientation with >=1 nt in between, and either match yet unassigned mitochondrial coding regions (within cassettes) or must be the predominant type of cloned amplicons. 'Spurious sequences' include (i) mitochondrial sequences where one or both primers have annealed unspecifically at a site with <60% sequence identity on the sense (coding) or antisense strand; (ii) nuclear sequences; (iii) sequences not matching the available mtDNA or nuclear genome sequences or the vector. A detailed listing of the results, the characterization of spurious sequences, and the actual sequences of relevant clones are compiled in Supplementary Table S6 and the corresponding footnote.

Table 2**.** Computational detection of *cox1* trans-splicing and editing guides.[a]

Data set[b]	Size of data set (nt)	Number of guides	Number of guides (mean)[c]			
		Topology 1; L=0; d1,d2=0;	Topology 1; L=0..50; d1,d2=0..83;	Topology 1; L=0..5; d1,d2=0..83;	Topology 2; L=0..50; d1,d2=0..83;	Topology 2; L=0..50; d1,d2=0..5;
nuc genome	93,641,047	2,873	N. d.	367,275,143	N. d.	27,321,323
nuc ESTs	2,816,174	135	N. d.	3,916,582	N. d.	197,463
mt genome	633,395	34	15,653,575	N. d.	15,596,645	N. d.
mt cDNAs	68,395	0	37,520	N. d.	35,799	N. d.

[a]For an illustrative description of parameters, see Fig. 3C, D. Minimum match length of paired regions was set to a=6. The number of allowed G:U pairs is ≤3. Guides that have the potential to direct also joining of non-cognate modules have been eliminated. For each data set, the minimum number of junctions covered by guides of a given structural class is six (out of eight); otherwise the number of detected guides is set to zero. Numbers of guides count those with distinct sequence. N. d., not determined.

[b]nuc genome, nuclear genome sequences; mt genome, mitochondrial genome sequences; mt cDNA, mitochondrial cDNA sequences; nuc ESTs, nuclear EST sequences of *D. papillatum*. For details, see Methods Section.

[c]The minimum and maximum values deviate from the mean by ~ 5% (see Supplementary Table S8).

3.2.7 Figure legends

Figure 1. Transcription and transcript processing in *Diplonema* mitochondria. (A) Structure of mitochondrial chromosomes showing regions unique to a given chromosome ('cassette'), which include coding ('gene module') and non-coding sequence ('unique flanking regions'), regions common to a given chromosome class ('class-specific constant regions'), and the portion of the constant regions that is shared by A- and B-class chromosomes ('shared'). (B) The strandedness of the modules relative to the constant region. (C)-(E) Observed types and abundance of putative RNA processing intermediates including a single module (mono-module transcripts). 5', 5'-terminal (first) modules; i, internal modules; 3', 3'-terminal (last) modules of all genes combined. Note that the 5'-terminal module includes a non-coding 5'UTR of ~25 nt. Data are taken from Supplementary Table S1. The total counts of detected 5'-terminal, internal, and 3'terminal mono-modules with defined length of adjacent regions are four, 35, and 38, respectively. In E, the number of polyadenylated 3'modules is overestimated (indicated by <26%, <21%), since one library was enriched in poly-A RNA. The low number of 5'-terminal mono-modules is due to the experimental design. Although 5'-terminal mono-modules carrying both adjacent regions or a 3'-adjacent region have not been detected among the total number of four 5'-terminal mono-modules, this type of intermediate is likely among transcripts whose 3'end remained unknown due to the choice of RT-PCR primers (see Supplementary Table S1, e. g., *cox1*-m1 (dp4030), *cox2*-m1 (dp9347), and *cox3*-m1 (dp9346,56,69). (F) Length distribution of 5' and 3'-adjacent regions of mono-modules. The grey shades represent the percentage of observed intermediates with 5' and/or 3'extensions of length 0 nt, 1-25 nt, 1-100 nt, 1-300 nt, 1-500, etc, up to 1-1,100 nt. Note a steep drop of percentage immediately adjacent to the module boundaries, except for 5'modules that include a 5'UTR of 26-27 nt. The reason why the longest observed 5'extensions of 5'modules is shorter than those of internal plus 3' modules is due to the smaller sample size of the former modules. For details, see Supplementary Table S1.

Figure 2. Transcripts bearing a 5'-triphosphate. (A) Radiogram of total RNA capped by α-[^{32}P]-GTP and guanylyl-transferase, and separated on a denaturating polyacrylamide gel (5%). The band of ~0.12 kb is believed to be cytosolic 5S rRNA that, across eukaryotes, possesses a 5'-triphosphate. The band of ~0.3 kb is most likely mitochondrial 7S RNA based on abundance and size. (B) Ethidium bromide-stained total RNA separated on the same gel as capped RNA. The prominent bands are cytosolic LSU rRNA (~3.5 kb), SSU rRNA (~2 kb), 5.8S rRNA (0.17 kb), and 5S rRNA (0.12 kb) as determined earlier by others (Sturm*, et al.*, 2001).

Figure 3. Design of the experimental and computational search for trans-splicing and RNA-editing guides. m_i, upstream module; m_{i+1}, downstream module. Upper grey bars, module transcripts. Black lines, hypothetical guides. Vertical thin lines, pairing between module transcripts and guide. (A, B) RT-PCR experiments. Arrows, location of primers

for reverse transcriptase (RT) and PCR reactions. Lower grey bar, the resulting amplicon, where light grey shade indicates the sequences that originate from the primers, and dark grey shade indicates the central region that originates from the template. For primers used, see Supplementary Table S9. (A) RT-PCR after RNA circularization to detect the distal regions of antisense RNAs. The short vertical bar indicates the circularization point. (B) RT-PCR to detect the central portion of antisense RNAs that are complementary to module junctions. (C, D) In silico search. L, bridge length; d1, d2, distance of match from 3'-module boundary and 5'-module boundary; a ('anchors'), stretch of 100% sequence complementarity between module and guide. (C) 'Regular' conformation (Topology 1). (D) Permutated conformation (Topology 2).

Figure 4. Potential transcript interactions and maturation pathway. (A) Potential pairing of module precursor transcripts encoded in opposite orientation on a given chromosome class. Transcripts from (+)-orientation chromosomes have the propensity to pair with counterparts from (-)-orientation chromosomes of the same class. Such pairing could form foci of precursors in the organelle, but not align cognate modules. (B)-(D), inferred transcript maturation pathways for 'first' (5') modules (B), internal modules (C), and last' (3') modules (D). Note that 5'modules include a non-coding 5'UTR of ~25 nt. Boxes in black or striped represent modules. Grey boxes are unique flanking regions. Thin bars represent constant regions. Interrupted thin bars indicate that a neighboring module may or may not be present. 5', 5'-terminal module; i, internal module; 3', 3'-terminal module of a given gene. AAAA, poly(A) tail.

Figure 5. Inconsistencies between gene and transcript sequences in 3'-end regions of genes. (A) RNA editing (red letters) and completion of the stop codon by polyadenylation at the 3'end of *cob*. (B), (C), Completion of stop codons by polyadenylation. The genome-encoded 3'-terminal U of the last module from *atp6* and *cox1* is completed to UAA by the A-tail, generating the only stop codon in the *atp6* reading frame and a second stop codon, immediately following UAG, in the *cox1* reading frame. Completion of stop codons by polyadenylation most probably occurs in all protein-coding genes of *Diplonema* mtDNA. (B), lower part: the square bracket with sequences in grey font color indicate artifacts due to the usage of anchored oligo-dT primer for cDNA synthesis; see Supplementary Information. Bold font style and underscoring highlights nucleotides that are identical in genomic and transcriptomic sequence. Lower-case letters in genome sequences show non-coding regions. For adenines (As) in a transcript sequence set in bold but not underlined, it cannot be inferred whether they are encoded or added post-transcriptionally. Genomic, sequence of clones from mtDNA; cDNA/polyAlib, cDNA sequences of clones in cDNA libraries generated by reverse-transcription of poly(A) RNA using an anchored oligo-dT primer; cDNA/RNAcirc, cDNA sequences of clones obtained from circularized RNA that was reverse-transcribed and amplified using gene-specific primers (see Methods). When only a single clone exists for a given sequence, the clone ID is given in parentheses. When multiple clones share the same sequence, the total number of such clones is indicated in

parentheses. The corresponding clones are as follows. *cob*-m6 genomic: dp4155, dp4608, dp4735, dp4941, dp4980, dp4984; *cob*-m6 cDNA/polyAlib: dp0205, dp0314, dp0317, dp1021, dp4278; *atp6*-m4 genomic, dp4241, dp4242, dp4246, dp4887, dp4896; *atp6*-m4 cDNA/RNAcirc: dp9537, dp10201, dp10202, dp10245; *atp6*-m4 cDNA/polyAlib: dp1971, dp0414; *cox1*-m9 genomic, dp4216, dp3328-4, dp3207; *cox1*-m9 cDNA, dp6005 and 83 additional clones.

Figure 6. Experimentally detected RNAs potentially guiding trans-splicing and RNA editing of *cox1*. The cDNA sequence at junctions is shown in blue (upstream module) and green (downstream module). -|, 3' end of upstream module. |-, 5' end of downstream module. dp9241 to dp8689 are RT-PCR clones representing RNAs that are complementary (antisense) to module junctions. The sequence portion in black originates from primers, and the underlined stretch colored red is the sequence originating from the anti-sense RNA. oli88(RT) etc., oligonucleotide primers used for priming the reverse transcriptase reaction. oli80 etc., oligonucleotide primers used in the PCR reaction, together with the RT primer. For primers, see Supplementary Table S9.

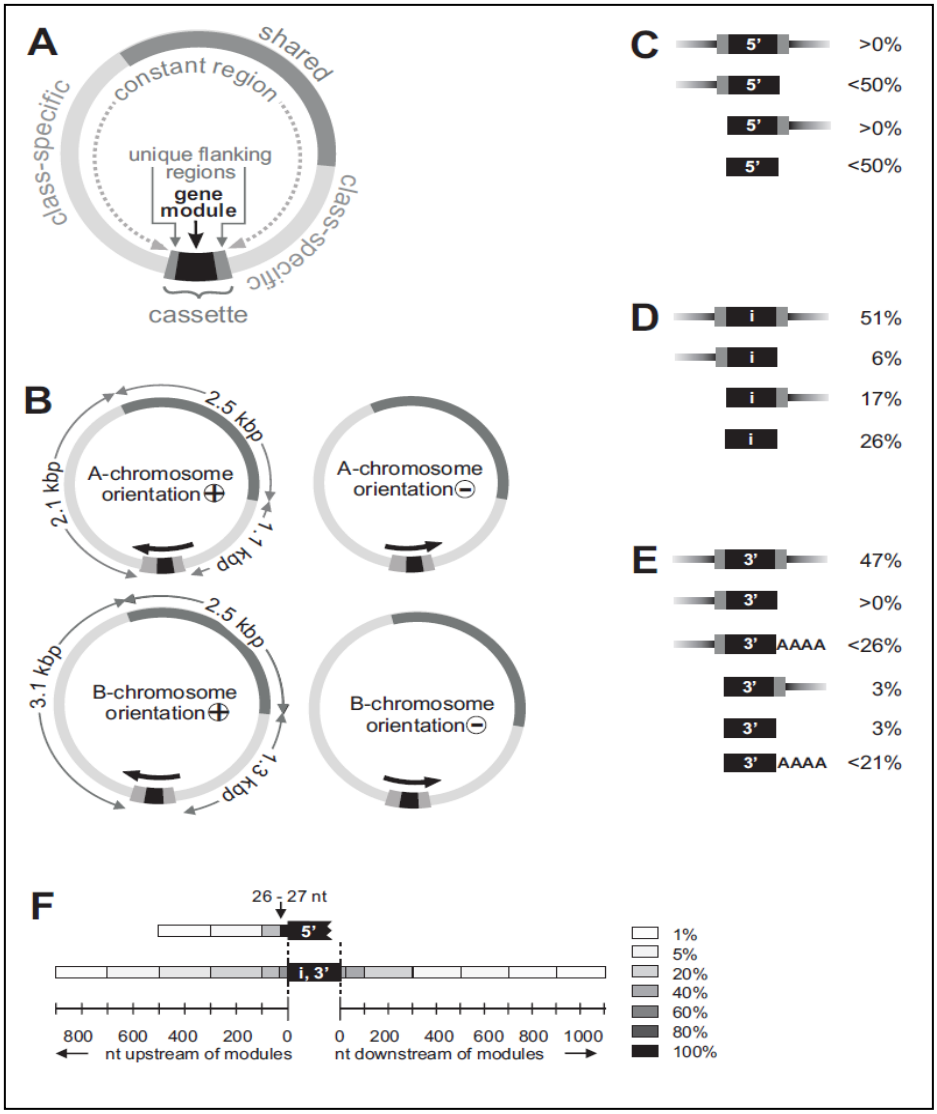


Figure 1. Transcription and transcript processing in *Diplonema* mitochondria.

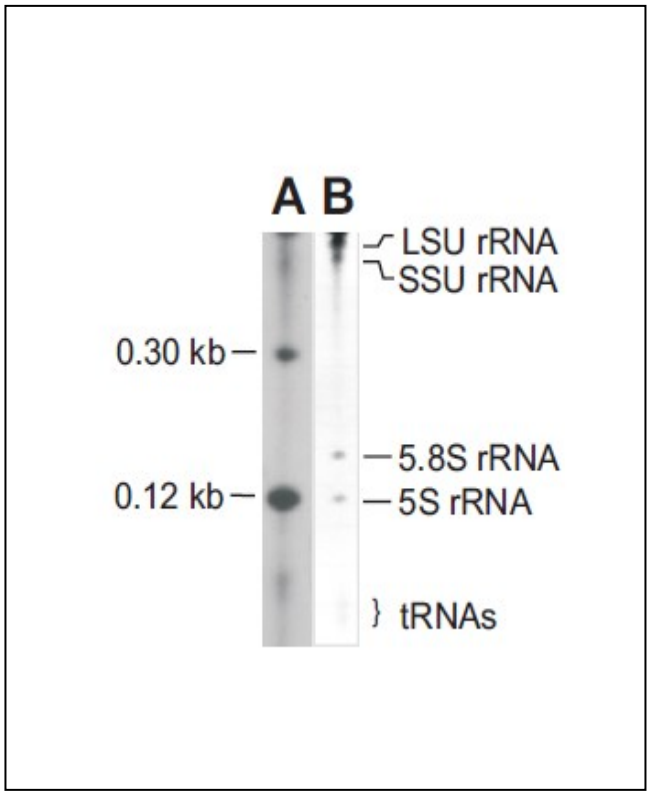


Figure 2. Transcripts bearing a 5'-triphosphate.

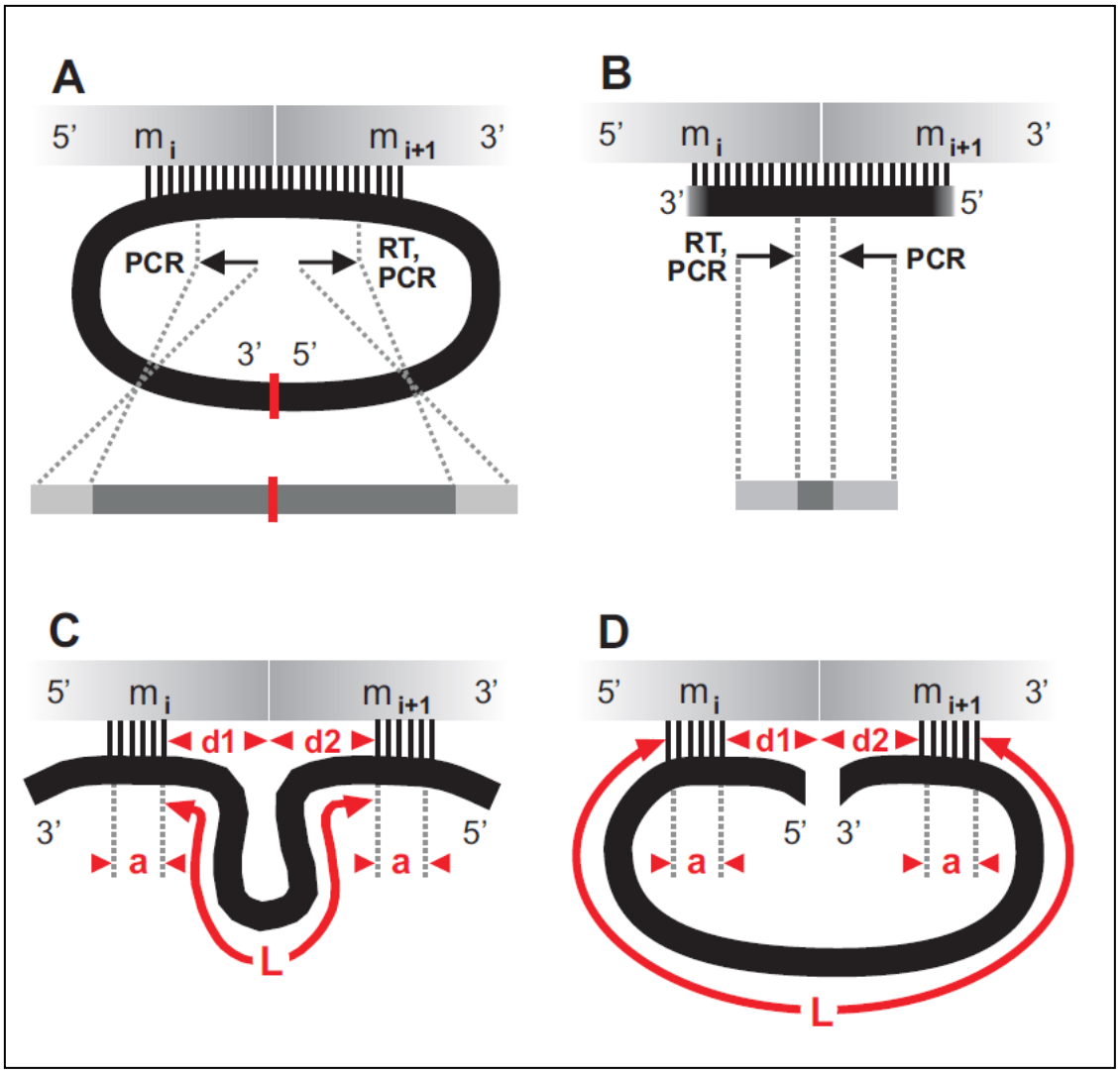


Figure 3. Design of the experimental and computational search for trans-splicing and RNA-editing guides.

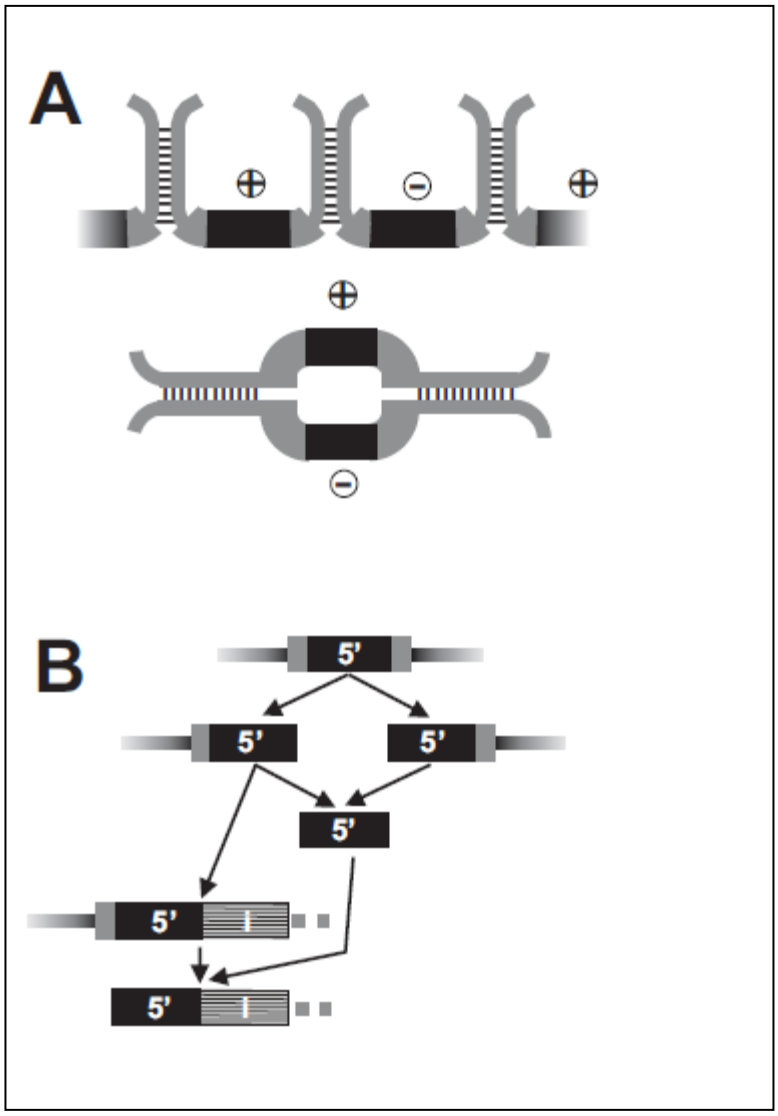


Figure 4 A-B. Potential transcript interactions and maturation pathway.

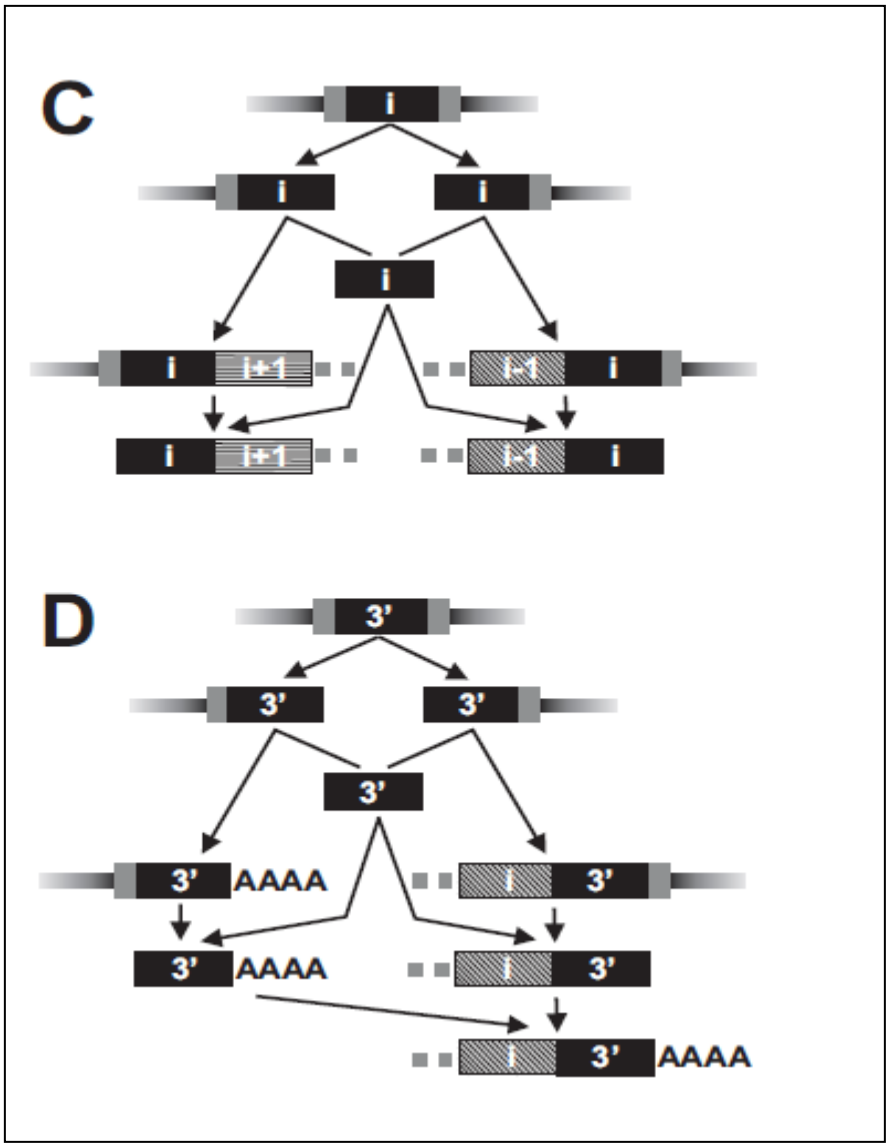


Figure 4 C-D. Potential transcript interactions and maturation pathway.

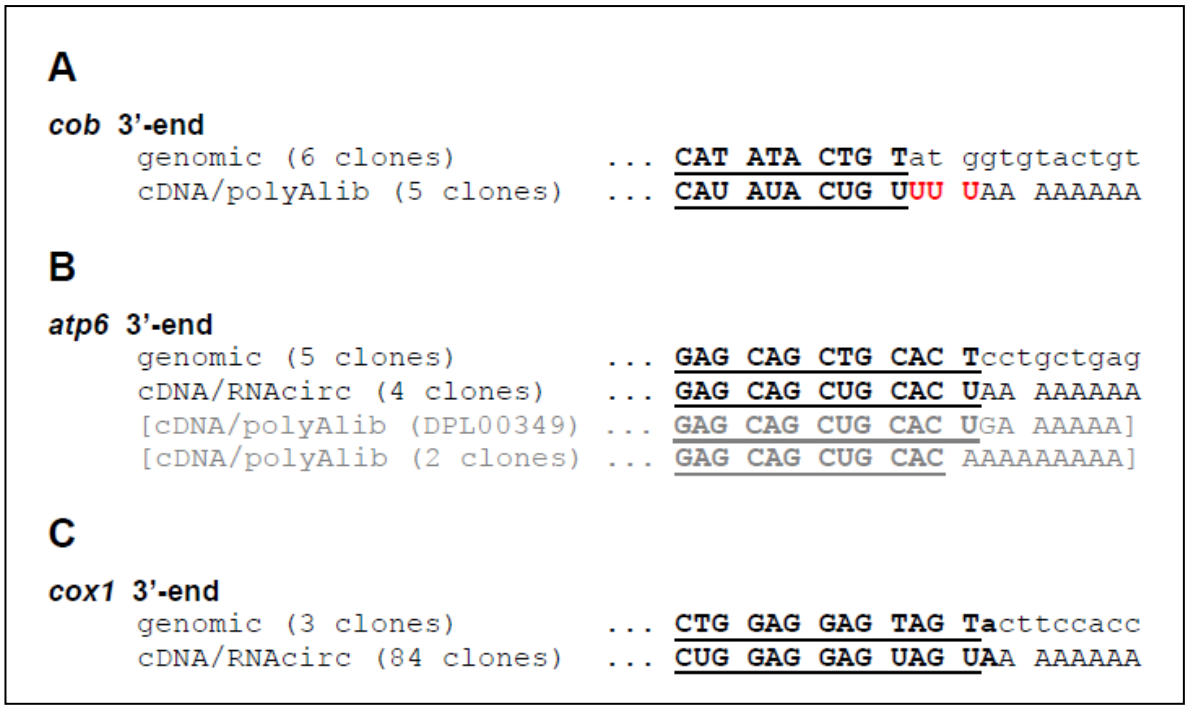


Figure 5. Inconsistencies between gene and transcript sequences in 3'-end regions of genes.

Figure 6. Experimentally detected RNAs potentially guiding trans-splicing and RNA editing of *cox1*.

3.2.8 Acknowledgments

We thank S. Teijeiro and M. Aoulad-Aissa (Université de Montréal) for excellent technical assistance. Further we acknowledge W. Marande (Museum National d'Histoire Naturelle, Paris, France), Sivakumar Kannan (NCBI, Bethesda, USA), and Amir Malekpour (University of Teheran, Iran) for conducting preliminary experiments in the context of their Ph. D. and post-doctoral training under the supervision of GB: RNA circularization and poisoned primer experiments (MW), and in silico searches of RNA and DNA trans-factors (SK, AM). We also thank B. Franz Lang (Université de Montréal) and Julius Lukes (University of South Bohemia, Czech Republic) for advice and discussions, B. Franz Lang and Matus Valach for critical comments on the manuscript, and Matus Valach for help in primer extension experiments. We acknowledge the Institute of Molecular Genetics, Academy of Sciences of the Czech Republic in Prague, in particular C. Vlcek, J. Paces and J. Ridl, for 454 DNA sequencing.

3.2.9 Funding

This work was supported by operating grants from the Canadian Institute for Health Research (CIHR, grant MOP-79309; GB) and the National Science and Engineering Research Council, Canada (NSERC, grant 250909-2006; MT), a graduate student award from the Faculty of graduate and post-doctoral studies (FESP; Université de Montréal; YY), and a Ph. D. scholarship from the Programme Canadien de Bourses de la Francophonie (PCBF scholarship; GNK).

3.2.10 References

Simpson AG and Roger AJ (2004) Protein phylogenies robustly resolve the deep-levelrelationships within Euglenozoa. *Molecular Phylogenetic and Evolution*, **30**, 201-212.

Maslov DA, Yasuhira S and Simpson L (1999) Phylogenetic affinities of *Diplonema* within the Euglenozoa as inferred from the SSU rRNA gene and partial COI protein sequences. *Protist*, **150**, 33-42.

Marande W, Lukeš J and Burger G (2005) Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids. *Eukaryotic Cell*, **4**, 1137-1146.

Marande W and Burger G (2007) Mitochondrial DNA as a genomic jigsaw puzzle. *Science*, **318**, 415.

Kiethega G, Turcotte M and Burger G (2011) Conserved *cox1* trans-splicing and RNA editing lacking conserved sequence patterns. *Molecular Biology and Evolution*, **28**, 2425-2458.

Vlcek C, Marande W, Teijeiro S, Lukeš J and Burger G (2011) Systematically fragmented genes in a multipartite mitochondrial genome. *Nucleic Acids Research*, **39**, 979-988.

Bonen L (1993) Trans-splicing of pre-mRNA in plants, animals, and protists. *FASEB Journal*, **7**, 40-46.

Moreira S, Breton S and Burger G (2012) Unscrambling of genetic information at the RNA level. *Wiley Interdisciplinary Review RNA*.

Lang BF and Burger G (2007) Purification of mitochondrial and plastid DNA. *Nature Protocoles*, **2**, 652-660.

Rodriguez-Ezpeleta N, Teijeiro S, Forget L, Burger G and Lang B.F (2009) In Parkinson J (ed.), *Methods in Molecular Biology: Expressed Sequence Tags (ESTs)*. Humana Press, Totowa, NJ, Vol. 533, pp. 33-47.

Huang Y, Niu B, Gao Y, Fu L and Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, **26**, 680-682.

Ewing B and Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, **8**, 186-194.

Gordon D (2003) Viewing and editing assembled sequences using Consed. *Curr Protoc Bioinformatics*, **Chapter 11**, Unit 11.12.

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

Pearson WR (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods in Molecular Biology*, **132**, 185-219.

Falkenberg M, Larsson NG and Gustafsson CM (2007) DNA replication and transcription in mammalian mitochondria. *Annual Review of Biochemistry*, **76**, 679-699.

Kennell JC. and Lambowitz AM (1989) Development of an in vitro transcription system for *Neurospora crassa* mitochondrial DNA and identification of transcription initiation sites. *Molecular and Cellular Biology*, **9**, 3603-3613.

Sturm NR, Maslov DA, Grisard EC and Campbell DA (2001) *Diplonema* spp. possess spliced leader RNA genes similar to the Kinetoplastida. *Journal of Eukaryotic Microbiology*, **48**, 325-331.

Lee DY and Clayton DA (1998) Initiation of mitochondrial DNA replication by transcription and R-loop processing. *Journal of Biological Chemistry*, **273**,

30614-30621.

Bonawitz ND, Clayton DA and Shadel GS (2006) Initiation and beyond: multiple functions of the human mitochondrial transcription machinery. *Molecular Cell*, **24**, 813-825.

Jackson CJ, Norman JE, Schnare MN, Gray MW, Keeling PJ and Waller RF (2007) Broad genomic and transcriptional analysis reveals a highly derived genome in dinoflagellate mitochondria. *BMC Biology*, **5**, 41.

Gray, MW (2003) Diversity and evolution of mitochondrial RNA editing systems. *IUBMB Life*, **55**, 227-233.

Slamovits CH, Saldarriaga JF, Larocque A and Keeling PJ (2007) The highly reduced and fragmented mitochondrial genome of the early-branching dinoflagellate *Oxyrrhis marina* shares characteristics with both apicomplexan and dinoflagellate mitochondrial genomes. *Journal of Molecular Biology*, **372**, 356-368.

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F. *et al.* (1981) Sequence and organization of the human mitochondrial genome. *Nature*, **290**, 457-465.

Blum B and Simpson L (1990) Guide RNAs in kinetoplastid mitochondria have a nonencoded 3' oligo(U) tail involved in recognition of the preedited region. *Cell*, **62**, 391-397.

Nowacki M, Vijayan V, Zhou Y, Schotanus K, Doak TG. and Landweber LF (2008) RNA-mediated epigenetic programming of a genome-rearrangement pathway. *Nature*, **451**, 153-158.

Koulintchenko M, Konstantinov Y and Dietrich A (2003) Plant mitochondria actively

import DNA via the permeability transition pore complex. *EMBO Journal*, **22**, 1245-1254.

Salinas T, Duchene AM. and Marechal-Drouard L (2008) Recent advances in tRNA mitochondrial import. *Trends in Biochemical Sciences*, **33**, 320-329.

Keiler KC, Shapiro L and Williams KP (2000) tmRNAs that encode proteolysis-inducing tags are found in all known bacterial genomes: A two-piece tmRNA functions in *Caulobacter*. *Proceedings National Academy of Sciences of Unites States of America*, **97**, 7778-7783.

Soma A, Onodera A, Sugahara J, Kanai A, Yachie N, Tomita M, Kawamura F and Sekine Y (2007) Permuted tRNA genes expressed via a circular RNA intermediate in *Cyanidioschyzon merolae*. *Science*, **318**, 450-453.

Stuart KD, Schnaufer A, Ernst NL. and Panigrahi AK. (2005) Complex management: RNA editing in trypanosomes. *Trends in Biochemical Sciences*, **30**, 97-105.

Byrne EM, Connell GJ. and Simpson L (1996) Guide RNA-directed uridine insertion RNA editing in vitro. *EMBO Journal*, **15**, 6758-6765.

Brecht M, Niemann M, Schluter E, Muller UF, Stuart K and Goringer HU (2005) TbMP42, a protein component of the RNA editing complex in African trypanosomes, has endo-exoribonuclease activity. *Molecular Cell*, **17**, 621-630.

Kang X, Rogers K, Gao G, Falick AM, Zhou S and Simpson L (2005) Reconstitution of uridine-deletion precleaved RNA editing with two recombinant enzymes.
Proceedings National Academy of Sciences of Unites States of America, **102**, 1017-1022.

Niemann M, Kaibel H, Schluter E, Weitzel K, Brecht M and Goringer, HU. (2009) Kinetoplastid RNA editing involves a 3' nucleotidyl phosphatase activity. *Nucleic Acids Research*, **37**, 1897-1906.

3.2.11 Supplementary data

Table of content

3.2.12 Supplementary results

Poisoned primer extension

The aim of this experiment was to test for the existence of a transcript carrying *cox1* Modules 4 and 5 without six Us in between (abbreviated 'm4-m5'), which would imply that RNA editing takes place after trans-splicing. In the poisoned primer extension assay, RNA is reverse transcribed starting from a radiolabeled primer, in the presence of

the four nucleotides whereof one is a chain-terminating ddNTP (in one reaction, we used ddTTP, and in the other ddCTP). Consequently, the chain elongation will stop upon incorporation of the ddNTP opposite to the first occurrence of the complementary nucleotide in the template, notably G or A, respectively. The size of the reaction product, determined by electrophoretic separation on a polyacrylamide gel, thus pinpoints the first occurrence of a given nucleotide upstream of the priming site. The primer used (dp109) anneals with positions +2 to +30 in the 5'terminal region of *cox1* Module 5. The band sizes observed correspond to products templated by (i) Module-4_UUUUUU_Module-5 and (ii) Module 5 with a unprocessed 5'end, but are not compatible with the m4-m5 scenario. This provides independent corroboration to the experiments described in the main text, that *cox1* RNA editing in *Diplonema* mitochondria proceeds not as in kinetoplastids by nucleotide insertion, but rather by terminal nucleotide addition (attachment scenario, see main text).

Sequence polymorphism in mitochondrial transcripts?

Considerable sequence variation was noted among mitochondrial cDNA clones of *Diplonema*, in the region directly upstream of the poly(A) tail, but not among corresponding genomic clones (for example *atp6*; Fig. 5B, main text, sequences in brackets). Sequence polymorphism adjacent to the poly(A) tail may arise artefactually from first-strand synthesis during EST library construction, where an 'anchored' oligo-dT-primer (5'-30xT[A or C or G]-3') was used to assure that annealing occurs immediately downstream of the coding region and not elsewhere in an mRNA's A-tail. In fact, critical inspection of sequence reads shows that in all cDNA-library clones including the 3' end of *atp6*, the A-tract originates from the oligo-dT-primer sequence, suggesting that the apparent sequence polymorphism is an artefact. For additional experimental confirmation, we performed for *atp6* an RT-PCR experiment on circularized poly(A) RNA using divergent primers annealing within coding regions, notably one in the last module and the other in a more upstream module (see Methods). The resulting cDNA sequences concurred with that of genomic clones. This confirms

that the seeming sequence polymorphism in the initial cDNA clones is a spurious result. Therefore, caution must be employed when interpreting cDNA 3'-end data derived from libraries constructed with an anchored oligo-dT-primer.

3.2.13 Supplementary material and methods

Poisoned primer extension

The assay followed essentially the original procedure (Driscoll*, et al.*, 1989). Briefly, enriched mitochondrial RNA was reverse transcribed with AMV reverse transcriptase using as primer the oligo- nucleotide dp109 (see Supplementary Table S9) that was labeled at its 5' end with γ-[^{32}P]-ATP and T4 polynucleotide kinase. The primer anneals with transcripts that include the *cox1* Module 5, and its sequence is complementary to the 5' region of the module. The reaction was incubated in the presence of dATP, ddCTP, dGTP, and dTTP. An additional experiment was performed using dATP, dCTP, dGTP, and ddTTP. Strand-termination occurs after incorporation of the first dideoxy nucleotide. For ddCTP, the 5'-end unprocessed Module 5 transcript, the hypothetical Module 4-Module 5 transcript, and the edited Module 4-6xU-Module 5 transcript would yield products that are 9 nt, 9 nt, and 3 nt, respectively, longer than the primer. For ddTTP, these three transcripts would yield products that are 3 nt, 2 nt, and 8 nt, respectively, longer than the primer. The reaction products were separated on a 10% denaturizing polyacrylamide gel. Labeled oligonucleotides, untreated and hydrolyzed by NaOH, served as size markers.

Computational search for guiding RNAs or DNAs

A guiding element is defined as an RNA or DNA molecule that directs module trans-splicing, as well as RNA editing. We postulate that there is at least one guiding element for each module junction. The element's sequence comprises two fixed-length segments, each interacting with one of two adjacent modules. The two interacting regions are located at distance **d1** from the end of the upstream module and distance **d2**

from the start of the downstream module. As Figures 3C and D (main text) illustrate, guiding elements can have Topology 1 (e.g. like kinetoplastid gRNAs or microRNAs) or Topology 2 (permutation of Topology 1). The two segments are not necessarily adjacent to one another, but may be separated by a 'bridge', whose length is denoted **L**. Furthermore, we assume that functional guiding elements have a common structure, i.e., that they all bind at the same distance from the junction and will all have the same bridge length and topology.

The algorithm that identifies potential guiding elements for Topology 1 is outlined below.

1. For each **L**=0..50, **d1**=0..83 and **d2**=0..83

1.1 Let **n**=0 be the number of junctions covered

1.2 For junction **j**=1..8

1.2.1 Let **s1** be the segment of length 6 at distance **d1** from the end of module **j**

1.2.2 Let **s2** be the segment of length 6 at distance **d2** from the start of module **j+1**

1.2.3 Let **e** be a regular expression consisting of **s2' -** a bridge of length **L - s1',** where **s2'** and **s1'** are the reverse complement of **s2** and **s1**, respectively

1.2.4 Find all the occurrences of **e** in the given set of sequences

1.2.5 Filter the occurrences for number of G:U pairs, miss-assembly, and uniqueness

1.2.6 If the number of remaining occurrences is greater than 0, increment **n**

1.3 If **n** is greater than or equals to 6, save the results.

For Topology 2, the expression constructed at line 1.2.3 is **s1' -** a bridge of length **L - s2'**. In the case of genomic sequences, the complementary strands are searched as well (line 1.2.4). Up to three G:U pairs are allowed per duplex. The above algorithm is implemented in Java using ExecutorService for parallelism.

3.2.14 Supplementary tables

Supplementary Table S1. Observed mitochondrial transcripts containing gene modules.[a]

Gene module contained in transcript (module size)	Chromosome type, orientation (length of unique 5' & 3'-flanking regions)	Length of transcript's region upstream (5') of module	Length of transcript's region downstream (3') of module	Experiment[b]	Clone ID	Transcript category
atp6-m2	A-	0 nt	/	circRT-PCR (dpcirc)	dp9571	oligo-module: m2..m3(partial)
(195 nt)	(~4 nt & 36 nt)			circRT-PCR (dpcirc)	dp10245	oligo-module: m2..m4+A-tail
				circRT-PCR (dpcirc)	dp10226	oligo-module: m2..m4(partial)
		1 nt	/	circRT-PCR (dpcirc)	dp10237	oligo-module: m2..m3(partial)
atp6-m3	A+	73 nt	/	cDNA library (dpape_s)	dp0414	oligo-module: m3..m4+A-tail
(271 nt)	(10 nt & 18 nt)	118 nt	/	cDNA library (dpape_s)	dp1977	oligo-module: m3..m4+A-tail
atp6-m4*	A-	/	0 nt+A-tail	cDNA library (dpape_s)	dp1977	oligo-module: m3..m4+A-tail
(139 nt)	(50 nt & 62 nt)	/	0 nt+A-tail	circRT-PCR (dpcirc)	dp9537	number of modules unknown
				circRT-PCR (dpcirc)	dp10201	oligo-module:m3(partial)..m4+A-tail
				circRT-PCR (dpcirc)	dp10202	oligo-module: m2(partial)..m4+A-tail
				circRT-PCR (dpcirc)	dp10245	oligo-module: m2..m4+A-tail
		/	0 nt+A-tail	cDNA library (dpape_s)	dp0414	oligo-module: m3..m4+A-tail
		/	0 nt+A-tail	cDNA library (dpape_s)	DPL0349	number of modules unknown
cob-m1	A+	26 nt	0 nt	circRT-PCR (dpmod9)	dp5996	mono-module
(198 nt)	(24 nt & 54 nt)	27 nt	/	cDNA library (dpape_s)	dp0315	complete transcript
				cDNA library (dpape_s)	dp4278	complete transcript
				cDNA library (dpape_s)	dp0811	complete transcript
cob-m4	A-	0 nt	/	cDNA library (dpape_s)	dp0314	oligo-module: m4..m6+A-tail
(198 nt)	(42 nt & 54 nt)					
cob-m5	A-	0 nt	/	cDNA library (dpape_s)	**dp0317**	oligo-module: m5..m6+A-tail
	(15 nt & 22 nt)					
cob-m6*	B-	/	0 nt+A-tail	cDNA library (dpape_s)	dp0205w	number of modules unknown
	(39 nt & 44 nt)			cDNA library (dpape_s)	dp1021	number of modules unknown
				cDNA library (dpape_s)	dp4278	complete transcript
				circRT-PCR (dpcirc)	dp10666	mono-module
cox1-m1	B+	>25 nt	/	RT-PCR 5'-race (dpape_s)	dp4023	number of modules unknown
(195 nt)	(52 nt & 35 nt)				dp4024	
				RT-PCR 5'-race (dpape_s)		
		26 nt	/	circRT-PCR (dpmod59)	dp6210	complete transcript
				circRT-PCR (dpmod59)	dp6230	complete transcript
				circRT-PCR (dpmod59)	dp6290	complete transcript
		>26 nt	/	RT-PCR 5'-race (dpape_s)	dp4022	
		27 nt	/	circRT-PCR (dpmod59)	dp6228	complete transcript
				circRT-PCR (dpmod59)	dp6231	complete transcript

				circRT-PCR (dpmod59)	dp6259	complete transcript
				circRT-PCR (dpmod59)	dp6283	complete transcript
		27 nt	/	circRT-PCR (dpmod9)	dp5954	complete transcript
				circRT-PCR (dpmod9)	dp5996	complete transcript
			/	circRT-PCR (dpmod29)	dp5692	complete transcript
		28 nt	/	circRT-PCR (dpmod59)	dp6279	complete transcript
		29 nt	/	circRT-PCR (dpmod59)	dp6207	complete transcript
		31 nt	0 nt	circRT-PCR (dpmod9)	dp5929	mono-module
		>383 nt	/	RT-PCR 5'-race (dpape_s)	dp4030	number of modules unknown
cox1-m2	A+	0 nt	6 nt	circRT-PCR (dpmod2)	dp5555	mono-module
(124 nt)	(35 nt & 63 nt)	0 nt	30 nt	circRT-PCR (dpmod2)	dp5557	mono-module
		21 nt	2 nt	circRT-PCR (dpmod2)	dp5589	mono-module
		22 nt	31 nt	circRT-PCR (dpmod2)	dp5553	mono-module
		30 nt	40 nt	circRT-PCR (dpmod2)	dp5554	mono-module
		40 nt	38 nt	circRT-PCR (dpmod2)	dp5585	mono-module
		41 nt	42 nt	circRT-PCR (dpmod2)	dp5594	mono-module
		47 nt	88 nt	circRT-PCR (dpmod2)	dp5558	mono-module
		49 nt	54 nt	circRT-PCR (dpmod2)	dp5551	mono-module
		57 nt	32 nt	circRT-PCR (dpmod2)	dp5560	mono-module
		62 nt	47 nt	circRT-PCR (dpmod2)	dp5587	mono-module
		63 nt	41 nt	circRT-PCR (dpmod2)	dp5549	mono-module
		67 nt	42 nt	circRT-PCR (dpmod2)	dp5592	mono-module
		73 nt	0 nt	circRT-PCR (dpmod2)	dp5595	mono-module
		74 nt	44 nt	circRT-PCR (dpmod2)	dp5593	mono-module
		127 nt	/	circRT-PCR (dpmod59)	dp6275	oligo-module: m2.. ≥m9
		>136 nt	/	RT-PCR (dpape_s)	dp4027	number of modules unknown
				RT-PCR (dpape_s)	dp4025	number of modules unknown
		>140 nt	/	RT-PCR (dpape_s)	dp4026	number of modules unknown
		>146 nt	/	RT-PCR (dpapg)	dp9355	number of modules unknown
				RT-PCR (dpapg)	dp9366	number of modules unknown
		149 nt	/	circRT-PCR (dpmod59)	dp6287	oligo-module: m2..m9+A-tail
		384 nt	/	circRT-PCR (dpmod59)	dp6281	oligo-module: m2..m9+A-tail
		470 nt	56 nt	circRT-PCR (dpmod2)	dp5556	mono-module
		>631 nt	/	RT-PCR (dpape_s)	dp4031	number of modules unknown
				RT-PCR (dpape_s)	dp4032	number of modules unknown
				RT-PCR (dpape_s)	dp4033	number of modules unknown
		943 nt	/	RT-PCR (dpape_s) 5'race	dp4028	number of modules unknown
cox1-m3	A+	0 nt	/	circRT-PCR (dpmod45)	dp7008	oligo-module: m3..m5
(263 nt)	(26 nt & 32 nt)			circRT-PCR (dpmod45)	**dp7011**	oligo-module: m3..m6
				RT-PCR (dpapg)	dp8025	oligo-module: m3..m9(partial)
		6 nt	/	circRT-PCR (dpmod59)	dp6282	oligo-module: m3..m9+A-tail
		7 nt	/	circRT-PCR (dpmod59)	dp6277	oligo-module: m3..m9+A-tail
cox1-m4	B+	0 nt	121 nt	circRT-PCR (dpmod45)	dp7357	mono-module
(226 nt)	(15 nt & 47 nt)	0 nt	146 nt	circRT-PCR (dpmod45)	dp7304	mono-module

		0 nt	151 nt	circRT-PCR (dpmod45)	dp7311	mono-module
		47 nt	100 nt	circRT-PCR (dpmod45)	dp7073	mono-module
		/	>344 nt	circRT-PCR (dpmod45)	**dp7337**	number of modules unknown
				circRT-PCR (dpmod45)	dp7346	number of modules unknown
cox1-m5	A+	0 nt	0 nt	circRT-PCR (dpmod45)	dp7341	mono-module
(179 nt)	(63 nt & 24 nt)			circRT-PCR (dpmod45)	dp7382	mono-module
				circRT-PCR (dpmod59)	dp6256	mono-module
				circRT-PCR (dpmod5)	dp6101	mono-module
				circRT-PCR (dpmod5)	& 11 clones	mono-module
		0 nt	/	cDNA library (dpape_s)	dp0585	oligo-module: m5.. ≥m8
		35 nt	13 nt	circRT-PCR (dpmod45)	dp7005	mono-module
		65 nt	0 nt	circRT-PCR (dpmod45)	dp7373	mono-module
		276 nt	/	cDNA library (dpape_s)	**dp0110**	oligo-module: m5..≥m7
		/	0 nt	circRT-PCR (dpmod45)	dp7008	oligo-module: m3..m5
cox1-m6	A+	0 nt	/	cDNA library (dpape_s)	dp0333	oligo-module: m6..m9+A-tail
(169 nt)	(35 nt & 80 nt)			cDNA library (dpape_s)	dp0464	number of modules unknown
				cDNA library (dpape_s)	**dp0655**	oligo-module: m6..m9+A-tail
				cDNA library (dpape_s)	dp0748	oligo-module: m6..m9+A-tail
				cDNA library (dpape_s)	dp1026	oligo-module: m6..m9+A-tail
		/	0 nt	circRT-PCR (dpmod45)	**dp7011**	oligo-module: m3..m6
		>316 nt	/	RT-PCR (dpapg)	dp9365	number of modules unknown
cox1-m8	A+	>4 nt	/	RT-PCR (dpapg)	dp8025	oligo-module: m8..m9(partial)
(111 nt)	(102 nt & 22 nt)			RT-PCR (dpapg)	dp8026	oligo-module: m8..m9(partial)
				RT-PCR (dpapg)	dp8054	oligo-module: m8..m9(partial)
		>4 nt	0 nt	RT-PCR (dpapg)	dp8058	mono-module
		/	2 nt	RT-PCR (dpapg)	dp8029	number of modules unknown
cox1-m9*	A-	0 nt	0 nt+A-tail	circRT-PCR (dpmod9)	**dp5927**	mono-module
(251 nt)	(9 nt & 51 nt)			circRT-PCR (dpmod9)	& 5 clones	
		12 nt	17 nt	circRT-PCR (dpmod9)	**dp5939**	mono-module
		14 nt	54 nt	circRT-PCR (dpmod9)	dp5538	mono-module
		19 nt	16 nt	circRT-PCR (dpmod9)	dp5539	mono-module
		35 nt	59 nt	circRT-PCR (dpmod9)	dp5521	mono-module
		37 nt	48 nt	circRT-PCR (dpmod9)	dp5514	mono-module
		43 nt	33 nt	circRT-PCR (dpmod9)	dp5519	mono-module
		56 nt	80 nt	circRT-PCR (dpmod9)	dp5524	mono-module
		113 nt	2 nt	circRT-PCR (dpmod9)	dp5518	mono-module
		185 nt	0 nt+A-tail	cDNA library (dpape_s)	dp0402	mono-module
		222 nt	0 nt+A-tail	cDNA library (dpape_s)	**dp0285**	mono-module
		282 nt	9 nt	circRT-PCR (dpmod9)	dp5922	mono-module
		382 nt	11 nt	circRT-PCR (dpmod9)	dp5542	mono-module
		637 nt	85 nt	circRT-PCR (dpmod9)	dp5934	mono-module
		643 nt	391 nt	circRT-PCR (dpmod9)	dp5956	mono-module
		651 nt	463 nt	circRT-PCR (dpmod9)	dp6043	mono-module

		719 nt	325 nt	circRT-PCR (dpmod9)	dp6034	mono-module
		/	0 nt	circRT-PCR (dpmod9)	**dp5977**	oligo-module: m6(partial)..m9
		/	80 nt	circRT-PCR (dpmod59)	dp5622	number of modules unknown
		/	709 nt	circRT-PCR (dpmod9)	dp6006	number of modules unknown
		/	0 nt+A-tail	cDNA library (dpape_s)	**dp0655**	oligo-module: m6..m9
cox2-m1	A+	29 nt	/	cDNA library (dpape_s)	dp1341	complete transcript
				cDNA library (dpape_s)	dp1342	complete transcript
(237 nt)	(41 nt & 28±2 nt)	40 nt	/	cDNA library (dpape_s)	dp2760	complete transcript
		>153 nt	/	RT-PCR (dpapg)	dp9347	number of modules unknown
		243 nt	/	cDNA library (dpape_s)	dp0321	complete transcript
cox2-m3	A+	0 nt	/	cDNA library (dpape_s)	dp0932	oligo-module: m3..m4+A-tail
(76 nt)	(57 nt & 155 nt)			cDNA library (dpape_s)	dp1158	oligo-module: m3..m4+A-tail
		>316 nt	/	RT-PCR (dpapg)	dp9337	number of modules unknown
				RT-PCR (dpapg)	dp9364	number of modules unknown
				RT-PCR (dpapg)	dp9368	number of modules unknown
cox2-m4*	A-	0 nt	0 nt+A-tail	cDNA library (dpape_s)	dp0107	mono-module
(125 nt)	(26 nt & 133 nt)	/	0 nt+A-tail	cDNA library (dpape_s)	dp0510	number of modules unknown
				cDNA library (dpape_s)	dp0519	complete transcript
				cDNA library (dpape_s)	dp1341	complete transcript
				cDNA library (dpape_s)	dp1342	complete transcript
				cDNA library (dpape_s)	dp2760	complete transcript
cox3-m1	A+	22 nt	/	cDNA library (dpape_s)	dp0371	complete transcript
		25 nt	/	cDNA library (dpape_s)	dp1050	complete transcript
(344 nt)	(4 nt & 9 nt)	26 nt	0 nt	circRT-PCR (dpmod9)	dp5993	mono-module
		>92 nt	/	RT-PCR (dpapg)	dp9346	number of modules unknown
		>108 nt	/	RT-PCR (dpapg)	dp9369	number of modules unknown
				RT-PCR (dpapg)	dp9356	number of modules unknown
		116 nt	0 nt	circRT-PCR (dpmod9)	dp5932	mono-module
		142 nt	/	circRT-PCR (dpmod9)	dp5907	(low sequence quality from nt.270 on)
		142 nt	/	cDNA library (dpape_s)	dp0301	complete transcript
				cDNA library (dpape_s)	dp0305	complete transcript
cox3-m2	A+	0 nt	/	cDNA library (dpape_s)	dp0660	oligo-module: m2..m3+A-tail
(266 nt)	(55 nt & 12 nt)			cDNA library (dpape_s)	dp1020	oligo-module: m2..m3+A-tail
cox3-m3*	A-	0 nt	0 nt+A-tail	cDNA library (dpape_s)	dp0453	mono-module
(230 nt)	(17 nt & 57 nt)					
nad1-m1	A+	5-6 nt	/	circRT-PCR (dpmod29)	dp5640	oligo-module: m1..m2(partial)
(214 nt)	(26 nt & 38 nt)			circRT-PCR (dpmod29)	dp5685	oligo-module: m1..m2(partial)
nad1-m4	A-	0 nt	0 nt	circRT-PCR (dpmod45)	dp7303	mono-module
(228 nt)	(19 nt & 65 nt)	0 nt	/	cDNA library (dpape_s)	dp1251	oligo-module: m4..m5+A-tail
		>20 nt	115 nt	circRT-PCR (dpcirc)	dp9818	mono-module
nad1-m5*	A-	183 nt	0 nt+A-tail	cDNA library (dpape_s)	dp3250	mono-module
(204 nt)	(24 nt & 52 nt)	667 nt	0 nt+A-tail	cDNA library (dpape_s)	dp0975	mono-module
nad4-m4	A+	0 nt	0 nt	circRT-PCR (dpmod5)	dp6108	mono-module
(168 nt)	(55 nt & 33 nt)	0 nt	/	cDNA library (dpape_s)	dp0031	oligo-module: m4.. ≥m5
nad4-m6	A-	0 nt	/	cDNA library (dpape_s)	dp4754	oligo-module: m6..m8+A-tail

(165 nt)	(25 nt & 43 nt)			cDNA library (dpape_s)	dp0716	oligo-module: m6..m8+A-tail
				cDNA library (dpape_s)	dp0115	oligo-module: m6..m8+A-tail
				cDNA library (dpape_s)	dp0840	oligo-module: m6..m8+A-tail
nad4-m7	B+	0 nt	0 nt	circRT-PCR (dpmod45)	dp7308	mono-module
(120 nt)	(31 nt & 127 nt)	/	969 nt	circRT-PCR (dpcirc)	dp9410	number of modules unknown
		/	>693 nt	circRT-PCR (dpcirc)	dp9436	number of modules unknown
		/	**1143 nt**	circRT-PCR (dpcirc)	dp9661	number of modules unknown
nad4-m8*	A-	147 nt	0 nt+A-tail	cDNA library (dpape_s)	dp3212	mono-module
(173 nt)	(13 nt & 73 nt)	213 nt	0 nt+A-tail	cDNA library (dpape_s)	dp0726	mono-module
		449 nt	0 nt+A-tail	cDNA library (dpape_s)	dp1114	mono-module
nad5-m2	B+	~0 nt	/	circRT-PCR (dpcirc)	dp10288	oligo-module: m2..m12+A-tail
(318 nt)	(15 nt & 41 nt)					
nad5-m6	A-	0 nt	0 nt	circRT-PCR (dpcirc)	dp9894	mono-module
(136 nt)	(18 nt & 157 nt)	0 nt	/	cDNA library (dpape_s)	dp0104	oligo-module: m6..≥m11
				cDNA library (dpape_s)	dp0211	oligo-module: m6..≥m11
				cDNA library (dpape_s)	dp0863	oligo-module: m6..≥m11
				cDNA library (dpape_s)	dp0944	oligo-module: m6..≥m11
nad5-m9	A+	~0 nt	240 nt	cDNA (dpape_s)	dp0745	mono-module; 1 nt missing at 5'
(90 nt)	(116 nt & 3 nt)	>224 nt	>3 nt	RT-PCR (dpapg)	dp9351	mono-module
		/	0 nt	circRT-PCR (dpcirc)	dp9985	mono-module; 4 nt missing at 5'
nad5-m10	A-	/	0 nt	circRT-PCR (dpcirc)	dp9985	number of modules unknown
(192 nt)	(32 nt & 53 nt)					
nad5-m11	A-	0 nt	/	cDNA library (dpape_s)	dp0346	oligo-module: m11..m12+A-tail
(113 nt)	(68 nt & 92 nt)			cDNA library (dpape_s)	dp0450	oligo-module: m11..m12+A-tail
				cDNA library (dpape_s)	dp4569	oligo-module: m11..m12+A-tail
		712 nt	/	cDNA library (dpape_s)	dp0975	number of modules unknown
nad5-	A-	/	0 nt+A-tail	cDNA library (dpape_s)	dp2493	oligo-module: m11..m12+A-tail
m12*				cDNA library (dpape_s)	dp2904	oligo-module: m11..m12+A-tail
				cDNA library (dpape_s)	dp3036	oligo-module: m11..m12+A-tail
(76 nt)	(68 nt & 92 nt)	0 nt	374 nt	circRT-PCR (dpcirc)	dp10256	mono-module
		383 nt	~0 nt+A-tail	cDNA library (dpape_s)	dp4430	mono-module
		/	386 nt	circRT-PCR (dpcirc)	dp10016	number of modules unknown
nad7-m1	A+	44 nt	/	cDNA library (dpape_s)	dp4285	complete transcript
(221 nt)	(40 nt & 35 nt)	211 nt	/	circRT-PCR (dpmod9)	dp5936	oligo-module: m1..≥m5
nad7-m5	A+	0 nt	/	circRT-PCR (dpcirc)	dp9873	oligo-module: m5..m8
(192 nt)	(81 nt & 11-12 nt)			circRT-PCR (dpmod9)	dp5548	oligo-module: m5..m7(partial)
		222 nt	/	cDNA library (dpape_s)	dp4280	oligo-module: m5..m6(partial)
		223 nt	/	cDNA library (dpape_s)	dp4282	oligo-module: m5..m9+A-tail
nad7-m6	A-	0 nt	/	circRT-PCR (dpcirc)	dp9728	oligo-module: m6..m7(partial)
(102 nt)	(36 nt & 194 nt)			circRT-PCR (dpcirc)	dp9706	oligo-module: m6..m7(partial)
nad7-m7	B+	5 nt	/	circRT-PCR (dpmod9)	dp5927	oligo-module: m7..m8(partial)
(169 nt)	(31 nt & 53 nt)	/	0 nt	circRT-PCR (dpcirc)	dp9818	number of modules unknown
		/	403 nt	circRT-PCR (dpcirc)	dp9947	number of modules unknown
nad7-m8	A-	/	0 nt	circRT-PCR (dpcirc)	dp9873	oligo-module: m5..m8
	(34 nt & 3 nt)					

nad7-m9*	A-	/	0 nt+A-tail	cDNA library (dpape_s)	dp0669	number of modules unknown
(79 nt)	(66 nt & 128 nt)			cDNA library (dpape_s)	dp3636	number of modules unknown
rnl-m(k)*	B-	0 nt	0 nt	circRT-PCR (dpcirc)	dp10439r	mono-module
(352 nt)	(36 nt & 46 nt)			circRT-PCR (dpcirc)	dp10526	mono-module
		0 nt	0 nt+A-tail	cDNA library (dpape_s)	≥50	mono-module (e.g. dp0268)
				circRT-PCR (dpcirc)	clones	mono-module (e.g. dp10439)
					14 clones	
		>76 nt	>45 nt	circRT-PCR (dpcirc)	dp10586	mono-module
		108 nt	88 nt	circRT-PCR (dpcirc)	dp10574	mono-module
		516 nt	>31 nt	circRT-PCR (dpcirc)	dp9408	mono-module
X1-m(k-1)	A+	0 nt	/	cDNA library (dpape_s)	dp1172	oligo-module: m(k-1)..m(k)+A-tail
(258 nt)	(176 nt & 1 nt)	>34 nt	/	RT-PCR (dpapg)	dp9354	number of modules unknown
X1-m(k)*	A+	272 nt	/	cDNA library (dpape_s)	dp3976	number of modules unknown
(101 nt)	(40 nt & 3 nt)	/	0 nt+A-tail	cDNA library (dpape_s)	dp1115	oligo-module: m(k-1)(part)..m(k)+A-
				cDNA library (dpape_s)	dp1172	tail
						oligo-module: m(k-1)..m(k)+A-tail
X2-m(k)*	A-	176 nt	0 nt+A-tail	cDNA library (dpape_s)	dp2611	mono-module
(130 nt)	(48 nt & 41 nt)	429 nt	35 nt	circRT-PCR (dpmod9)	dp6005	mono-module
		613 nt	/	cDNA library (dpape_s)	dp2179	mono-module
		/	0 nt+A-tail	cDNA library (dpape_s)	dp1778	number of modules unknown
				cDNA library (dpape_s)	dp2105	number of modules unknown
		/	763 nt	circRT-PCR (dpcirc)	dp10251	number of modules unknown
X3-m(k)*	A-	53 nt	12 nt	circRT-PCR (dpmod9)	dp5988	mono-module
(239 nt)	(~0 nt & ~31 nt)	80 nt	0 nt+A-tail	cDNA library (dpape_s)	dp0245	mono-module

[a]For 5'-terminal modules (m1), all observed module transcripts are listed, i.e., also those in mature RNAs, in order to document the 5' UTR. For internal and last modules, only those transcripts are listed that have at least one unjoined end (i.e., excluding mature RNAs). Tilt, approximate value. Clone identifiers and region lengths in bold highlight those mentioned in the main text. Definition of a module's orientation in chromosome: 'A-' corresponds to that of *cox1*-m9 in A3208 (GenBank acc. no. HQ288823) and 'B+' to that of *cox1*-m4 in B3209 (GenBank acc. no. EU123537). The most 3' modules are marked by an asterisk. (k), unknown module number. For annotated gene whose 5' modules are unknown, module numbers have been estimated. X1-X3, yet unidentified genes detected in cDNA libraries and mapped back to mtDNA. Slash, not applicable.

[b]dpape_s, cDNA library, made from cellular poly(A) RNA. dpcirc, dpmod2, dpmod45, dpmod5, dpmod9, dpmod59, libraries of amplicons from RT-PCR performed on circularized RNA, using divergent primers that anneal in various modules (RNA was not 3' dephosphorylated prior to circularization). dpapg, library of amplicons from RT-PCR using convergent primers, to target hypothetical ppRNAs.

Supplementary Table S2. Transcripts of flanking regions cleaved off from putative module precursor RNAs.

Transcript regions	Transcript start / end positions relative to module	Clone ID
upstream of module *rnl*-m(k)	-1 / -226	dp5971
upstream of module *cox1*-m4	-10 / -569	dp5654

Supplementary Table S3**.** Relative abundance of observed putative processing intermediates of mono-module transcripts[a]

	Number of transcripts detected					
Library	Type 1 ---□---	Type 2 ---□	Type 3 ---□AAA	Type 4 □---	Type 5 □	Type 6 □AAA
cDNAlib (dpape_s)	/	/	10	/	/	2[b]
circRT-PCR (dpmod2)	14	0	0	2	0	0
circRT-PCR (dpmod9)	18	2	2	0	2	6
circRT-PCR (dpmod45)	2	1	0	3	7	0
circRT-PCR (dpcirc)	4	0	0	1	2	0
circRT-PCR (dpmod5)	0	0	0	0	1	0
RT-PCR (dpapg)	1	1	0	0	0	0

[a]□, module; --, module-adjacent sequence; AAA, poly(A) tail. Slash, type of RNA intermediate not targeted by the corresponding experiment. The table enumerates transcripts including a single module (5', internal, and terminal combined) of a given library (which also contains oligo-module transcripts and mature transcripts). For a description of libraries, see Methods of main text. Note that the observed >65 transcripts of *rnl*-m(k)* (see Supplementary Table S1) were not counted here as processing intermediates, because it is yet unknown whether mitochondrial LSU-rRNA of *Diplonema* is trans-spliced or naturally fragmented as seen in a number of other protists.

[b]Bias against this type is due to the selection of long inserts chosen for cDNA library sequencing.

Supplementary Table S4. Termini of oligo-module transcripts.[a]

Type of intermediate	Included gene modules[b]	Clone ID[c]
□□□	*cox1*-m3..m5	dp7008
	cox1-m3..m6	dp7011
	nad7-m5..m8	dp9873
□□□--	/	/
□□□AAA	*atp6*-m3..m4*	dp10245
	cob-m4..m6*	dp0314
	cob-m5..m6*	dp0317
	cox1-m6..m9*	dp0333
	nad4-m6..m8*	dp0115
	nad5-m2..m12*	dp10288
---□□□---	*cox1*-m2..m3	+[d]
	nad5-m6..m7	+[d]
---□□□	/	/
---□□□AAA	*atp6*-m3..m4*	dp0414, dp1977
	cox1-m2..m9*	dp6281, dp6287
	cox1-m3..m9*	dp6277, dp6282
	nad7-m5..m9*	dp4282

[a]□□□, transcripts consisting of ≥2 modules. For other symbols, see legend of Supplementary Table S3. Slash, not detected.
[b]Asterisks indicate the terminal modules.
[c]Intermediates detected in cDNA libraries and RT-PCR experiments, and identified by sequencing.
[d]Intermediates detected by targeted RT-PCR and identified by their amplicon sizes.

Supplementary Table S5. Experiments testing different scenarios of *cox1* RNA editing.[a]

Experiment[b]	Clone series	Number of clones examined[c]	Number of clones for scenario			
			..M4-M5..	..M4-≠6xU-M5..	..M4-6xU	6xU-M5..
Poisoned primer extension	/	/	None	None	/	/
RT-PCR on non-circularized RNA, 5' phosphorylated	dp54xx	91	None	None	/	/
RT-PCR on circularized RNA, 5' phosphorylated	dp56xx	10	None	None	None	None
	dp61xx	34				
	dp62xx	50				
RT-PCR on circularized RNA, 5' phosphorylated, 3' dephosphorylated	dp70xx	30	None	None	**12**	None
	dp73xx	90			**17**	

[a]Scenarios are depicted in Supplementary Fig. S1. Slash means not applicable. None, not found; ..M4, Module 4 that may or may not be attached to upstream modules; M5.., Module 5 that may or may not be attached to downstream modules; M4-M5, both modules joined without intervening Us; M4-≠6xU-M5, both modules joined with either more or less than six intervening Us; M4-6xU, Module 4 with six Us appended to its 3' end; 6xU-M5, Module 5 with six Us attached to its 5' end.
[b]5' end phosphorylation and 3' end dephosphorylation was performed by treatment of RNA with T4 polynucleotide kinase possessing or lacking 3'-phosphatase activity, prior to ligation (see Material and Methods of main text).
[c]Total number of clones with insert and readable sequence.
[d]The remaining 18 and 73 clones of the experiment represent the scenario '..M4-6xU-M5..' (which corresponds to the mature transcript), as well as fully and partially processed modules.

Supplementary Table S6. Experimental detection of RNAs potentially guiding trans-splicing and RNA editing of *cox1*.

cox1-junction	Targeted ppRNA region[a]	Primers used	Clone series from separate experiments (nr. of clones holding insert/having unique sequence)[b]	Nr. of total (unique) sequen-ces combined[c]	Nr. of clones and sequence type across all experiments for a given junction[d]	Interpretation
M2/M3	Central	dp88 (RT) + dp80	dp7901-96 (24/3)	53 (15)	**32 x *cox1*-m2/m3-junct_antisense-primed**, 1 type;	**ppRNA candidate**
			dp8373-96 (1/1)		16 x nuclear (1 *rns*), 11 types;	No predominant sequence
			dp8401-24 (1/1)		3 x *cox1*-m3-m1+flanking_antisense-primed, 1 type;	Unspecific priming of dp80
			dp9237-72 (27/14)		2 x unknown, 2 types;	Unspecific priming (nuclear?)
M3/M4	Central	dp146 (RT) + dp147	dp8425-72 (4/3)	4 (3)	**1x *cox1*-m3/m4-junct_antisense-primed**, 1 type;	**ppRNA candidate**
			dp9273-84 (0/0)		2 x unknown, 1 type;	Unspecific priming (nuclear?)
					1 x module-flanking;	Unspecific priming of both pr.
M4/M5	Central	dp129 (RT) + dp109	dp6401-96 (2/2)	126 (38)	**67 x cox1-m4/m5-junct_antisense-primed**, 1 type;	**ppRNA candidate**
			dp6501-96 (6/6)		22 x nuclear (8 *rns;* 1 spliced leader), 20 types;	No predominant sequence
			dp6601-96 (63/7)		14 x primers, 1 type;	Artefact
			dp7101-96 (28/13)		12 x module-flanking, 8 types;	Unspecific priming of dp129
		dp129(RT) + SMART	dp6901-96 (27/24)		8 x unknown, 6 types;	Unspecific priming of both pr.
					3 x short (>15 nt), 2 types;	No primer recognizable
M5/M6	Central	dp150 (RT) + dp151	dp8473-96 (1/1)	6 (2)	**5 x *cox1*-m5/m6-junct_antisense-primed**, 1 type;	**ppRNA candidate**
			dp9285-96 (5/2)		1 x short sequence (<15 nt)	Not interpretable
M8/M9	Central	dp154 (RT) + dp41	dp8649-96 (1/1)	62 (17)	31 x module-flanking, 1 type;	Unspecific priming of both pr.
			dp9337-96 (61/17)		8 x nuclear, 4 types;	Unspecific priming of both pr.
					5 x *cox3*-m1+flanking_antisense-primed, 1 type;	Unspecific priming of both pr.
					4 x *cox2*-m3_sense-primed, 1 type;	Unspecific priming of both pr.
					2 x *atp6*-m4+flanking_antisense-primed, 1 type;	Unspecific priming of both pr.
					2 x *cox1*-m2+flanking_sense-primed, 1 type;	Unspecific priming of both pr.
					2 x *cox1*-m6+flanking_antisense-primed, 1 type;	Unspecific priming of both pr.
					2 x *nad1*-m2+flanking_antisense-primed, 1 type;	Unspecific priming of both pr.
					1 x *cox1*-m8/m9-junct_antisense_primed;	**ppRNA candidate**
					1 x *nad5*-m9+flanking_sense-primed;	Unspecific priming of both pr.
					1 x *nad5*-m11+flanking_antisense-primed;	Unspecific priming of both pr.
					1 x *cox3*-m1+flanking_sense-primed;	Unspecific priming of both pr.
					1 x *geneX1*-m(k)+flanking_sense-primed;	Unspecific priming of both pr.
					1 x *geneX11*-m(k)+flanking_antisense-primed;	Unspecific priming of both pr.
M2/M3	Distal	dp138 (RT) + dp139	dp8001-24,37-48 (30/7)	30 (7)	16 x *cox*-m2+flanking, 2 types;	Unspecific priming of dp139
					10 x nuclear (≤7 reads in 1 cluster; rep: dp8044), 3 types;	Unspecific priming of both pr.
					3 x primers, 1 type;	Unspecific priming of dp138
					1 x *cox*-m2_sense-primed;	Artefact
M3/M4	Distal	dp148 (RT) + dp149	dp8149-72 (24/2)	24 (2)	24 x *cox1*-m3-portion_antisense-primed, 2 types	Unspecific priming of dp149
M5/M6	Distal	dp152 (RT) + dp153	dp8173-96 (15/7)	15 (7)	7 x *cox1*-m5+flanking_antisense-primed, 1 type;	Unspecific priming of dp152
					7 x nuclear (1 *rns*), 5 types;	No predominant sequence
					1 x *cox1*-m5/m6_antisense-primed (dp8189)	**ppRNA candidate** (includes 62 nt of Module-5 3'end and 67 nt of Module-6

5'end).

M7/M8	Distal	dp84 (RT)	dp7801-96 (36/22)	73 (34)	27 x nuclear (2 *rns*; ≤6 reads in 1 cluster; rep: dp8074), 16	No predominant sequence
		+ dp41			t.s;	Unspecific priming of dp84
		dp141 (RT)	dp8025-36,49-96		8 x *cox1*-m8+flanking_antisense-primed, 1 type;	Unspecific priming of both pr.
		+ dp140	(37/22)		7 x gene*X11*-m(k)+flanking_antisense-primed, 1 type;	No predominant sequence
					6 x unknown (nuclear?), 3 types;	Unspecific priming of dp84
					5 x *cox1*-m3+m8-9_sense-primed, 1 type;	Unspecific priming of dp84
					5 x *cox1*-m8+flanking_sense-primed, 1 type;	Unspecific priming of both pr.
					4 x module-flanking, 3 types;	Unspecific priming of both pr.
					4 x *cox1*-m2+flanking_antisense-primed, 1 type;	No primer recognizable
					2 x short (>15 nt), 2 types;	Clone mix-up
					1 x *cox1*-m2/m3 junction (dp88+80), 1 type;	Unspecific priming of dp84
					1 x *cox1*-m8_antisense-primed;	Unspecific priming of both pr.
					1 x *nad5*-m5_antisense-primed;	Artefact
					1 x primers;	Clone mix-up
					1 x clone from unrelated experiment;	

[a]Determination of central region by 'convergent' RT-PCR, and of distal regions by 'divergent' RT-PCR on circularized RNA (see Fig. 3A, B, main text).

[b]Each clone series (dp7901-96, dp8373-96, etc.) were obtained from separate RT-PCR reactions. Only clones with non-vector inserts are counted; clones holding vector-derived inserts were (mostly) eliminated by the phred vector clipping utility (see Methods in main text). Unique sequences are cluster representatives after clustering with CD-Hit at 90% identity. The different experiments yielded unequal proportions of total vs unique sequences and these variations are most likely due to PCR bias. In two instances, considerably unequal outcomes are also observed between replicates of the same experiment (see dp7901-96 compared to dp9237-72, and between dp6601-96 and dp6901-96). These variations might be due to a different enzyme batch used or to minor (unintentional) differences in the experimental protocol employed.

[c]Unique sequences (types) across all experiments for a given junction.

[d]Sequences referred to as spurious in Table 1 of the main text are characterized here in more detail. m2/m3-junct, etc., sequence at the junction of Modules 2 and 3, etc. Sense-primed, amplicons derived from mitochondrial transcripts where the RT-primer annealed with the coding strand. Antisense-primed, annealing of the RT-primer with antisense RNA. Nuclear, amplicon matched fully with the available nuclear sequence of *Diplonema*. Module-flanking, amplicon sequence is identical with sequence up- or downstream of gene modules in the available mtDNA sequence. The sequence of clones highlighted in red is as follows (compl_dp153 means the reverse complementary sequence of primer dp153):

dp8044: 5'dp138-GGGTGGTCAAATT-compl_dp139

dp8074: 5'dp84-GGGTTCCCGACTACCGCAATCGTGAGCAGCACGAATG-compl_dp141

dp8189:5'dp152-TTTG-compl_dp153-dp153-
GATGGATGTAGGTAGCGCTATGAGCAACGTGGCACTGGAGAAGGACATGAGG
AGGAAGGCCACGACGTACCACGCAGCGCGTGCTGGC-compl_dp152

(Two amplicons cloned into one vector molecule). In experiments targeting the central region of the hypothetical ppRNA, candidates must carry both primers in correct orientation, the primer used for RT must prime the antisense transcript, and the sequence in between these primers must match that of the corresponding module junction (U:G pairs allowed, but not insertions/deletions). In the case of experiments targeting the distal region of the hypothetical ppRNA, candidates must carry both primers in correct orientation with >=1 nt in between, and either match yet unassigned mitochondrial coding regions (within cassettes) or must be the predominant type of cloned amplicons. Representative clones derived from ppRNA-candidates are: dp9264 (Module2/Module3-central), dp8451 (Module3/Module4-central),

dp7136 (Module4/Module5-central), dp9289 (Module5/Module6-central), dp8689 (Module8/Module9-central), and dp8189 (Module5/Module6-distal).

Supplementary Table S7. Guide candidates (simple model) predicted in mitochondrial genome sequences.[a]

Guide for *cox1* junction	Number of hits	Genome data set	JUNCTION SEQUENCE ppRNA sequence
M1/M2	0	/	/
M2/M3	3	dpapimt.all	5'-TCGGGA CATGGC-3' ||:||| |||||| 3'-agtcct gtaccg-5'
M3/M4	0	/	/
M4/M5	6	dpapimt.all	5'-TTTTTT CGCTCT-3'
	4	prag-mt-readings	:|:||| |:||||
			3'-gagaaa gtgagg-5'
M5/M6	5	dpapimt.all	5'-ATGGTG GGACTG-3'
	2	prag-mt-readings	|::|:| ||||||
			3'-tgtcgc cctgac-5'
M6/M7	1	prag-mt-readings	5'-TCTCAT TAGGAG-3' :||||| :||||| 3'-ggagta gtcctc-5'
M7/M8	2	dpapimt.all	5'-CGTGTA CAGGTG-3'
	5	prag-mt-readings	|:||:| ||||:|
			3'-gtacgt gtccgc-5'
M8/M9	3	dpapimt.all	5'-CCTAGG TACAGT-3'
	2	prag-mt-readings	||||| ||||||
	1	chromosome dp3207	3'-ggatcc atgtca-5'

[a]For datasets see Methods section of the main text. The sequence of one candidate (in blue) is indicated for a given junction (black); the junction sequence shown consists of the six terminal residues of Module i and the six first residues of Module i+1. The datasets dpapimt.all and prag-mt-readings contain sequences from partial chromosomes. dp3207 is the sequence of a complete chromosome (see Methods section of main text).

Supplementary Table S8. Statistics of the computational search for RNA and DNA molecules guiding *cox1* trans-splicing and RNA editing.[a]

Data set[b]	Parameters	Total number of distinct guides	Min. / mean / max. number of distinct guides per guide class[c]	Number of guide classes [Number of guide groups contained in each class][d]
Nuclear genome	Topology 1; L=0..5; d1, d2=0..83;	367,275,143	1,474,000/ 6,121,000/ 8,665,000/	6 classes [84 groups each]
	Topology 2; L=0..50; d1, d2=0..5;	27,321,323	89,130/ 535,700/ 566,800/	51 classes [6 groups each]
Nuclear ESTs	Topology 1; L=0..5; d1, d2=0..83;	3,719,119	582,200/ 619,900/ 633,400/	6 classes [84 groups each]
	Topology 2; L=0..50; d1, d2=0..5;	197,463	3,574/ 3,872/ 4,169/	51 classes [6 groups each]
Mitochondrial genome	Topology 1; L=0..50; d1, d2=0..83;	15,653,575	163,600/ 306,900/ 412,900/	51 classes [84 groups each]
	Topology 2; L=0..50; d1, d2=0..83;	15,596,645	162,700/ 305,800/ 407,700/	51 classes [84 groups each]
Mitochondrial cDNAs	Topology 1; L=0..50; d1, d2=0..83;	37,520	571/ 735/ 990/	51 classes [84 groups each]
	Topology 2; L=0..50; d1, d2=0..83;	35,799	514/ 701/ 937/	51 classes [84 groups each]

[a]For an illustrative description of parameters, see Fig. 3C, D, main text. Minimum match length of paired regions is a=6. The number of allowed G:U pairs is ≤3. Guides that have the potential to also direct joining of non-cognate modules have been eliminated. The minimum number of junctions covered by guides in a given group is six (out of eight), otherwise the number of guides reported is set to zero. Each single data set contains guides for all classes tested. Numbers of guides count those with distinct sequence.

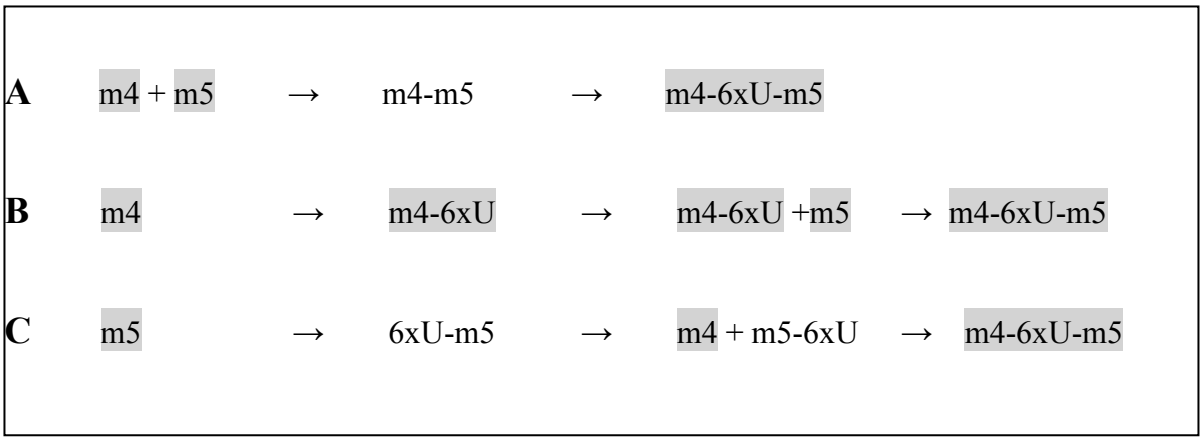
[b]For details on the data sets, see Methods Section in main text.
[c]The minimum, mean and maximum numbers of guides within a class. A class is defined by sharing a specific topology and having identical bridge length, L.
[d]A guide group contained in a class is defined by identical d1, d2.

Supplementary Table S9. Primers used in this study.

Primer-ID	Sequence	Targeted gene; anneals with strand
dp26	GGTCTTCGGGCACCCAGAG	*cox1*-Module 4; antisense
dp33	CCAGGAGAACACCTTGATGGA	*cox1*-Module 5; sense
dp41	CCACAGGTGGTGGATACCACT	*cox1*-Module 9; sense
dp46	CCGTTGAGCAGACCGAAGACCGCTCCT	*cox1*-Module 7, sense
dp48	GGCCTACTAGTGGTATGACTG	*cob*-Module 6; sense
dp49	CGTGGTAGGATGCCTAGCTGGG	*cob*-Module 6; antisense
dp50	GCTGTAACCCATCCACTAACGT	*cob*-Module 6; sense
dp55	GCTACCAGCGCAGGAGCTCA	*nad5*-Module 12; sense
dp68	CGTAGTAGCAGTACCAGCTGTA	upstream cassette B; 'sense'
dp69	GGAGCAGGAGTATGCTGTAGA	downstream cassette B; 'antisense'
dp80	CCACTAGCAGCCATG	*cox1*-Module 3, antisense
dp82	CCATCAAGGTGTTCTCCTGGA	*cox1*-Module 5, sense
dp88	CCCTAAGGTGAACAACGTCGG	*cox1*-Module 2; sense
dp90	GCTGATGTGCAGCTCTCCCTTTGGA	*nad7*-Module 9; sense
dp91	GAGTGCGTCCTGCAGACCTAGGA	*nad7*-Module 9; antisense
dp92	TACCTACGGGTACTGATAGCACGG	*nad7*-Module 9; sense
dp94	CGCTCCGGAGTAGGACACCAGATG	*nad7*-Module 9; sense
dp95	ACCCGTAGGTACATACGGTGACA	*nad7*-Module 9; antisense
dp96	GTTGATAGGCACAGTGGACGT	*nad7*-Module 9; antisense
dp97	CATGACAGCGTCTGCTAAGGATGT	*nad7*-Module 9; sense
dp101	GCAGTAGCGTAGTGATACAG	*nad7*-Module 9; antisense
dp102	TGATAGCACGGTACAGCTGT	*nad7*-Module 9; sense
dp109	CGTACACCATGCCCAGCATGTTGTAGAGC	*cox1*-Module 5; sense
dp111	GGCCCAGGTACAGATGTCCTA	*atp6*-Module 4; sense
dp115	GAGGAGGTGCTATGAGCCTAA	*nad7*-Module 9; antisense
dp116	CAGCTGTACCGTGCTATCAG	*nad7*-Module 9; antisense
dp118	CCTCCAGAGTGTCCTAGCTGT	*nad7*-Module 9; antisense
dp119	CAGTGCATGTAGTCCTAGGA	*nad7*-Module 9; sense
dp120	TGCTAGGGACATCCTTAGCA	*nad7*-Module 9; antisense
dp126	CCGTACCAGTGTTCCTATCAGTGA	*cox1*-Module 4; sense

dp129	ATTCCCTACATCGAGGAGGA	*cox1*-Module 4; antisense
dp133	CCTGAGTAGCAGCTAACAGCA	*nad5*-Module 12; sense
dp134	ATACTCCACGGCATGGCAT	*nad5*-Module 12; antisense
dp147	TACTAGCAGTGCACCTGTGA	*cox1*-Module 4; antisense
dp148	TCACAGGTGCACTGCTAGTA	*cox1*-Module 3; sense
dp150	AAGGTGTTCTCCTGGATGGT	*cox1*-Module 5; sense
dp154	TGCACCTGATGGATACCTAG	*cox1*-Module 8; sense
dp164	GGTGTCCACCTAACATGCTG	*nad7*-Module 9; antisense
dp165	CCTTGTAGCATGACTCCCTG	*nad4*-Module 8; antisense
dp166	CCTTGTAGCATGACTCCCTG	*nad4*-Module; sense
dp167	GGCGCTGTCGTGTACCTTCC	*atp6*-Module 4; antisense
dp170	GCTGGTGCACTGTACATGG	*cob*-Module 6; sense
dp171	GTGGCTGTATCCACCCATAT	*cob*-Module 6; antisense
dp172	GGTGTACTGCCTGTGATGG	*nad5*-Module 12; antisense
dp173	CTGCTGATGGACTGGGTGAT	*nad5*-Module 12; sense
dp174	CCATGAACCAGCTCAGCTGT	*nad5*-Module 12; antisense
dp175	AGTACGAGCTGCTGTCCTG	*nad7*-Module 9; sense
dp207	ATTCCCTACATCGAGGAGGACTTTTTTCGCTCT	*cox1*-junction M4/M5; sense

3.2.15 Supplementary figure



Supplementary Figure S1. Tested scenarios of RNA editing in *cox1*. (A) U-Insertion editing, after joining of Modules 4 and 5. (B, C) Editing by addition of Us either to the 3' end of Module 4 (B) or to the 5' end of Module 5 (C). Experimentally detected transcript intermediates are highlighted by a grey background. Note that m4-6xUs is only observed when RNA is dephosphorylated at the 3' terminus prior to circularization. See also Supplementary Table S5.

3.2.16 Supplementary references

Driscoll DM, Wynne JK, Wallis SC, and Scott J 1989. An in vitro system for the editing of apolipoprotein B mRNA. *Cell* **58**:519-525.

4 Travaux non publiés sur l'édition des ARNm et l'existence d'ARN antisens

Les résultats présentés dans cette section seront l'objet d'une publication en préparation qui sera soumise après validation par l'analyse de nos résultats obtenus par RNA-Seq.

4.1 Editions supplémentaires dans la mitochondrie de *D. papillatum*

La séquence génomique de chaque module a été alignée à l'ADNc du gène correspondant pour identifier des sites d'éditions probables. Nous avons ainsi découvert que les gènes *nad1* et *X1* sont apparemment édités par l`addition de 15-16 et trois uridines respectivement, au dernier module (Figures 22, 23). Aucune édition n'a été retrouvée dans les séquences ADNc des gènes *atp6*, *cox2*, *cox3*, *nad4*, *nad5*, *nad7* disponibles à ce jour (Tableau VII).

4.2 Les modules 5' des gènes mitochondriaux de *D. papillatum*

Nos tentatives de compléter par RT-PCR les modules prédits en 5' des gènes *nad4* et *atp6* (ces gènes ont été prédits par comparaison à ceux de la plupart des eucaryotes) n'ont pas réussi. Les analyses préliminaires des données RNA-Seq effectuées par la Dr Burger, indiquent qu'il n'existe pas d'autres modules en amont des modules désignés provisoirement *atp6*-m2 et *nad4*-m2.

Apparemment les protéines correspondant à ces deux gènes sont plus courtes que celles de la plupart des eucaryotes mais comparables en longueur avec celles des kinétoplastides.

4.3 Absence de longs ARN antisens de *cox1* chez tous les diplonémides

Nous avons testé si un ARN antisens du transcrit complet de *cox1* existe et pourrait servir de matrice pour lier les neufs modules de *cox1* à la fois. Cette question a

été abordée chez *D. papillatum* mais aussi chez *D. ambulator*, *D.* sp. 2 et *R. euleeides*. Notons que des ARN antisens sont impliqués réarrangement du génome nucléaire chez les ciliés (Nowacki*, et al.*, 2011). Les expériences RT-PCR utilisant des amorces couvrant les neuf modules de *cox1* n'ont pas permis d'amplifier des molécules d'ARN antisens chez les trois diplonémides (Tableau VIII). Donc nous concluons que ces molécules n'existent pas chez les diplonémides ou sont présentes en concentration trop faible pour être détectées.

4.4 Absence d'ARN antisens de gènes *atp6*, *cob*, *cox2*, *nad5* et *nad7*

Nous avons également voulu vérifier s'il existe des ARN antisens de transcrits d'autres gènes mitochondriaux que *cox1* chez *D. papillatum*. Les RT-PCR réalisées avec des amorces spécifiques couvrant au minimum deux modules n'ont pas permis d'identifier de potentiels ARN antisens (Tableau VIII). Comme pour *cox1*, nous concluons qu'il n'y a pas d'ARN antisens pour les autres gènes mitochondriaux testés ou qu'ils sont présents en concentration trop faible pour être détectés.

4.5 Les transcrits de *rnl* détectés par Northern Blot

Pour identifier les transcrits du gène *rnl* de *D. papillatum*, nous avons réalisé des expériences de Northern Blot. En utilisant de l'ARN total et comme sonde, le module terminal, nous avons obtenu deux produits: un d'une taille de 0.3 kb correspondant au module terminal et un autre, prédominant, à environ 0.9 kb. La même hybridation avec de l'ARN poly (A) donne uniquement un produit d'environ 0.3 kb correspondant au module terminal polyadenylé (Figure 25). Ce résultat était inattendu et difficile à interpréter. Soit le transcrit de 0.3 kb qui est polyadenylé est un de plusieurs fragments d'ARNr matures comme chez certains organismes où l'ARNr est fragmenté et parfois découpé en plus de 10 morceaux (Milbury*, et al.*, 2010, Feagin*, et al.*, 2012, Lavrov*, et al.*, 2012); le transcrit de 0.9 kb serait ainsi un précurseur avant la maturation. Soit le produit de 0.9 kb est la forme mature de *rnl* et le transcrit de 0.3 kb est un module avant

l'épissage en *trans* ce qui impliquerait que sa queue poly (A) est enlevée simultanément à l'épissage en *trans*. Ce résultat sera discuté plus en détails dans la section 4.7.

4.6 Séquençage du transcrit du gène *rnl* de 0.9 kb mitochondrial de *D. papillatum*

Pour caractériser le transcrit de 0.9 kb du gène *rnl* obtenu par Northern Blot nous avons réalisé une RT-PCR après circularisation de l'ARN total. L'amplification a été avec deux amorces situées aux extrémités 3' et 5' du module terminal de *rnl*, nous avons obtenu un produit de plus de 0.9 kb. Le séquençage montre, en amont du module terminal, une série d'environ 30 uridines puis un module non identifié auparavant. Ce transcrit est apparemment également édité entre deux modules (Figure 24). Nous avons donc finalement réussi à détecter le gène et le transcrit entier de la LSU de l'ARNr chez *D. papillatum.*

4.7 La maturation de la LSU-ARNr mitochondrial chez *D. papillatum*

Nos tentatives infructueuses initiales pour amplifier l'extrémité 5' du gène *rnl* ont mené à une première hypothèse selon laquelle la LSU-ARNr mature est fragmentée elle aussi et que ses fragments ne sont pas joints par épissage en *trans* (Vlcek*, et al.*, 2011). Mais récemment, nous avons pu montrer que le transcrit de 0.9 kb est constitué de deux modules de 535 et 354 pb et qu'il y a de l'édition entre le module 1 et 2 par l'ajout de 27-30 uridines. Cette édition est confirmée par la comparaison de la séquence ADNc avec la séquence génomique. L'expérience de Northern Blot mentionnée ci-haut indique que le transcrit de 0.9 kb de *rnl* n'est pas polyadénylé car aucun transcrit de 0.9 kb n'est visible dans la fraction poly (A). Notons que les deux séquences de la LSU-ARNr obtenues par RT-PCR portent 17-19 As au 3' terminal tandis que le transcrit du module 2 porte jusqu'à 47 As. Le fait que le transcrit de 0.9 kb n'est pas détectable dans l'ARN poly (A) utilisé pour l'expérience de Northern Blot concorde avec le fait que la colonne cellulose oligo dT utilisée, n'est censée retenir que des transcrits avec des queues poly

(A) d'une longueur de plus de 15 nt. Apparemment la plupart des LSU-ARNr portent une queue poly (A) inférieure à 15 nt. Ce qui impliquerait que durant la maturation de *rnl*, la queue poly (A) du module 2 est considérablement raccourcie (Aphasizhev & Aphasizheva, 2011).

Des données RNA-Seq confirment l'épissage en *trans* et l'édition du transcrit de la LSU-ARNr. Cet ARNr apparait plus court (0.9 kb) comparé à celui des kinétoplastides qui fait 1.1 kb. Le transcrit de 0.9 kb es-t-il la LSU-ARNr mature ? Ou y aurait-il d'autres fragments de la LSU-ARNr mitochondriale de *D. papillatum* ? Chez les euglénides, les ARNr sont fragmentés en deux pièces liées par appariement de bases (Spencer & Gray, 2011). Pour élucider la structure des ARNr, nous tentons de purifier des ribosomes mitochondriaux de *D. papillatum*.

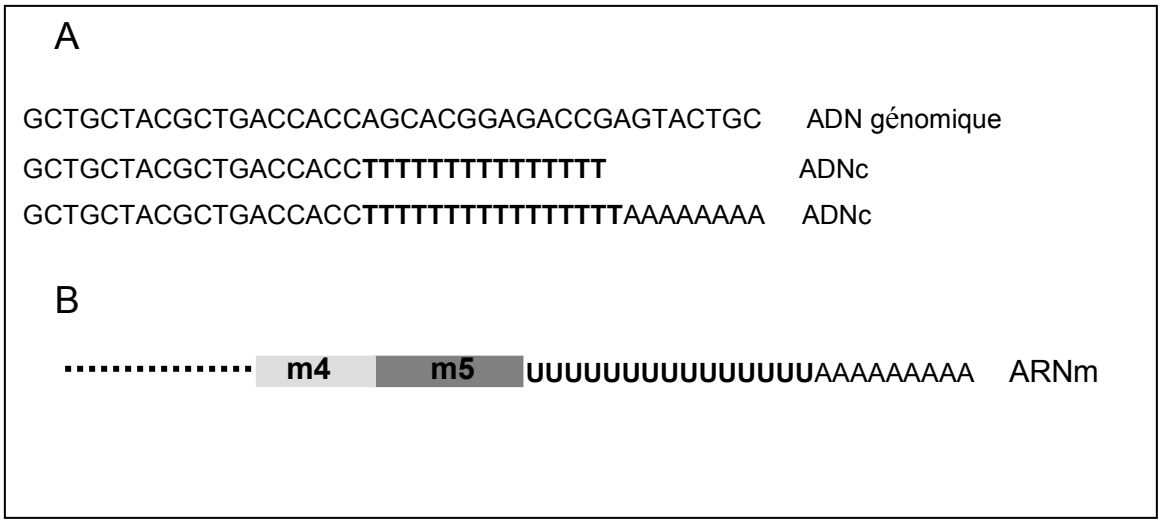


Figure 22. Édition du gène *nad1* par addition de 15-16 uridines au module 5. A. Alignement de l'ADN génomique (6 clones) avec ADNc (deux clones) du module 5. Les T ajoutés par l'édition sont en gras dans l'ADNc. B. ARNm incomplet du gène *nad1* avec les modules 4 et 5 en gris.

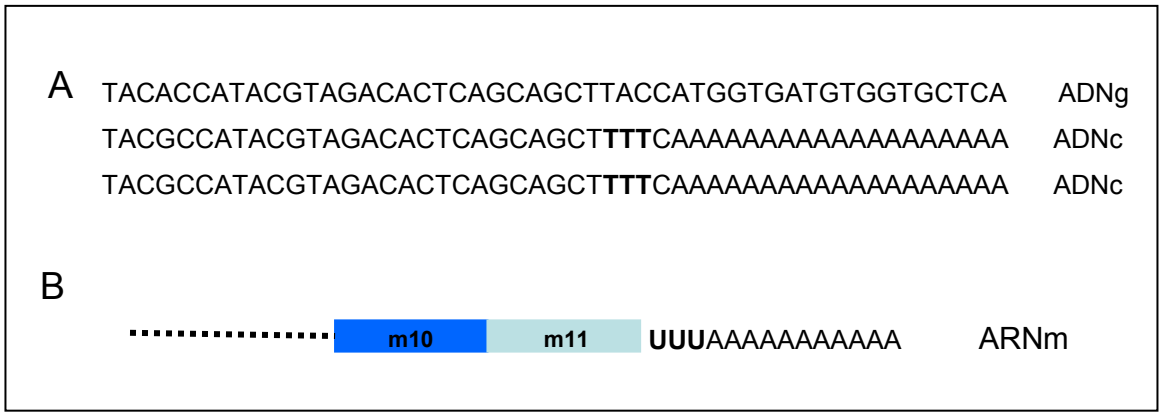


Figure 23. Édition du gène *XI* par addition de trois uridines au module 11.
A. Alignement ADN génomique (4 clones) avec ADNc (2 clones) du module 11. Les trois T ajoutés par l'édition sont en gras dans l'ADNc. B. ARNm incomplet du gène *XI* avec les modules 10 et 11 en bleu.

Figure 24. Édition du gène *rnl* par addition de 27-30 uridines au module 1. A. Alignement de l'ADN génomique (29 clones) et ADNc (deux clones du module 1). B. ARNm du gène *rnl* avec les U ajoutés en gras, les modules 1 et 2 sont en violet et ocre.

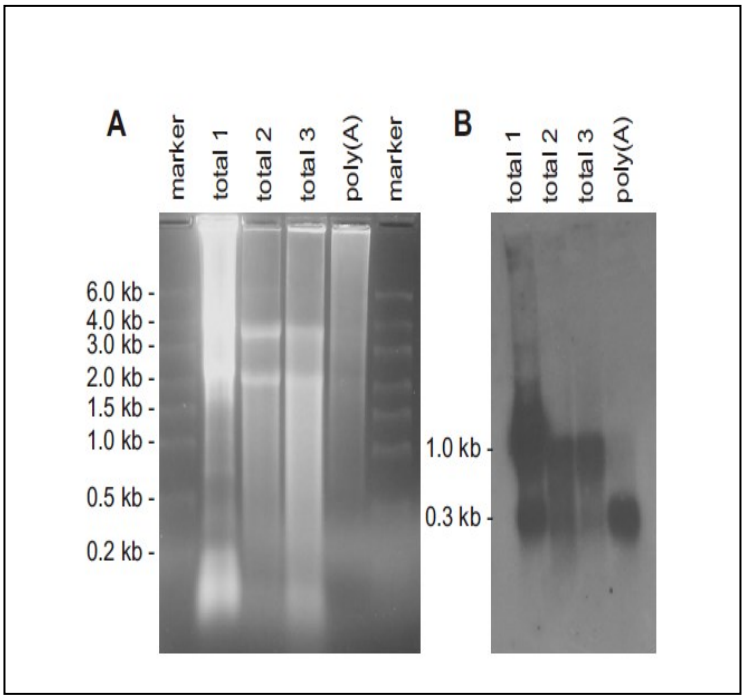


Figure 25. Détection de transcrits mitochondriaux de *rnl chez D. papillatum* par Northern Blot. A. Les différents ARN sont séparés sur gel agarose 1.5% dénaturant (37% formaldéhyde). B. Hybridation des ARN avec le module terminal de *rnl*. Transcrits du gène *rnl* détectés après exposition de 6 H au film. L'hybridation avec les ARN totaux donne deux transcrits; 0.3 kb et 0.9 kb tandis que l'hybridation avec le poly (A) montre seulement le transcrit de 0.3 kb.

Tableau VII. Gènes mitochondriaux de *D. papillatum* analysés pour rechercher des sites d'édition.

gènes	ADNc	Modules[a] génomiques
atp6	M1-M3	M1, 2, 3
cox2	M1-M4	M1, 2, 3,4
cox3	M1-M3	M1, 2, 3
cob	M1-M6	M1, 2, 3, 4, 5, 6
nad1	M4-M5	M1, 2, 3, 4, 5
nad4	M6-M8	M1, 2, 3, 4, 5, 6, 7, 8
nad5	M2-M11	M2, 3, 4, 5, 6, 7, 8, 9, 10, 11
nad7	M1-M9	M1, 2, 3, 4, 5, 6, 7, 8, 9
rnl	M1-M2	M1, 2
x1[b]	M10-M11	M10, 11

[a] M correspond à la nouvelle numérotation des modules
[b] Gène non identifié

Tableau VIII. Expériences d'amplification d'ARN mitochondriaux antisens de diplonémides.

Espèces	Gènes	Jonctions ciblées	Produits obtenus
D. ambulator	*cox1*	M1/2, M2/3, M3/4, M4/5, M5/6, M6/7, M7/8, M8/9	Non
D. sp. 2	*cox1*	M1/2, M2/3, M3/4, M4/5, M5/6, M6/7, M7/8, M8/9	Non
R. euleeides	*cox1*	M1/2, M2/3, M3/4, M4/5, M5/6, M6/7, M7/8, M8/9	Non
D. papillatum	*atp6*	M2/3	Non
	cob	M5/6	Non
	cox2	M2/3, M3/4	Non
	nad5	M6/7, M8/9, M10/11, M11/12	Non
	nad7	M6/7, M7/8, M8/9	Non

Chapitre 4. Discussion

Le but de cette étude était la caractérisation des processus post-transcriptionnels dans la mitochondrie des diplonémides. Notre étude d'eucaryotes peu connus a révélé une architecture de génome et une expression des gènes exceptionnelles. Dans l'article 1, nous avons démontré que la fragmentation des gènes, l'épissage en *trans* et l'édition des transcrits mitochondriaux sont des caractères communs aux diplonémides.

Dans l'article 2, nous avons caractérisé les étapes de maturation post-transcriptionnelles des transcrits des gènes fragmentés. Ces étapes incluent l'excision des extrémités 5' et 3', l'épissage en *trans* et /ou l'édition et la polyadénylation.

Bien que la fragmentation des gènes, l'épissage en *trans* et l'édition d'ARN ont été décrits auparavant, ces phénomènes s'avèrent être très particuliers dans les mitochondries des diplonémides.

1 Gènes fragmentés-produits fragmentés, gènes fragmentés-produits contigus

La fragmentation des SSU et LSU ARNr mitochondriaux existe dans des organismes tels que les algues vertes, les apicomplexés, les ciliés et les dinoflagellés. Or des gènes fragmentés codant pour des protéines et une fragmentation systématique telle qu'observée dans les mitochondries des diplonémides, sont rares (Burger*, et al.*, 2003, Burger*, et al.*, 2011). Les quelques exemples connus à l'extérieur des diplonémides produisent apparemment plusieurs ARNm ainsi qu'une protéine fragmentée; par exemple, le gène *nad1* codant pour la sous unité 1 de la NADH déshydrogénase des ciliés (Burger*, et al.*, 2011).

Un seul autre cas de fragmentation d'un gène donnant un produit contigu a été découvert dans la mitochondrie de certains dinoflagellés (Waller & Jackson, 2009, Jackson*, et al.*, 2012). Leur gène *cox3* possède trois formes de transcrits différentes : deux transcrits courts et polyadénylés et un troisième qui est le résultat de la liaison des deux premiers (Waller & Jackson, 2009). Il y a également une queue oligo adénine au niveau du site de liaison et dont la longueur varie d`une espèce à une autre. Cela pourrait

impliquer l`existence de molécules guides dans ce mécanisme d`épissage en *trans* (Jackson & Waller, 2013).

2 Mécanismes de l'épissage en *trans*

L'épissage en *trans* impliquant des introns non contigus de groupe I, II, d`introns impliqués dans le splicéosome ou d`introns d`ARNt, met en jeu des séquences *cis* présentes sur les ARN à lier. Or dans la mitochondrie des diplonémides, aucun de ces introns n'a été trouvé, pas même les introns de groupe I et II qui sont habituellement présents dans les organelles (Tableau IX). L'épissage en *trans* chez les diplonémides, constitue un processus unique, qui est précis, non directionnel et coordonné à la maturation des extrémités 5' et 3'et à l'édition.

3 La nature des facteurs *trans* impliqués dans l'épissage en *trans*

L'absence de signatures d'introns et de séquences complémentaires ou conservées dans les régions flanquantes des modules voisins de *cox1*, nous a permis de conclure que la mitochondrie des diplonémides réalise probablement un nouveau type d'épissage en *trans* impliquant des facteurs *trans*.

Notre recherche a permis la découverte des petits ARN antisens, complémentaires aux jonctions des modules. Toutefois, il reste à démontrer leur fonction; c'est-à-dire que ces molécules dirigent l'épissage en *trans* et servent de matrice pour l'édition. Alternativement à des petits ARN, des protéines pourraient également guider l'épissage en *trans* et l'édition dans la mitochondrie de *D. papillatum*. Comme l`épissage en *trans* est très fiable (nous n'avons jamais observé de liaison inapropriée de modules), il y a probablement une protéine pour chaque jonction (Figure 26 B). Avec une centaine de modules à assembler cela ferait une centaine de protéines. Ces protéines pourraient appartenir à la famille des Pentatrico Peptides Repeat proteins (PPRs). Comme mentionné dans l'introduction, ces PPRs sont impliquées dans plusieurs processus post-

transcriptionnels tels que l'épissage en *trans* et l'édition chez les eucaryotes. Selon nos données actuelles, le génome nucléaire de *D. papillatum* encode 42 PPRs (Moreira et Burger non publié).

4 Edition d'ARN

4.1 Types d'édition connus

Telle que décrite dans l'introduction, l'édition des ARN dans la mitochondrie consiste généralement en des insertions, des délétions (d'un ou de plusieurs nucléotides), des substitutions et des conversions dans la séquence de l'ARN. Chez les diplonémides, l'édition apparait comme issue d'une insertion de nucléotides mais en réalité, il s'agit de l'addition de nucléotides à l'extrémité 3' d'un transcrit. Ce type d'édition n'a pas été observé dans d'autres systèmes. Toutefois elle présente des ressemblances avec l'édition dans la mitochondrie des kinétoplastides comme résumé dans le paragraphe suivant.

4.2 Comparaison de l'édition chez *D. papillatum* avec celle qui existe chez les kinétoplastides

L'édition par insertion/délétion d'uridines dans la mitochondrie des kinétoplastides utilise des ARNg comme matrices. Chez les kinétoplastides, l'ARNg a une séquence complémentaire au pré-ARNm. Une des hypothèses est que chez *Diplonema*, un ARNg réalisant l'édition du module 4 de *cox1* se fixerait sur l'extrémité 3'du transcrit. Un mauvais appariement avec ce dernier entrainerait l'ajout de six uridines par une TUTase. La jonction entre les modules 4 et 5 pourrait nécessiter deux ARNg. Un premier qui se fixerait sur le module 4, servant de matrice pour l'édition et un autre qui se fixerait sur les uridines ajoutées au module 4 et sur le module 5 et lierait les deux modules (Figure 27 A). L'édition précéderait l'épissage en *trans* comme nous l'avons démontré expérimentalement (Kiethega*, et al.*, 2013). Une hypothèse alternative postule qu'un

ARNg se fixe sur les deux modules, servant ainsi simultanément de matrice pour l'édition ainsi que pour la liaison (Figure 27 B).

4.3 Rôle biologique de l'édition

Comme chez les kinétoplastides, l'édition chez les diplonémides semble indispensable à la production d'une protéine fonctionnelle. Dans la structure secondaire de la séquence protéique de *cox1,* les deux codons ajoutés par l'édition codent pour des acides aminés qui forment une boucle conservée dans la protéine (Marande, 2007). Finalement l'alignement multiple de la séquence protéique de *cox1,* avec celles de plusieurs eucaryotes suggère fortement que cette édition est fondamentale à la fonction et à la structure de la sous-unité 1 de la cytochrome oxidase. Chez les gènes *cob* et *nad1,* l'édition spécifie une phénylalanine qui est bien conservée chez les euglénozoaires et participe aussi à la formation du codon stop qui est complété par la polyadénylation (Figure 4, annexe).

5 Rôles régulateurs des « petits ARN »

Nos études indiquent que l'épissage en *trans* et l'édition pourraient être guidés par des petits ARN dans la mitochondrie de *D. papillatum*. La mitochondrie de *D. papillatum* serait un exemple de plus dans lequel des petits ARN jouent des rôles spécifiques dans la régulation de l'expression des gènes.

Il est à noter qu'auparavant, on pensait que la régulation de l'expression des gènes était plutôt assurée par des protéines, en l'occurrence les facteurs de transcription.
Depuis quelques années, un nombre croissant de travaux rapporte les rôles des petits ARN dans la régulation de l'expression des gènes procaryotes et eucaryotes.

Chez les bactéries, ancêtres des mitochondries, des ARN de 50-300 pb contrôlent la stabilité de l'ARN et la traduction, par appariement court et imparfait en *trans* avec l'ARNm, soit à proximité du site de fixation des ribosomes, soit sur les RBS (Ribosome Binding Site) (Waters & Storz, 2009, Storz*, et al.*, 2011).

Dans le noyau des eucaryotes, des snRNAs (small nuclear RNA) de 100-200 pb sont impliqués dans l'épissage (Valadkhan, 2010). Ces ARN agissent au sein de complexes ribonucléoprotéiques et jouent un rôle dans l'identification du site d'épissage, l'alignement des exons et la régulation de l'activité catalytique du splicéosome par appariement avec l'ARNm.

Des petits ARN régulateurs localisés dans la mitochondrie sont les ARNg des kinétoplastides servant de matrices pour l'édition des ARNm (Lukes*, et al.*, 2005). L'implication de petits ARNg dans l'épissage en *trans* et dans l'édition chez les diplonémides proposée par notre équipe, renforcera la notion que les petits ARN jouent un rôle central dans l'expression des gènes.

6 Le niveau de régulation de l'expression des gènes

À travers les différents systèmes biologiques, l'expression des gènes est régulée tant au niveau transcriptionnel que post-transcriptionnel. La régulation transcriptionnelle fait intervenir des séquences *cis* notamment les promoteurs, autant que des facteurs *trans* tels que les facteurs de transcription. Quant à la régulation post-transcriptionnelle, les processus impliqués sont l'épissage des introns, l'édition, la maturation des extrémités et l'interférence ARN; chez les ARN structuraux, cette régulation implique aussi la modification de nucléotides. Ces processus permettent de réguler l'abondance, la stabilité et la traduction des ARN (Motorin & Helm, 2011, Rorbach & Minczuk, 2012). Parmi les intervenants qui régulent la traduction dans la mitochondrie on trouve les protéines PPRs. Ces protéines se lient probablement au 5' UTR de l'ARN pour effectuer cette régulation (Xu*, et al.*, 2004, Davies*, et al.*, 2009, Rackham*, et al.*, 2009, Rackham & Filipovska, 2012, Rackham*, et al.*, 2012, Ruzzenente*, et al.*, 2012).

Dans la mitochondrie de *D. papillatum,* la régulation de l'expression des gènes semble avoir lieu principalement au niveau post-transcriptionnel. De toute évidence, l'épissage en *trans* et l'édition sont essentiels à la production de transcrits chez les diplonémides, le second jouant le rôle de signal pour le premier. Ces deux processus

semblent être très précis en ce sens où il n'y a ni déviation dans l'ordre des modules assemblés, ni dans le nombre de nucléotides ajoutés aux ARNm par l'édition (l'édition du transcrit de *rnl* étant une exception, (Figure 28). Ce type de régulation est probablement plus rapide que celle agissant au niveau de la transcription.

7 Chronologie des processus post-transcriptionnels dans la mitochondrie

On a peu de connaissances sur la séquence et la coordination des processus de maturation des transcrits (Bonen, 2011). Par exemple des études dans la mitochondrie des trypanosomes n`ont pas permis de définir clairement l'ordre des différentes étapes de maturation post-transcriptionnelle (Aphasizhev & Aphasizheva, 2011).

Dans la mitochondrie des plantes, l'épissage semble précéder dans certains cas l'édition et même créer des sites d`édition (Bonen, 2011). Aussi l`ordre inverse semble plus fréquent (Burger*, et al.*, 2009, Castandet*, et al.*, 2010, Bonen, 2011, Farre*, et al.*, 2012). Dans ces cas, tout comme dans la mitochondrie de *D. papillatum*, l'édition précède l'épissage (Kiethega*, et al.*, 2013).

8 Évolution des gènes mitochondriaux fragmentés et édités chez les diplonémides

Nous avons montré que la fragmentation des gènes mitochondriaux et l'édition des transcrits existaient chez toutes les espèces de diplonémides étudiées. Il ne s'agit donc pas de caractères acquis récemment par *D. papillatum*, mais plutôt que ces caractères existaient déjà dans l'ancêtre commun des diplonémides. Ceci pose la question s'il y aurait un avantage à ce que ces phénomènes aient été maintenus durant des millions d'années. Si ils ont été conservés, ils ne peuvent pas être nuisibles et il pourrait s'agir de sélection neutre (Lynch, 2007).

La fragmentation et l'édition des gènes mitochondriaux sont des caractères partagés de façon variable par les deux autres groupes d'euglénozoaires. Chez les kinétoplastides l'édition est prédominante mais aucun gène mitochondrial fragmenté n'a été trouvé. En contraste chez *E. gracilis* le seul euglénide dont l'ADNmt a été étudié, la fragmentation des gènes est observée mais pas l'édition (Spencer & Gray, 2011). Trois scénarios pourraient expliquer l'émergence de ces deux caractères au cours de l'évolution. Soit la fragmentation et l'édition des gènes mitochondriaux existaient avant la séparation des trois groupes d'euglénozoaires et l'une ou l'autre a été perdue chez les euglénides et les kinétoplastides (Figure 29A) (Marande*, et al.*, 2005, Flegontov*, et al.*, 2011, Vlcek*, et al.*, 2011). Soit, l'édition a été acquise par l'ancêtre commun des diplonémides et des kinétoplastides après la séparation d'avec les euglénides (Figure 29B). Le troisième scénario serait que la fragmentation et l'édition ont été acquises de façon indépendante chez les euglénozoaires (Figure 29C). Pour déterminer le ou les scénarios d`évolution, il faudrait plus d'informations moléculaires sur les euglénides et les kinétoplastides divergeant tôt dans l'évolution.

Tableau IX. Introns impliqués dans l'épissage en *trans* mitochondrial.

Groupes	l'épissage en *trans* de gènes mitochondriaux		
	Types d'introns		Autres
Animaux	Intron I [a,b]	Intron II [h]	
Plantae	Intron I [c,d,g]	Intron II [i,j]	
Champignons	Intron I [e,f]	Intron II [k]	
Alvéolates	/	/	Inconnu [l]
Euglénozoaires	/	/	Facteurs *trans* [m]

[a-g] Introns de groupe I impliqués dans l'épissage en *trans* (Sinniger*, et al.*, 2007, Burger*, et al.*, 2009, Grewe*, et al.*, 2009, Pombert & Keeling, 2010, Hecht*, et al.*, 2011, Nadimi*, et al.*, 2012, Nishimura*, et al.*, 2012, Pelin*, et al.*, 2012). [h-k] Introns de groupe II impliqués dans l'épissage en *trans* (Bonen, 2008, Valles*, et al.*, 2008, Glanz & Kuck, 2009). [l, m] Épissage en *trans* n'impliquant pas des introns (Waller & Jackson, 2009, Kiethega*, et al.*, 2011).

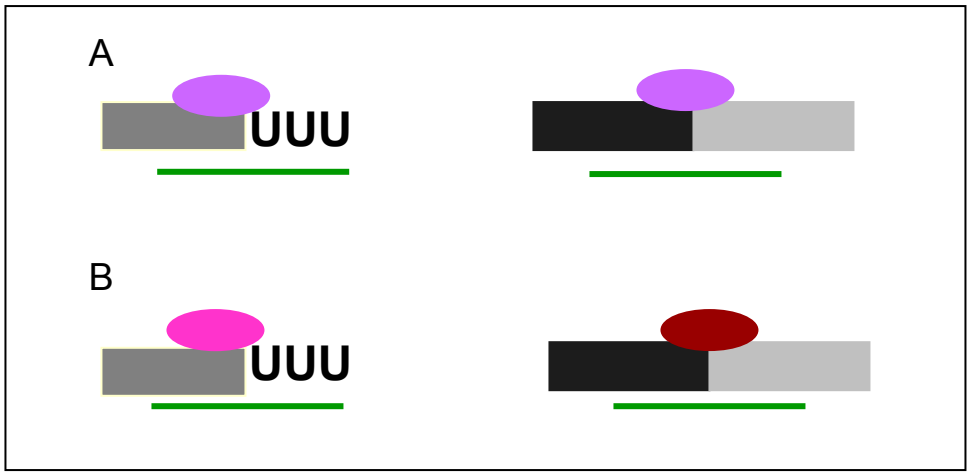


Figure 26. Hypothèses sur les facteurs *trans* impliqués dans l'édition et l'épissage en *trans*. A. Le même facteur protéique en violet est impliqué dans l'édition et l'épissage en *trans*. B. Des facteurs protéiques différents (PPRs) sont impliqués dans l'édition (rose) et l'épissage en *trans* (mauve), un facteur protéique pour chaque jonction; l'ARNg qui sert de matrice est en vert.

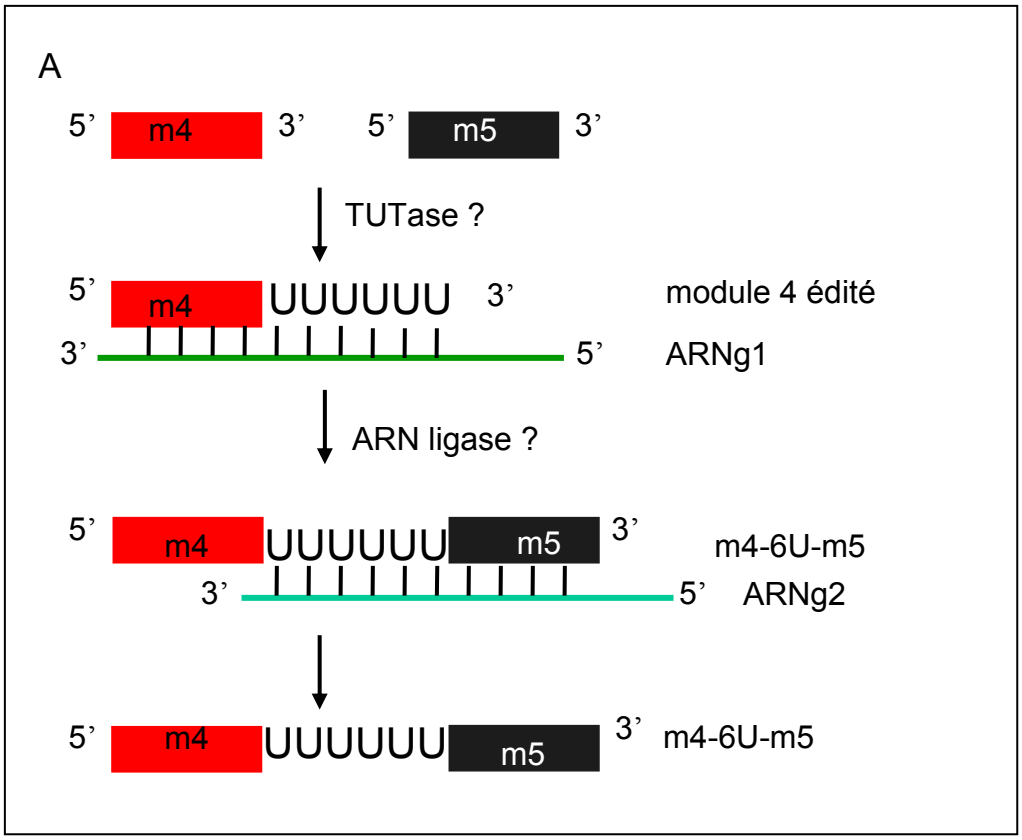


Figure 27 A. Mécanisme d’édition et d’épissage en *trans* de *cox1* dans la mitochondrie de *D. papillatum*. Hypothèse 1. Les modules 4 et 5 sont transcrits et le module 4 avec une extrémité 3’ excisée est édité. Cette édition implique un ARNg1 comme matrice et une hypothétique TUTase pour ajouter les six Us. Le module 4 édité est ensuite lié au module 5 avec une extrémité 5’ excisée grâce à une ARN ligase hypothétique. Un ARNg2 sert de matrice pour l’épissage en *trans*.

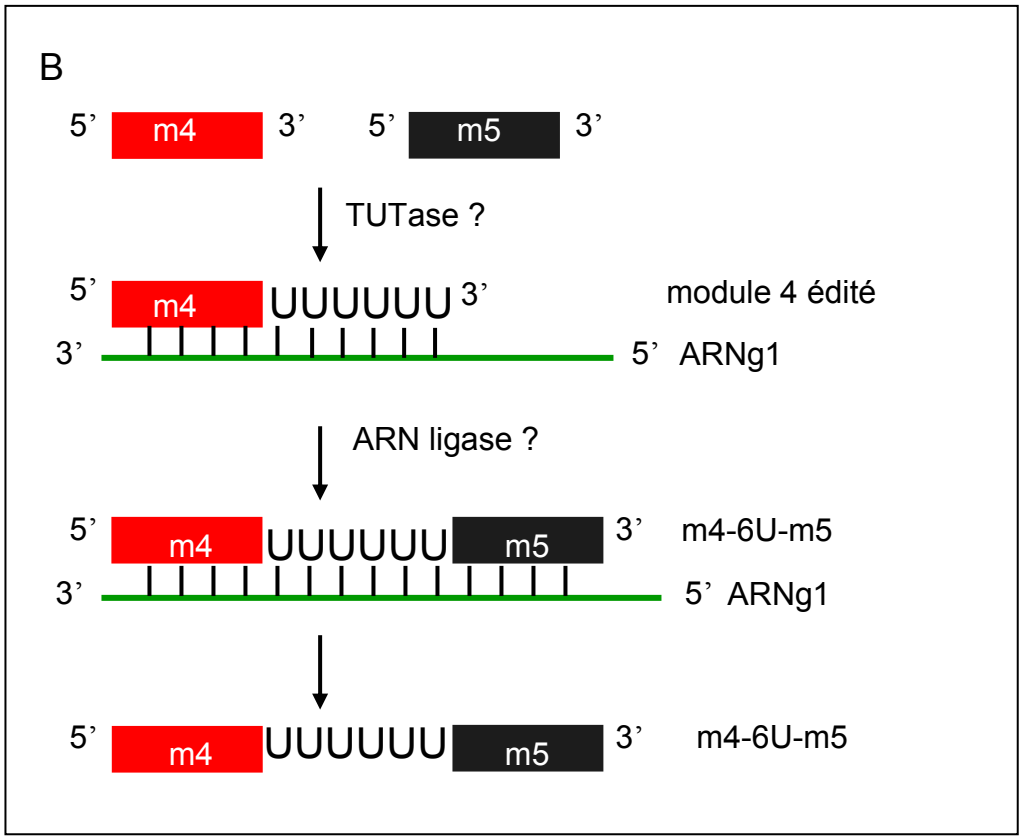


Figure 27 B. Mécanisme d’édition et d’épissage en *trans* de *cox1* dans la mitochondrie de *D. papillatum*. Hypothèse 2 Les modules 4 et 5 sont transcrits et le module 4 avec une extrémité 3’ excisée est édité. Cette édition implique un ARNg 1 comme matrice et une hypothétique TUTase pour ajouter les six U. Le module 4 édité est ensuite lié au module 5 avec une extrémité 5’ excisée grâce à une hypothétique ARN ligase. Le même ARNg 1 sert de matrice pour l’épissage en *trans*.

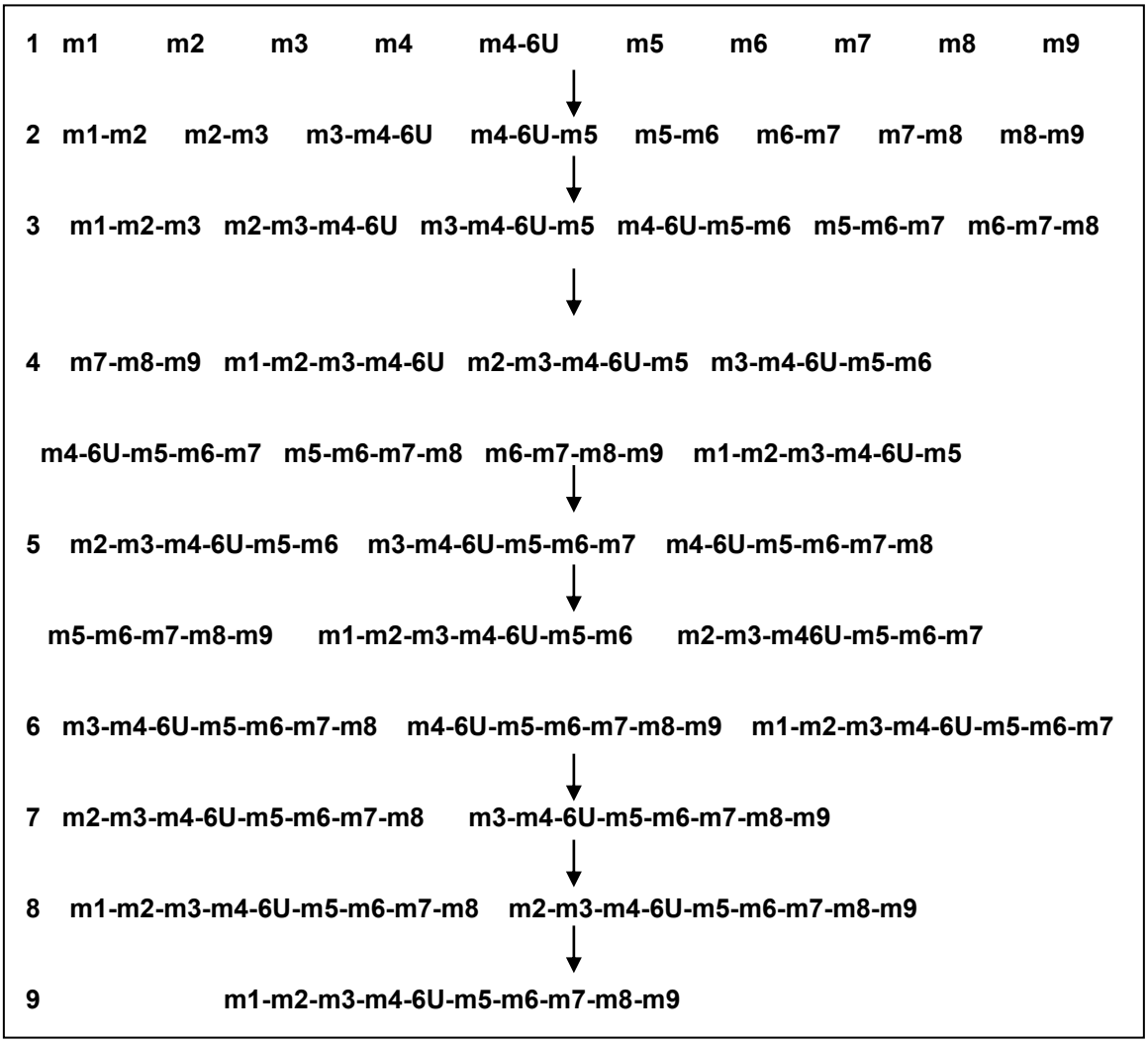


Figure 28. Les intermédiaires possibles de l'épissage en *trans* de *cox1*. À partir des neufs modules de *cox1* épissés (ligne 1), 36 intermédiaires possibles de l'épissage en *trans* peuvent être produits par la mitochondrie. Les lignes 2, 3, 4, 5, 6, 7, 8, 9 montrent des intermédiaires de l'épissage en *trans* formés de 2-9 modules.

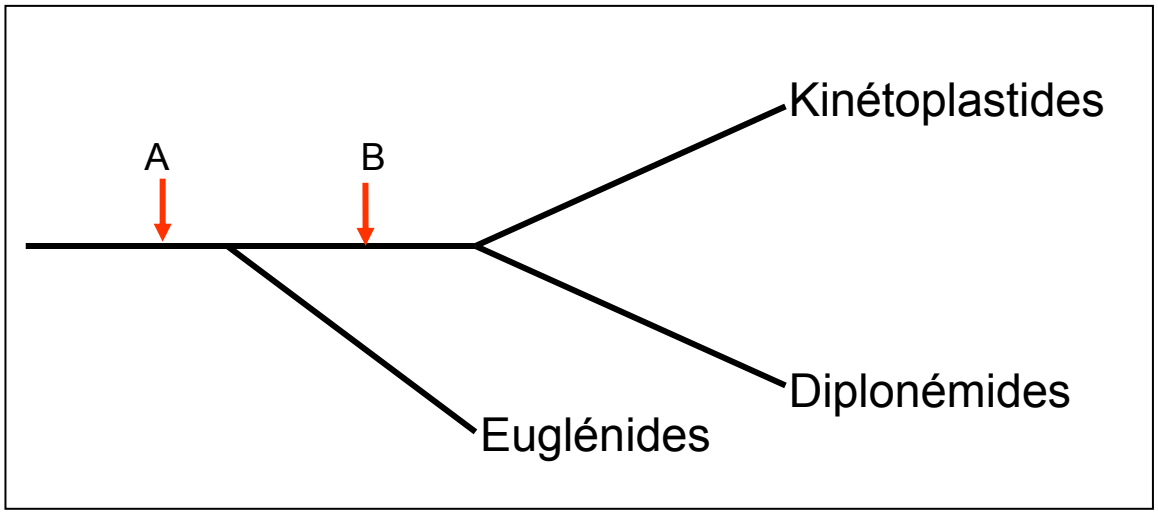


Figure 29. Evolution de la fragmentation et l'édition des gènes mitochondriaux au sein des euglénozoaires. A. Hypothèse 1. L'édition et la fragmentation existaient avant la séparation des trois groupes d'euglénozoaires (flèche rouge). B. Hypothèse 2. L'édition et la fragmentation existaient avant la séparation des kinétoplastides et diplonémides (flèche rouge).

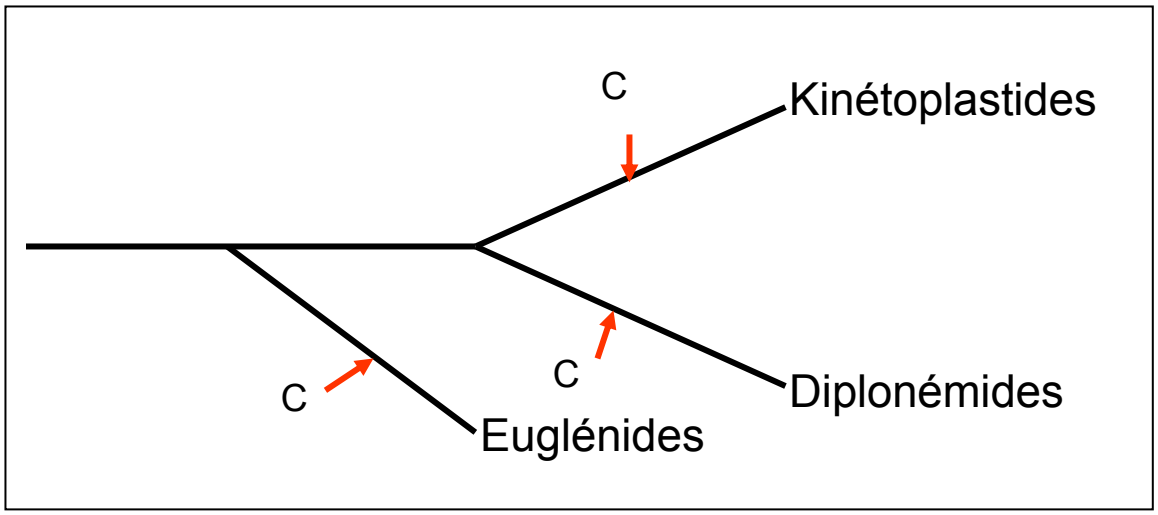


Figure 29. Evolution de la fragmentation et l'édition des gènes mitochondriaux au sein des euglénozoaires. C. Hypothèse 3. L'édition et la fragmentation sont acquises de façon indépendante par chaque groupe d'euglénozoaires (flèche rouge).

Chapitre 5. Conclusion et perspectives

L'étude des transcrits des gènes fragmentés dans la mitochondrie de quatre espèces de diplonémides a permis de caractériser l'épissage en *trans* et l'édition des ARN et de conclure que ces processus post-transcriptionnels sont communs au groupe entier.

La description des étapes de maturation des ARN mitochondriaux chez *D. papillatum*, montre que la maturation des extrémités 5' et 3', l`epissage en *trans* et/ou l'édition sont coordonnés et précis.

Un seul type d'édition a été caractérisé et consiste en l'ajout d'uridines à l'extrémité 3' d'un module. Cette édition ne se limite pas au gène *cox1*; nous avons également découvert quatre autres gènes mitochondriaux édités chez *D. papillatum*.

Nos recherches d'ARN antisens chez *D. papillatum* suggèrent que ces ARN pourraient servir de matrices pour l'épissage en *trans* et l'édition des pré-ARNm et des pré-ARNr. Toutefois nous ne pouvons exclure que d'autres molécules (protéines ou ADN) puissent servir de matrices. Nos résultats suscitent de nouvelles interrogations.

- Quelle est la structure exacte et la fonction des petits ARN antisens des diplonémides et comment sont-ils produits?
- Quels sont les composantes des machineries effectuant l'édition et l'épissage en *trans*?

Les nouvelles perspectives sont en cours d'étude au laboratoire et sont décrits dans les paragraphes suivants.

1 Travaux en cours

1.1 Analyse du séquençage à haut débit du transcriptome mitochondrial de *D. papillatum*

Pour pouvoir répondre aux questions concernant l'édition des gènes mitochondriaux de *D. papillatum* et pour caractériser d'avantage les petits ARN antisens, nous avons séquencé le transcriptome de cet organisme en utilisant une méthode récente et puissante : le RNA-Seq. Le principe du RNA-Seq repose sur le séquençage à haut débit d'une banque d'ADNc. Grâce à des millions de lectures d'ADNc, le RNA-Seq permet de cartographier, répertorier et quantifier le transcriptome entier d'un organisme. On peut ainsi cataloguer les différentes espèces d'ARN, identifier au nucléotide près les extrémités 3' et 5' des ARN, les sites d'initiation de la transcription, les processus post-transcriptionnels tels que l'édition et l'épissage (Costa*, et al.*, 2010, Garber*, et al.*, 2011, Ozsolak & Milos, 2011). Les objectifs de nos expériences RNA-Seq sont multiples : Valider les régions 5' des gènes *atp6*, *nad1*, *nad4* et *nad5*, les extrémités 5'UTR et les jonctions des modules; identifier les gènes mitochondriaux non annotés et trouver d'autres gènes édités et d'autres types d'éditions et répertorier d'une façon globale les petits ARN antisens.

Le séquencage RNA-Seq a été effectué durant la rédaction de cette thèse, avec le Dr Valach (stagiaire postdoctoral) dès que notre laboratoire a réussi à mettre au point l'isolation des mitochondries relativement pures de *D. papillatum*. L'analyse de cette immense banque de données constitue un défi.

1.2 Identification *in silico* des protéines impliquées dans l'épissage en *trans* et l'édition dans la mitochondrie de *D. papillatum*

La question de l'existence d'un éditosome mitochondrial chez *Diplonema*, similaire à celui des kinétoplastides, est également en cours d'étude dans notre laboratoire. Dans le cadre de son doctorat en bioinformatique, Sandrine Moreira effectue la recherche *in sillico* de protéines qui sont des homologues des composantes de l'éditosome chez les kinétoplastides. Nous avons postulé qu'au moins une ARN ligase et une TUTase sont impliquées dans l'épissage en *trans* et l'édition de *D. papillatum*. Sandrine Moreira a identifié un gène candidat d'ARN ligase de *D. papillatum* dans le

génome nucléaire et notre laboratoire tente présentement de démontrer que ce gène est traduit en une protéine importée dans la mitochondrie.

1.3 Identification de l'activité ARN ligase dans la mitochondrie de *D. papillatum*

Tel que mentionné précédemment, l'épissage en *trans* chez *D. papillatum* doit impliquer une activité ARN ligase. La Dr Breton dans le cadre de son stage postdoctoral a tenté d'identifier une telle activité enzymatique. Une caractéristique des ARN ligases est qu'elles lient l'ATP de façon covalente. Ainsi des activités d'adénylation de protéines ont été testées de même que la réaction de ligation catalysée par des fractions mitochondriales de *D. papillatum.* Les résultats préliminaires montrent la présence d'une activité ARN ligase mais la localisation subcellulaire n'est pas encore connue.

1.4 Isolation des complexes mitochondriaux

En nous basant sur ce qui a été décrit chez les kinétoplastides, nous supposons que les complexes effectuant l'édition et l'épissage en *trans* se trouvent dans la matrice plutôt que dans la membrane mitochondriale. La composition de ces complexes matriciels chez *D. papillatum* est en train d'être étudiée par spectrométrie de masse. Notre but est de trouver les complexes contenant une ARN ligase, une TUTase et/ou des protéines PPRs, donc des candidats potentiels du trans-splicéosome et de l'éditosome. Le principal défi de ce projet est d'obtenir une quantité suffisante de complexes matriciels pour être soumise à la spectrométrie de masse.

2 Projets futurs

Les résultats obtenus dans le cadre de mes recherches doctorales constituent une base solide pour avancer nos connaissances des mécanismes moléculaires impliqués dans les processus post-transcriptionnels dans la mitochondrie des diplonémides. Plusieurs lignes d'expérimentation sont proposées comme prochaines étapes.

2.1 Développer des tests fonctionnels *in vitro* de l'épissage en *trans* et d'édition de gènes mitochondriaux de *D. papillatum*

La question qui se pose est de savoir si les petits ARN antisens mitochondriaux interviennent dans les deux processus post-transcriptionnels. L'implication de ces ARN endogènes dans l'épissage en *trans* et l'édition pourrait être mise en évidence par des tests *in vitro*.

2.1.1 Test *in vitro* de l'épissage en *trans*

Il faudrait développer un test qui permettrait de déterminer si un extrait mitochondrial a la capacité de lier deux modules de *cox1* synthétiques, marqués par radioactivité. Le produit de la liaison, dont la taille doit être la somme de celles des deux modules marqués, serait séparé par migration sur gel d'acrylamide et visualisé par radiographie. La comparaison des tests de l'épissage en *trans* avec des extraits mitochondriaux traités et non traités avec l'ARNase nous indiquerait si des ARN endogènes sont impliqués dans l'épissage en *trans* de *cox1*. L'expérience pourrait être répétée avec des extraits mitochondriaux enrichis en petits ARN endogènes afin de confirmer le rôle de ces ARN dans l'épissage en *trans*.

Finalement si l'analyse des données RNA-Seq permet de caractériser la structure des ARN antisens, on pourrait utiliser des ARN antisens synthétiques pour diriger la liaison de modules qui ne sont pas joints *in vivo*.

2.1.2 Test in *vitro* d'édition

Il faudra aussi développer un test d'édition *in vitro* qui permettra de mesurer si des extraits mitochondriaux sont capables d'ajouter des uridines (marquées par radioactivité), au bout 3' d'un module édité. Les ARNr étant abondants, la détection de l'édition du module 1 du gène *rnl* sera privilégiée. Une fois de plus, la comparaison de l'édition avec des extraits mitochondriaux traités ou non avec l'ARNase démontrerait si

des ARN endogènes sont impliqués dans l'édition. La même approche que dans 2.1.1 pourrait être utilisée pour provoquer l'édition des modules qui ne sont pas édités in *vivo*.

2.2 Manipulation génétique de *D. papillatum*

La mise au point de méthodes de transfection stable chez *D. papillatum* faciliterait des études fonctionnelles sur les protéines impliquées dans les processus post-transcriptionnels mitochondriaux. La génétique inverse par knockdown de gènes codant pour des protéines cibles permettra de déterminer leur fonction comme cela a été fait chez les trypanosomes pour étudier l'éditosome (Clayton, 1999). Le rôle de la ARN ligase et des PPRs trouvés dans le génome nucléaire de *D. papillatum,* pourrait ainsi être investigué.

2.3 L'approche ARNi chez *D. papillatum*

Les ARNi sont des petits ARN régulateurs eucaryotes d`environ 20 nt incluant les miRNA (microRNA) et les siRNA (small interfering RNA) et les piwiRNA (piwi interacting RNA) (Siomi*, et al.*, 2011). L'interférence ARN est un mécanisme par lequel des ARNi (ARN interférents) interagissent avec des ARNm cibles au sein d'un complexe RISC (RNAi Induced Silencing Complex) permettant ainsi de réguler l'expression d'un gène.

L'interférence ARN, si elle existe naturellement dans l'organisme, peut être exploitée en utilisant des ARN synthétiques. Ceci est une méthode de choix pour étudier la fonction des gènes chez les kinétoplastides et ainsi identifier les composants le l'éditosome. Pour investiguer si de l'interférence ARN existe chez *D. papillatum,* il faudrait chercher des protéines Dicer et argonautes (impliquées dans l'interférence ARN) dans les données génomiques nucléaires disponibles. Il s'agirait de rechercher les domaines protéiques conservées tels que PAZ et PIWI, RNAse III ou de fixation du double brin ARN caractéristiques des argonautes et des Dicer (Batista & Marques, 2011).

2.4 Les processus post-transcriptionnels dans la mitochondrie des euglénozoaires

Nos travaux sur les processus post-transcriptionnels dans la mitochondrie des diplonémides suscitent aussi de nouvelles questions sur les euglénozoaires. Si de nombreuses études ont permis de comprendre l'éditosome chez les kinétoplastides, ces processus sont complètement inconnus chez les euglénides. Dans le passé, notre laboratoire a étudié les génomes mitochondriaux de deux membres de ce groupe, soient *Petalomonas cantuscygni* et *Peranema trichophorum*. Les données partielles indiquent que l'ADNmt de *P. cantuscygni* est constitué d'un seul chromosome circulaire de 40 kb portant des gènes fragmentés. Le génome mitochondrial de *P. tricophorum* apparait plutôt formé de nombreuses molécules linéaires de différentes tailles encodant des gènes contigus (Roy, 2006, Roy*, et al.*, 2007). Une analyse du transcriptome de ces espèces par RNA-Seq permettrait de lever le voile sur les gènes mitochondriaux des euglénides et finalement retracer l'évolution de l'architecture du génome mitochondrial et la structure des gènes chez les euglénozoaires.

Bibliographie

Adams KL & Palmer JD (2003) Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. *Molecular Phylogenetics and Evolution* **29**: 380-395.

Adl SM, Simpson AG, Lane CE*, et al.* (2012) The revised classification of eukaryotes. *Journal of Eukaryotic Microbiology* **59**: 429-493.

Alfonzo JD, Blanc V, Estevez AM, Rubio MA & Simpson L (1999) C to U editing of the anticodon of imported mitochondrial tRNA(Trp) allows decoding of the UGA stop codon in *Leishmania tarentolae*. *The EMBO Journal* **18**: 7056-7062.

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

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W & Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**: 3389-3402.

Alverson AJ, Wei X, Rice DW, Stern DB, Barry K & Palmer JD (2010) Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). *Molecular Biology and Evolution* **27**: 1436-1448.

Ammerman ML, Presnyak V, Fisk JC, Foda BM & Read LK (2010) TbRGG2 facilitates kinetoplastid RNA editing initiation and progression past intrinsic pause sites. *RNA* **16**: 2239-2251.

Anderson S, Bankier AT, Barrell BG*, et al.* (1981) Sequence and organization of the human mitochondrial genome. *Nature* **290**: 457-465.

Aphasizhev R & Simpson L (2001) Isolation and characterization of a U-specific 3'-5'-exonuclease from mitochondria of *Leishmania tarentolae*. *The Journal of Biological Chemistry* **276**: 21280-21284.

Aphasizhev R & Aphasizheva I (2008) Terminal RNA uridylyltransferases of trypanosomes. *Biochimica Biophysica Acta* **1779**: 270-280.

Aphasizhev R & Aphasizheva I (2011) Mitochondrial RNA processing in trypanosomes. *Research in Microbiology* **162**: 655-663.

Aphasizhev R & Aphasizheva I (2011) Uridine insertion/deletion editing in trypanosomes: a playground for RNA-guided information transfer. *Wiley interdisciplinary reviews. RNA* **2**: 669-685.

Aphasizhev R, Sbicego S, Peris M*, et al.* (2002) Trypanosome mitochondrial 3' terminal uridylyl transferase (TUTase): the key enzyme in U-insertion/deletion RNA editing. *Cell* **108**: 637-648.

Aphasizhev R, Aphasizheva I, Nelson RE*, et al.* (2003) Isolation of a U-insertion/deletion editing complex from Leishmania tarentolae mitochondria. *The EMBO Journal* **22**: 913-924.

Aphasizheva I, Ringpis GE, Weng J, Gershon PD, Lathrop RH & Aphasizhev R (2009) Novel TUTase associates with an editosome-like complex in mitochondria of *Trypanosoma brucei*. *RNA* **15**: 1322-1337.

Asin-Cayuela J & Gustafsson CM (2007) Mitochondrial transcription and its regulation in mammalian cells. *Trends in Biochemical Sciences* **32**: 111-117.

Barbrook AC, Howe CJ, Kurniawan DP & Tarr SJ (2010) Organization and expression of organellar genomes. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences* **365**: 785-797.

Batista TM & Marques JT (2011) RNAi pathways in parasitic protists and worms. *Journal of Proteomics* **74**: 1504-1514.

Baurain D, Brinkmann H, Petersen J*, et al.* (2010) Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. *Molecular Biology and Evolution* **27**: 1698-1709.

Beck CF & Warren RA (1988) Divergent promoters, a common form of gene organization. *Microbiological Reviews* **52**: 318-326.

Bendich AJ (1993) Reaching for the ring: the study of mitochondrial genome structure. *Current Genetics* **24**: 279-290.

Benne R, Van den Burg J, Brakenhoff JP, Sloof P, Van Boom JH & Tromp MC (1986) Major transcript of the frameshifted coxII gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* **46**: 819-826.

Blanc V & Davidson NO (2003) C-to-U RNA editing: mechanisms leading to genetic diversity. *The Journal of Biological Chemistry* **278**: 1395-1398.

Blanc V & Davidson NO (2010) APOBEC-1-mediated RNA editing. *Wiley Interdisciplinary Reviews. Systems Biology and Medicine* **2**: 594-602.

Blom D, de Haan A, van den Berg M, Sloof P, Jirku M, Lukes J & Benne R (1998) RNA editing in the free-living bodonid *Bodo saltans*. *Nucleic Acids Research* **26**: 1205-1213.

Blum B & Simpson L (1990) Guide RNAs in kinetoplastid mitochondria have a nonencoded 3' oligo(U) tail involved in recognition of the preedited region. *Cell* **62**: 391-397.

Blum B, Bakalara N & Simpson L (1990) A model for RNA editing in kinetoplastid mitochondria: "guide" RNA molecules transcribed from maxicircle DNA provide the edited information. *Cell* **60**: 189-198.

Bolender N, Sickmann A, Wagner R, Meisinger C & Pfanner N (2008) Multiple pathways for sorting mitochondrial precursor proteins. *EMBO reports* **9**: 42-49.

Bonawitz ND, Clayton DA & Shadel GS (2006) Initiation and beyond: multiple functions of the human mitochondrial transcription machinery. *Molecular Cell* **24**: 813-825.

Bonen L (1993) Trans-splicing of pre-mRNA in plants, animals, and protists. *FASEB Journal* **7**: 40-46.

Bonen L (2008) Cis- and trans-splicing of group II introns in plant mitochondria. *Mitochondrion* **8**: 26-34.

Bonen L (2011) RNA Splicing in Plant Mitochondria. *Plant mitochondria,*ed.^eds.), p.^pp. 131-155.

Bonen L (2012) Evolution of mitochondrial intron in plants and photosynthetic microbes. *Mitochondrial genome evolution,* Vol. 63 (Maréchal-Drouard L, ed.^eds.), p.^pp. 155-186.

Bonen L & Vogel J (2001) The ins and outs of group II introns. *Trends in Genetics* **17**: 322-331.

Breathnach R & Chambon P (1981) Organization and expression of eucaryotic split genes coding for proteins. *Annual Review of Biochemistry* **50**: 349-383.

Brecht M, Niemann M, Schluter E, Muller UF, Stuart K & Goringer HU (2005) TbMP42, a protein component of the RNA editing complex in African trypanosomes, has endo-exoribonuclease activity. *Molecular Cell* **17**: 621-630.

Breglia SA, Slamovits CH & Leander BS (2007) Phylogeny of phagotrophic euglenids (Euglenozoa) as inferred from *hsp90* gene sequences. *Journal of Eukaryotic Microbiology* **54**: 86-92.

Bundschuh R, Altmuller J, Becker C, Nurnberg P & Gott JM (2011) Complete characterization of the edited transcriptome of the mitochondrion of *Physarum polycephalum* using deep sequencing of RNA. *Nucleic Acids Research* **39**: 6044-6055.

Burger G & Nedelcu A (2011) "Mitochondrial genomes of algae."

Genomics of chloroplasts and mitochondria, Series: Advances in photosynthesis and respiration Vol. 35 (R. Bock VK, Springer-Verlag Heidelberg, Germany, ed.^eds.), p.^pp. 127-157.

Burger G, Gray MW & Lang BF (2003) Mitochondrial genomes: anything goes. *Trends in Genetics* **19**: 709-716.

Burger G, Roy J & Teijeiro S (2008) [When genes break up: the mitochondrial genomes of diplonemids]. *Medecines Sciences(Paris)* **24**: 703-705.

Burger G, Jackson CJ & Waller RF (2011) "*Unusual mitochondrial genomes and genes*." *Organelle Genetics 2012, Part 2 : evolution of organelle genomes and gene expression* (C. Bullerwell eS-VH, Germany, ed.^eds.), p.41-77.

Burger G, Yan Y, Javadi P & Lang BF (2009) Group I-intron trans-splicing and mRNA editing in the mitochondria of placozoan animals. *Trends in Genetics* **25**: 381-386.

Burki F & Pawlowski J (2006) Monophyly of Rhizaria and multigene phylogeny of unicellular bikonts. *Molecular Biology and Evolution* **23**: 1922-1930.

Busse I & Preisfeld A (2002) Phylogenetic position of *Rhynchopus* sp. and *Diplonema ambulator* as indicated by analyses of euglenozoan small subunit ribosomal DNA. *Gene* **284**: 83-91.

Byrne EM, Connell GJ & Simpson L (1996) Guide RNA-directed uridine insertion RNA editing in vitro. *EMBO Journal* **15**: 6758-6765.

Carnes J, Soares CZ, Wickham C & Stuart K (2011) Endonuclease associations with three distinct editosomes in *Trypanosoma brucei*. *The Journal of Biological Chemistry* **286**: 19320-19330.

Carnes J, Trotter JR, Ernst NL, Steinberg A & Stuart K (2005) An essential RNase III insertion editing endonuclease in *Trypanosoma brucei*. *Proceedings of National Academy of Sciences of United States of America* **102**: 16614-16619.

Carnes J, Trotter JR, Peltan A, Fleck M & Stuart K (2008) RNA editing in *Trypanosoma brucei* requires three different editosomes. *Molecular and Cellular Biology* **28**: 122-130.

Castandet B & Araya A (2011) RNA editing in plant organelles. Why make it easy? *Biochemistry. Biokhimiia* **76**: 924-931.

Castandet B, Choury D, Begu D, Jordana X & Araya A (2010) Intron RNA editing is essential for splicing in plant mitochondria. *Nucleic Acids Research* **38**: 7112-7121.

Cavalier-Smith T (1998) A revised six-kingdom system of life. *Biology Reviews of the Cambridge Philosophical Society* **73**: 203-266.

Cavalier-Smith T (2002) The phagotrophic origin of eukaryotes and phylogenetic classification of Protozoa. *International Journal of Systematic and Evolutionary Microbiology* **52**: 297-354.

Chacinska A, Koehler CM, Milenkovic D, Lithgow T & Pfanner N (2009) Importing mitochondrial proteins: machineries and mechanisms. *Cell* **138**: 628-644.

Chang JH & Tong L (2012) Mitochondrial poly(A) polymerase and polyadenylation. *Biochimica et Biophysica Acta* **1819**: 992-997.

Chaput H, Wang Y & Morse D (2002) Polyadenylated transcripts containing random gene fragments are expressed in dinoflagellate mitochondria. *Protist* **153**: 111-122.

Chateigner-Boutin AL & Small I (2010) Plant RNA editing. *RNA Biology* **7**: 213-219.

Chateigner-Boutin AL & Small I (2011) Organellar RNA editing. *Wiley interdisciplinary reviews. RNA* **2**: 493-506.

Chateigner-Boutin AL, Colas des Francs-Small C, Fujii S, Okuda K, Tanz SK & Small I (2013) The E domains of pentatricopeptide repeat proteins from different organelles are not functionally equivalent for RNA editing. *Plant Journal*.

Chen J, Rauch CA, White JH, Englund PT & Cozzarelli NR (1995) The topology of the kinetoplast DNA network. *Cell* **80**: 61-69.

Cheng YW & Gott JM (2000) Transcription and RNA editing in a soluble in vitro system from Physarum mitochondria. *Nucleic Acids Research* **28**: 3695-3701.

Clayton CE (1999) Genetic manipulation of kinetoplastida. *Parasitology today* **15**: 372-378.

Costa V, Angelini C, De Feis I & Ciccodicola A (2010) Uncovering the complexity of transcriptomes with RNA-Seq. *Journal of Biomedicine & Biotechnology* **2010**: 853916.

Covello PS & Gray MW (1989) RNA editing in plant mitochondria. *Nature* **341**: 662-666.

Crooks GE, Hon G, Chandonia JM & Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Research* **14**: 1188-1190.

Cruz-Reyes J, Zhelonkina AG, Huang CE & Sollner-Webb B (2002) Distinct functions of two RNA ligases in active *Trypanosoma brucei* RNA editing complexes. *Molecular and Cellular Biology* **22**: 4652-4660.

Davies SM, Rackham O, Shearwood AM, Hamilton KL, Narsai R, Whelan J & Filipovska A (2009) Pentatricopeptide repeat domain protein 3 associates with the mitochondrial small ribosomal subunit and regulates translation. *FEBS letters* **583**: 1853-1858.

Driscoll DM, Wynne JK, Wallis SC & Scott J (1989) An in vitro system for the editing of apolipoprotein B mRNA. *Cell* **58**: 519-525.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**: 1792-1797.

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**: 113.

Edgell DR, Chalamcharla VR & Belfort M (2011) Learning to live together: mutualism between self-splicing introns and their hosts. *BMC Biology* **9**: 22.

Embley TM, van der Giezen M, Horner DS, Dyal PL, Bell S & Foster PG (2003) Hydrogenosomes, mitochondria and early eukaryotic evolution. *IUBMB Life* **55**: 387-395.

Ernst NL, Panicucci B, Carnes J & Stuart K (2009) Differential functions of two editosome exoUases in *Trypanosoma brucei*. *RNA* **15**: 947-957.

Ernst NL, Panicucci B, Igo RP, Jr., Panigrahi AK, Salavati R & Stuart K (2003) TbMP57 is a 3' terminal uridylyl transferase (TUTase) of the *Trypanosoma brucei* editosome. *Molecular Cell* **11**: 1525-1536.

Estevez AM & Simpson L (1999) Uridine insertion/deletion RNA editing in trypanosome mitochondria--a review. *Gene* **240**: 247-260.

Etheridge RD, Aphasizheva I, Gershon PD & Aphasizhev R (2008) 3' adenylation determines mRNA abundance and monitors completion of RNA editing in *T. brucei* mitochondria. *EMBO Journal* **27**: 1596-1608.

Ewing B & Green P (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research* **8**: 186-194.

Falkenberg M, Larsson NG & Gustafsson CM (2007) DNA replication and transcription in mammalian mitochondria. *Annual Review of Biochemistry* **76**: 679-699.

Farre JC, Aknin C, Araya A & Castandet B (2012) RNA editing in mitochondrial trans-introns is required for splicing. *PLoS One* **7**: e52644.

Feagin JE (1990) RNA editing in kinetoplastid mitochondria. *Journal of Biological Chemistry* **265**: 19373-19376.

Feagin JE, Abraham JM & Stuart K (1988) Extensive editing of the cytochrome c oxidase III transcript in *Trypanosoma brucei*. *Cell* **53**: 413-422.

Feagin JE, Gardner MJ, Williamson DH & Wilson RJ (1991) The putative mitochondrial genome of *Plasmodium falciparum*. *Journal of Protozoology* **38**: 243-245.

Feagin JE, Harrell MI, Lee JC*, et al.* (2012) The fragmented mitochondrial ribosomal RNAs of *Plasmodium falciparum*. *PLoS One* **7**: e38320.

Flegontov P, Gray MW, Burger G & Lukes J (2011) Gene fragmentation: a key to mitochondrial genome evolution in Euglenozoa? *Current genetics* **57**: 225-232.

Foldynova-Trantirkova S, Paris Z, Sturm NR, Campbell DA & Lukes J (2005) The *Trypanosoma brucei* La protein is a candidate poly(U) shield that impacts spliced leader RNA maturation and tRNA intron removal. *International Journal for Parasitology* **35**: 359-366.

Frohman MA, Dush MK & Martin GR (1988) Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. *Proceedings of National Academy of Sciences United States of America* **85**: 8998-9002.

Gagliardi D, Stepien PP, Temperley RJ, Lightowlers RN & Chrzanowska-Lightowlers ZM (2004) Messenger RNA stability in mitochondria: different means to an end. *Trends in Genetics* **20**: 260-267.

Garber M, Grabherr MG, Guttman M & Trapnell C (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods* **8**: 469-477.

Gillespie DE, Salazar NA, Rehkopf DH & Feagin JE (1999) The fragmented mitochondrial ribosomal RNAs of *Plasmodium falciparum* have short A tails. *Nucleic Acids Research* **27**: 2416-2422.

Glanz S & Kuck U (2009) Trans-splicing of organelle introns--a detour to continuous RNAs. *Bioessays* **31**: 921-934.

Golden DE & Hajduk SL (2005) The 3'-untranslated region of cytochrome oxidase II mRNA functions in RNA editing of African trypanosomes exclusively as a cis guide RNA. *RNA* **11**: 29-37.

Golden DE & Hajduk SL (2006) The importance of RNA structure in RNA editing and a potential proofreading mechanism for correct guide RNA:pre-mRNA binary complex formation. *Journal of Molecular Biology* **359**: 585-596.

Gordon D (2003) Viewing and editing assembled sequences using Consed. *Curr Protoc Bioinformatics* **Chapter 11**: Unit 11.12.

Gott JM, Parimi N & Bundschuh R (2005) Discovery of new genes and deletion editing in Physarum mitochondria enabled by a novel algorithm for finding edited mRNAs. *Nucleic Acids Research* **33**: 5063-5072.

Gott JM, Somerlot BH & Gray MW (2010) Two forms of RNA editing are required for tRNA maturation in Physarum mitochondria. *RNA* **16**: 482-488.

Graeber MB & Muller U (1998) Recent developments in the molecular genetics of mitochondrial disorders. *Journal of the Neurological Sciences* **153**: 251-263.

Gray MW (1999) Evolution of organellar genomes. *Current Opinion Genetics Development* **9**: 678-687.

Gray MW (2003) Diversity and evolution of mitochondrial RNA editing systems. *IUBMB Life* **55**: 227-233.

Gray MW, Lang BF & Burger G (2004) Mitochondria of protists. *Annual Reviews Genetics* **38**: 477-524.

Grewe F, Viehoever P, Weisshaar B & Knoop V (2009) A trans-splicing group I intron and tRNA-hyperediting in the mitochondrial genome of the lycophyte *Isoetes engelmannii*. *Nucleic Acids Research* **37**: 5093-5104.

Gualberto JM, Weil JH & Grienenberger JM (1990) Editing of the wheat coxIII transcript: evidence for twelve C to U and one U to C conversions and for sequence similarities around editing sites. *Nucleic Acids Research* **18**: 3771-3776.

Gualberto JM, Lamattina L, Bonnard G, Weil JH & Grienenberger JM (1989) RNA editing in wheat mitochondria results in the conservation of protein sequences. *Nature* **341**: 660-662.

Guo X, Carnes J, Ernst NL, Winkler M & Stuart K (2012) KREPB6, KREPB7, and KREPB8 are important for editing endonuclease function in *Trypanosoma brucei*. *RNA* **18**: 308-320.

Hajduk S & Ochsenreiter T (2010) RNA editing in kinetoplastids. *RNA Biology* **7**: 229-236.

Hajduk SL, Siqueira AM & Vickerman K (1986) Kinetoplast DNA of *Bodo caudatus*: a noncatenated structure. *Molecular and Cellular Biology* **6**: 4372-4378.

Hampl V, Hug L, Leigh JW, Dacks JB, Lang BF, Simpson AG & Roger AJ (2009) Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *Proceedings National Academy of Sciences of Unites States of America* **106**: 3859-3864.

Hans J, Hajduk SL & Madison-Antenucci S (2007) RNA-editing-associated protein 1 null mutant reveals link to mitochondrial RNA stability. *RNA* **13**: 881-889.

Hashimi H, Zikova A, Panigrahi AK, Stuart KD & Lukes J (2008) TbRGG1, an essential protein involved in kinetoplastid RNA metabolism that is associated with a novel multiprotein complex. *RNA* **14**: 970-980.

Hecht J, Grewe F & Knoop V (2011) Extreme RNA editing in coding islands and abundant microsatellites in repeat sequences of *Selaginella moellendorffii* mitochondria: the root of frequent plant mtDNA recombination in early tracheophytes. *Genome Biology and Evolution* **3**: 344-358.

Hedtke B, Borner T & Weihe A (1997) Mitochondrial and chloroplast phage-type RNA polymerases in *Arabidopsis*. *Science* **277**: 809-811.

Heinemann IU, Soll D & Randau L (2010) Transfer RNA processing in archaea: unusual pathways and enzymes. *FEBS letters* **584**: 303-309.

Hernandez A, Madina BR, Ro K, Wohlschlegel JA, Willard B, Kinter MT & Cruz-Reyes J (2010) REH2 RNA helicase in kinetoplastid mitochondria: ribonucleoprotein complexes and essential motifs for unwinding and guide RNA (gRNA) binding. *The Journal of Biological Chemistry* **285**: 1220-1228.

Hiesel R, Wissinger B, Schuster W & Brennicke A (1989) RNA editing in plant mitochondria. *Science* **246**: 1632-1634.

Hoch B, Maier RM, Appel K, Igloi GL & Kossel H (1991) Editing of a chloroplast mRNA by creation of an initiation codon. *Nature* **353**: 178-180.

Horton TL & Landweber LF (2002) Rewriting the information in DNA: RNA editing in kinetoplastids and myxomycetes. *Current Opinion in Microbiology* **5**: 620-626.

Huang Y, Niu B, Gao Y, Fu L & Li W (2010) CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* **26**: 680-682.

Jackson CJ & Waller RF (2013) A widespread and unusual RNA trans-splicing type in dinoflagellate mitochondria. *PLoS One* **8**: e56777.

Jackson CJ, Gornik SG & Waller RF (2012) The mitochondrial genome and transcriptome of the basal dinoflagellate *Hematodinium sp*.: character evolution within the highly derived mitochondrial genomes of dinoflagellates. *Genome Biology and Evolution* **4**: 59-72.

Jackson CJ, Norman JE, Schnare MN, Gray MW, Keeling PJ & Waller RF (2007) Broad genomic and transcriptional analysis reveals a highly derived genome in dinoflagellate mitochondria. *BMC Biology* **5**: 41.

Jepson JE & Reenan RA (2008) RNA editing in regulating gene expression in the brain. *Biochimica et Biophysica Acta* **1779**: 459-470.

Kable ML, Seiwert SD, Heidmann S & Stuart K (1996) RNA editing: a mechanism for gRNA-specified uridylate insertion into precursor mRNA. *Science* **273**: 1189-1195.

Kang X, Rogers K, Gao G, Falick AM, Zhou S & Simpson L (2005) Reconstitution of uridine-deletion precleaved RNA editing with two recombinant enzymes. *Proceedings of the National Academy of Sciences of the United States of America* **102**: 1017-1022.

Kayal E, Bentlage B, Collins AG, Kayal M, Pirro S & Lavrov DV (2012) Evolution of linear mitochondrial genomes in medusozoan cnidarians. *Genome Biology and Evolution* **4**: 1-12.

Keeling PJ, Burger G, Durnford DG*, et al.* (2005) The tree of eukaryotes. *Trends in Ecology and Evolution* **20**: 670-676.

Keiler KC, Shapiro L & Williams KP (2000) tmRNAs that encode proteolysis-inducing tags are found in all known bacterial genomes: A two-piece tmRNA functions in *Caulobacter*. *Proceedings National Academy of Sciences of Unites States of America* **97**: 7778-7783.

Kennell JC & Lambowitz AM (1989) Development of an in vitro transcription system for *Neurospora crassa* mitochondrial DNA and identification of transcription initiation sites. *Molecular and Cellular Biology* **9**: 3603-3613.

Kent ML, Elston RA, Nerad TA & Sawyer TK (1987) An *Isonema*-like flagellate (Protozoa: Mastigophora) infection in larval geoduck clams, *Panope abrupta*. *Journal of Invertebrates Pathology* **50**: 221-229.

Kiethega GN, Turcotte M & Burger G (2011) Evolutionarily conserved *cox1* trans-splicing without cis-motifs. *Molecular Biology and Evolution* **28**: 2425-2428.

Kiethega GN, Yan Y, Turcotte M & Burger G (2013) RNA-level unscrambling of fragmented genes in *Diplonema* mitochondria. *RNA Biology* **10**.

Kim KS, Teixeira SM, Kirchhoff LV & Donelson JE (1994) Transcription and editing of cytochrome oxidase II RNAs in *Trypanosoma cruzi*. *The Journal of Biological Chemistry* **269**: 1206-1211.

Knoop V (2011) When you can't trust the DNA: RNA editing changes transcript sequences. *Cellular and Molecular Life Sciences* **68**: 567-586.

Kobayashi K, Kawabata M, Hisano K, Kazama T, Matsuoka K, Sugita M & Nakamura T (2012) Identification and characterization of the RNA binding surface of the pentatricopeptide repeat protein. *Nucleic Acids Research* **40**: 2712-2723.

Kolakofsky D, Roux L, Garcin D & Ruigrok RW (2005) Paramyxovirus mRNA editing, the "rule of six" and error catastrophe: a hypothesis. *The Journal of General Virology* **86**: 1869-1877.

Koulintchenko M, Konstantinov Y & Dietrich A (2003) Plant mitochondria actively import DNA via the permeability transition pore complex. *EMBO Journal* **22**: 1245-1254.

Kroemer G, Dallaporta B & Resche-Rigon M (1998) The mitochondrial death/life regulator in apoptosis and necrosis. *Annual Review of Physiology* **60**: 619-642.

Kugita M, Yamamoto Y, Fujikawa T, Matsumoto T & Yoshinaga K (2003) RNA editing in hornwort chloroplasts makes more than half the genes functional. *Nucleic Acids Res* **31**: 2417-2423.

Lambowitz AM & Zimmerly S (2011) Group II introns: mobile ribozymes that invade DNA. *Cold Spring Harbor Perspectives in Biology* **3**: a003616.

Lang BF & Burger G (2007) Purification of mitochondrial and plastid DNA. *Nature Protocols* **2**: 652-660.

Lang BF, Gray MW & Burger G (1999) Mitochondrial genome evolution and the origin of eukaryotes. *Annual Review of Genetics* **33**: 351-397.

Lang BF, Laforest MJ & Burger G (2007) Mitochondrial introns: a critical view. *Trends in Genetics* **23**: 119-125.

Lang BF, Burger G, O'Kelly CJ*, et al.* (1997) An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature* **387**: 493-497.

Lange H, Sement FM, Canaday J & Gagliardi D (2009) Polyadenylation-assisted RNA degradation processes in plants. *Trends in Plant Science* **14**: 497-504.

Lavrov DV, Pett W, Voigt O, Worheide G, Forget L, Lang BF & Kayal E (2012) Mitochondrial DNA of *Clathrina clathrus* (Calcarea, Calcinea): six linear chromosomes, fragmented rRNAs, tRNA editing, and a novel genetic code. *Molecular Biology and Evolution*.

Leander BS, Esson HJ & Breglia SA (2007) Macroevolution of complex cytoskeletal systems in euglenids. *Bioessays* **29**: 987-1000.

Lee DY & Clayton DA (1998) Initiation of mitochondrial DNA replication by transcription and R-loop processing. *J Biol Chem* **273**: 30614-30621.

Lerch M, Carnes J, Acestor N, Guo X, Schnaufer A & Stuart K (2012) Editosome accessory factors KREPB9 and KREPB10 in *Trypanosoma brucei*. *Eukaryotic Cell* **11**: 832-843.

Leung SS & Koslowsky DJ (2001) RNA editing in *Trypanosoma brucei*: characterization of gRNA U-tail interactions with partially edited mRNA substrates. *Nucleic Acids Research* **29**: 703-709.

Leung SS & Koslowsky DJ (2001) Interactions of mRNAs and gRNAs involved in trypanosome mitochondrial RNA editing: structure probing of an mRNA bound to its cognate gRNA. *RNA* **7**: 1803-1816.

Li F, Herrera J, Zhou S, Maslov DA & Simpson L (2011) Trypanosome REH1 is an RNA helicase involved with the 3'-5' polarity of multiple gRNA-guided uridine insertion/deletion RNA editing. *Proceedings of the National Academy of Sciences of the United States of America* **108**: 3542-3547.

Liere K, Weihe A & Borner T (2011) The transcription machineries of plant mitochondria and chloroplasts: Composition, function, and regulation. *Journal of Plant Physiology* **168**: 1345-1360.

Lin S (2011) Genomic understanding of dinoflagellates. *Research in Microbiology* **162**: 551-569.

Lin S, Zhang H & Gray MW (2008) *RNA editing in dinoflagellates and its implications for the evolutionary history of the editing machinery.*

Liu B, Liu Y, Motyka SA, Agbo EE & Englund PT (2005) Fellowship of the rings: the replication of kinetoplast DNA. *Trends in Parasitology* **21**: 363-369.

Lukes J, Hashimi H & Zikova A (2005) Unexplained complexity of the mitochondrial genome and transcriptome in kinetoplastid flagellates. *Current Genetics* **48**: 277-299.

Lukes J, Guilbride DL, Votypka J, Zikova A, Benne R & Englund PT (2002) Kinetoplast DNA network: evolution of an improbable structure. *Eukaryotic Cell* **1**: 495-502.

Lukes J, Arts GJ, van den Burg J, de Haan A, Opperdoes F, Sloof P & Benne R (1994) Novel pattern of editing regions in mitochondrial transcripts of the cryptobiid *Trypanoplasma borreli*. *The EMBO Journal* **13**: 5086-5098.

Lukescaron J, Jirku M, Avliyakulov N & Benada O (1998) Pankinetoplast DNA structure in a primitive bodonid flagellate, *Cryptobia helicis*. *The EMBO journal* **17**: 838-846.

Lynch M (2007) *The Origins of Genome Architecture*.

Manning JE, Wolstenholme DR, Ryan RS, Hunter JA & Richards OC (1971) Circular chloroplast DNA from *Euglena gracilis*. *Proceedings National Academy of Sciences of United States of America* **68**: 1169-1173.

Marande W (2007) Structure et expression des gènes mitochondriaux de *Diplonema papillatum*. Thèse, 98p. Université de Montréal, Montreal (Canada).

Marande W & Burger G (2007) Mitochondrial DNA as a genomic jigsaw puzzle. *Science* **318**: 415.

Marande W, Lukes J & Burger G (2005) Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids. *Eukaryotic Cell* **4**: 1137-1146.

Maris C, Masse J, Chester A, Navaratnam N & Allain FH (2005) NMR structure of the apoB mRNA stem-loop and its interaction with the C to U editing APOBEC1 complementary factor. *RNA* **11**: 173-186.

Maslov DA, Yasuhira S & Simpson L (1999) Phylogenetic affinities of *Diplonema* within the Euglenozoa as inferred from the SSU rRNA gene and partial COI protein sequences. *Protist* **150**: 33-42.

Maslov DA, Podlipaev SA & Lukes J (2001) Phylogeny of the kinetoplastida: taxonomic problems and insights into the evolution of parasitism. *Memoria do Instituto Oswaldo Cruz* **96**: 397-402.

Masters BS, Stohl LL & Clayton DA (1987) Yeast mitochondrial RNA polymerase is homologous to those encoded by bacteriophages T3 and T7. *Cell* **51**: 89-99.

McBride HM, Neuspiel M & Wasiak S (2006) Mitochondria: more than just a powerhouse. *Current Biology* **16**: R551-560.

Meisinger C, Sickmann A & Pfanner N (2008) The mitochondrial proteome: from inventory to function. *Cell* **134**: 22-24.

Michel F, Costa M & Westhof E (2009) The ribozyme core of group II introns: a structure in want of partners. *Trends in Biochemical Sciences* **34**: 189-199.

Milbury CA, Lee JC, Cannone JJ, Gaffney PM & Gutell RR (2010) Fragmentation of the large subunit ribosomal RNA gene in oyster mitochondrial genomes. *BMC Genomics* **11**: 485.

Miller ML & Miller DL (2008) Non-DNA-templated addition of nucleotides to the 3' end of RNAs by the mitochondrial RNA polymerase of *Physarum polycephalum*. *Molecular and Cellular Biology* **28**: 5795-5802.

Missel A, Souza AE, Norskau G & Goringer HU (1997) Disruption of a gene encoding a novel mitochondrial DEAD-box protein in *Trypanosoma brucei* affects edited mRNAs. *Molecular and Cellular Biology* **17**: 4895-4903.

Miyazawa S & Jernigan RL (1999) Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins* **34**: 49-68.

Moreira S, Breton S & Burger G (2012) Unscrambling genetic information at the RNA level. *Wiley Interdisciplinary Reviews. RNA* **3**: 213-228.

Motorin Y & Helm M (2011) RNA nucleotide methylation. *Wiley Interdisciplinary Reviews. RNA* **2**: 611-631.

Nadimi M, Beaudet D, Forget L, Hijri M & Lang BF (2012) Group I intron-mediated trans-splicing in mitochondria of *Gigaspora rosea* and a robust phylogenetic affiliation of arbuscular mycorrhizal fungi with Mortierellales. *Molecular Biology and Evolution* **29**: 2199-2210.

Nagaike T, Suzuki T & Ueda T (2008) Polyadenylation in mammalian mitochondria: insights from recent studies. *Biochimica et Biophysica Acta* **1779**: 266-269.

Nakamura T, Yagi Y & Kobayashi K (2012) Mechanistic insight into pentatricopeptide repeat proteins as sequence-specific RNA-binding proteins for organellar RNAs in plants. *Plant & Cell Physiology* **53**: 1171-1179.

Nash EA, Nisbet RE, Barbrook AC & Howe CJ (2008) Dinoflagellates: a mitochondrial genome all at sea. *Trends in Genetics* **24**: 328-335.

Neupert W & Herrmann JM (2007) Translocation of proteins into mitochondria. *Annual Review of Biochemistry* **76**: 723-749.

Nielsen H & Johansen SD (2009) Group I introns: Moving in new directions. *RNA Biology* **6**: 375-383.

Niemann M, Kaibel H, Schluter E, Weitzel K, Brecht M & Goringer HU (2009) Kinetoplastid RNA editing involves a 3' nucleotidyl phosphatase activity. *Nucleic Acids Research* **37**: 1897-1906.

Nishikura K (2010) Functions and regulation of RNA editing by ADAR deaminases. *Annual Review of Biochemistry* **79**: 321-349.

Nishimura Y, Kamikawa R, Hashimoto T & Inagaki Y (2012) Separate origins of group I introns in two mitochondrial genes of the katablepharid *Leucocryptos marina*. *PLoS One* **7**: e37307.

Nosek J & Tomaska L (2003) Mitochondrial genome diversity: evolution of the molecular architecture and replication strategy. *Current Genetics* **44**: 73-84.

Nowacki M, Shetty K & Landweber LF (2011) RNA-Mediated Epigenetic Programming of Genome Rearrangements. *Annual Review of Genomics and Human Genetics* **12**: 367-389.

Nowacki M, Vijayan V, Zhou Y, Schotanus K, Doak TG & Landweber LF (2008) RNA-mediated epigenetic programming of a genome-rearrangement pathway. *Nature* **451**: 153-158.

Ohman M (2007) A-to-I editing challenger or ally to the microRNA process. *Biochimie* **89**: 1171-1176.

Okuda K & Shikanai T (2012) A pentatricopeptide repeat protein acts as a site-specificity factor at multiple RNA editing sites with unrelated cis-acting elements in plastids. *Nucleic Acids Research* **40**: 5052-5064.

Ostermeier C, Harrenga A, Ermler U & Michel H (1997) Structure at 2.7 A resolution of the *Paracoccus denitrificans* two-subunit cytochrome c oxidase complexed with an antibody FV fragment. *Proceedings National Academy of Sciences of United States of America* **94**: 10547-10553.

Ozsolak F & Milos PM (2011) RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics* **12**: 87-98.

Panigrahi AK, Ernst NL, Domingo GJ, Fleck M, Salavati R & Stuart KD (2006) Compositionally and functionally distinct editosomes in *Trypanosoma brucei*. *RNA* **12**: 1038-1049.

Panigrahi AK, Schnaufer A, Ernst NL, Wang B, Carmean N, Salavati R & Stuart K (2003) Identification of novel components of *Trypanosoma brucei* editosomes. *RNA* **9**: 484-492.

Panigrahi AK, Zikova A, Dalley RA*, et al.* (2008) Mitochondrial complexes in *Trypanosoma brucei*: a novel complex and a unique oxidoreductase complex. *Molecular and Cellular Proteomics* 7: 534-545.

Panigrahi AK, Schnaufer A, Carmean N*, et al.* (2001) Four related proteins of the *Trypanosoma brucei* RNA editing complex. *Molecular and Cellular Biology* **21**: 6833-6840.

Parfrey LW, Barbero E, Lasser E, Dunthorn M, Bhattacharya D, Patterson DJ & Katz LA (2006) Evaluating support for the current classification of eukaryotic diversity. *PLoS Genetics* **2**: e220.

Pearson WR (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol* **132**: 185-219.

Pearson WR (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods in Molecular Biology* **132**: 185-219.

Pelin A, Pombert JF, Salvioli A, Bonen L, Bonfante P & Corradi N (2012) The mitochondrial genome of the arbuscular mycorrhizal fungus *Gigaspora margarita* reveals two unsuspected trans-splicing events of group I introns. *The New Phytologist* **194**: 836-845.

Pelletier M & Read LK (2003) RBP16 is a multifunctional gene regulatory protein involved in editing and stabilization of specific mitochondrial mRNAs in *Trypanosoma brucei*. *RNA* **9**: 457-468.

Pollard VW, Rohrer SP, Michelotti EF, Hancock K & Hajduk SL (1990) Organization of minicircle genes for guide RNAs in *Trypanosoma brucei*. *Cell* **63**: 783-790.

Pombert JF & Keeling PJ (2010) The mitochondrial genome of the entomoparasitic green alga *Helicosporidium*. *PLoS One* **5**: e8954.

Preisfeld A, Busse I, Klingberg M, Talke S & Ruppel HG (2001) Phylogenetic position and inter-relationships of the osmotrophic euglenids based on SSU rDNA data, with emphasis on the *Rhabdomonadales* (Euglenozoa). *International Journal of Systematic and Evolutionary Microbiology* **51**: 751-758.

Price DH & Gray MW (1998) *Editing of tRNA*. In modification and editing of RNA. p. 377-393

Rackham O & Filipovska A (2012) The role of mammalian PPR domain proteins in the regulation of mitochondrial gene expression. *Biochimica et Biophysica Acta* **1819**: 1008-1016.

Rackham O, Mercer TR & Filipovska A (2012) The human mitochondrial transcriptome and the RNA-binding proteins that regulate its expression. *Wiley Interdiscip Rev RNA* **3**: 675-695.

Rackham O, Davies SM, Shearwood AM, Hamilton KL, Whelan J & Filipovska A (2009) Pentatricopeptide repeat domain protein 1 lowers the levels of mitochondrial leucine tRNAs in cells. *Nucleic Acids Research* **37**: 5859-5867.

Rhee AC, Somerlot BH, Parimi N & Gott JM (2009) Distinct roles for sequences upstream of and downstream from *Physarum* editing sites. *RNA* **15**: 1753-1765.

Rodriguez-Ezpeleta N, Teijeiro S, Forget L, Burger G & Lang BF (2009) 3. Generation of cDNA libraries: Protists and Fungi. *Methods in Molecular Biology: Expressed Sequence Tags (ESTs),* Vol. 533 (Parkinson J, ed.^eds.), p.^pp. 33-47. Humana Press, Totowa, NJ.

Rodriguez-Ezpeleta N, Teijeiro S, Forget L, Burger G & Lang BF (2009) 3. Generation of cDNA libraries: Protists and Fungi. *Methods in Molecular Biology: Expressed Sequence Tags (ESTs),* Vol. 533 (Parkinson J, ed.^eds.), p.^pp. Humana Press, Totowa, NJ.

Rodriguez-Ezpeleta N, Brinkmann H, Burey SC*, et al.* (2005) Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. *Current Biology* **15**: 1325-1330.

Rogers K, Gao G & Simpson L (2007) Uridylate-specific 3' 5'-exoribonucleases involved in uridylate-deletion RNA editing in trypanosomatid mitochondria. *The Journal of Biological Chemistry* **282**: 29073-29080.

Rorbach J & Minczuk M (2012) The post-transcriptional life of mammalian mitochondrial RNA. *The Biochemical Journal* **444**: 357-373.

Roy J (2006) Etude de l'évolution de la strcuture des génomes mitochondriaux chez les *Euglenozoa*. Thèse,78p. Université de Montréal.

Roy J, Faktorova D, Lukes J & Burger G (2007) Unusual mitochondrial genome structures throughout the Euglenozoa. *Protist* **158**: 385-396.

Roy J, Faktorova D, Benada O, Lukes J & Burger G (2007) Description of *Rhynchopus euleeides n. sp*. (Diplonemea), a free-living marine euglenozoan. *Journal of Eukaryotic Microbiology* **54**: 137-145.

Rudinger M, Polsakiewicz M & Knoop V (2008) Organellar RNA editing and plant-specific extensions of pentatricopeptide repeat proteins in jungermanniid but not in marchantiid liverworts. *Molecular Biology and Evolution* **25**: 1405-1414.

Rudinger M, Fritz-Laylin L, Polsakiewicz M & Knoop V (2011) Plant-type mitochondrial RNA editing in the protist *Naegleria gruberi*. *RNA* **17**: 2058-2062.

Rudinger M, Szovenyi P, Rensing SA & Knoop V (2011) Assigning DYW-type PPR proteins to RNA editing sites in the funariid mosses *Physcomitrella patens* and *Funaria hygrometrica*. *The Plant Journal for Cell and Molecular Biology* **67**: 370-380.

Ruzzenente B, Metodiev MD, Wredenberg A*, et al.* (2012) LRPPRC is necessary for polyadenylation and coordination of translation of mitochondrial mRNAs. *The EMBO Journal* **31**: 443-456.

Salinas T, Duchene AM & Marechal-Drouard L (2008) Recent advances in tRNA mitochondrial import. *Trends in Biochemical Science* **33**: 320-329.

Salone V, Rudinger M, Polsakiewicz M*, et al.* (2007) A hypothesis on the identification of the editing enzyme in plant organelles. *FEBS letters* **581**: 4132-4138.

Schafer B (2005) RNA maturation in mitochondria of *S. cerevisiae* and *S. pombe*. *Gene* **354**: 80-85.

Schmidt O, Pfanner N & Meisinger C (2010) Mitochondrial protein import: from proteomics to functional mechanisms. *Nature Reviews Molecular Cell Biology* **11**: 655-667.

Schmitz-Linneweber C & Small I (2008) Pentatricopeptide repeat proteins: a socket set for organelle gene expression. *Trends in Plant science* **13**: 663-670.

Schnaufer A, Ernst NL, Palazzo SS, O'Rear J, Salavati R & Stuart K (2003) Separate insertion and deletion subcomplexes of the *Trypanosoma brucei* RNA editing complex. *Molecular Cell* **12**: 307-319.

Schneider TD & Stephens RM (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research* **18**: 6097-6100.

Schuster W, Hiesel R, Wissinger B & Brennicke A (1990) RNA editing in the cytochrome b locus of the higher plant *Oenothera berteriana* includes a U-to-C transition. *Molecular and Cellular Biology* **10**: 2428-2431.

Shapiro TA, Klein VA & Englund PT (1999) Isolation of kinetoplast DNA. *Methods in Molecular Biology* **94**: 61-67.

Shikanai T (2006) RNA editing in plant organelles: machinery, physiological function and evolution. *Cellular and Molecular Life Sciences* **63**: 698-708.

Shutt TE & Gray MW (2006) Bacteriophage origins of mitochondrial replication and transcription proteins. *Trends in Genetics* **22**: 90-95.

Shutt TE, Lodeiro MF, Cotney J, Cameron CE & Shadel GS (2010) Core human mitochondrial transcription apparatus is a regulated two-component system in vitro. *Proceedings of the National Academy of Sciences of the United States of America* **107**: 12133-12138.

Simpson AG & Roger AJ (2004) Protein phylogenies robustly resolve the deep-level relationships within Euglenozoa. *Molecular Phylogenetic and Evolution* **30**: 201-212.

Simpson AG, Lukes J & Roger AJ (2002) The evolutionary history of kinetoplastids and their kinetoplasts. *Molecular Biology and Evolution* **19**: 2071-2083.

Simpson AG, Inagaki Y & Roger AJ (2006) Comprehensive multigene phylogenies of excavate protists reveal the evolutionary positions of "primitive" eukaryotes. *Molecular Biology and Evolution* **23**: 615-625.

Simpson AGB (1997) The identity and composition of the Euglenozoa. *Arch. Protistenkd.* **148**: 318–328.

Simpson L, Aphasizhev R, Gao G & Kang X (2004) Mitochondrial proteins and complexes in *Leishmania* and *Trypanosoma* involved in U-insertion/deletion RNA editing. *RNA* **10**: 159-170.

Sinniger F, Chevaldonne P & Pawlowski J (2007) Mitochondrial genome of *Savalia savaglia* (Cnidaria, Hexacorallia) and early metazoan phylogeny. *Journal of Molecular Evolution* **64**: 196-203.

Siomi MC, Sato K, Pezic D & Aravin AA (2011) PIWI-interacting small RNAs: the vanguard of genome defence. *Nature Reviews. Molecular Cell Biology* **12**: 246-258.

Slamovits CH, Saldarriaga JF, Larocque A & Keeling PJ (2007) The highly reduced and fragmented mitochondrial genome of the early-branching dinoflagellate *Oxyrrhis marina* shares characteristics with both apicomplexan and dinoflagellate mitochondrial genomes. *Journal of Molecular Biology* **372**: 356-368.

Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, Palmer JD & Taylor DR (2012) Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLoS Biology* **10**: e1001241.

Smith DR, Kayal E, Yanagihara AA, Collins AG, Pirro S & Keeling PJ (2012) First complete mitochondrial genome sequence from a box jellyfish reveals a highly fragmented linear architecture and insights into telomere evolution. *Genome Biology and Evolution* **4**: 52-58.

Smith SW, Overbeek R, Woese CR, Gilbert W & Gillevet PM (1994) The genetic data environment an expandable GUI for multiple sequence analysis. *Computer Applications in the Biosciences* **10**: 671-675.

Sogin ML, Morrison HG, Hinkle G & Silberman JD (1996) Ancestral relationships of the major eukaryotic lineages. *Microbiologia* **12**: 17-28.

Soma A, Onodera A, Sugahara J*, et al.* (2007) Permuted tRNA genes expressed via a circular RNA intermediate in *Cyanidioschyzon merolae*. *Science* **318**: 450-453.

Spencer DF & Gray MW (2011) Ribosomal RNA genes in *Euglena gracilis* mitochondrial DNA: fragmented genes in a seemingly fragmented genome. *Molecular Genetics and Genomics* **285**: 19-31.

Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22**: 2688-2690.

Stoddard BL (2005) Homing endonuclease structure and function. *Quarterly Reviews of Biophysics* **38**: 49-95.

Storz G, Vogel J & Wassarman KM (2011) Regulation by small RNAs in bacteria: expanding frontiers. *Molecular Cell* **43**: 880-891.

Stuart KD, Schnaufer A, Ernst NL & Panigrahi AK (2005) Complex management: RNA editing in trypanosomes. *Trends in Biochemical Sciences* **30**: 97-105.

Sturm NR & Simpson L (1990) Kinetoplast DNA minicircles encode guide RNAs for editing of cytochrome oxidase subunit III mRNA. *Cell* **61**: 879-884.

Sturm NR, Maslov DA, Grisard EC & Campbell DA (2001) *Diplonema spp*. possess spliced leader RNA genes similar to the Kinetoplastida. *Journal of Eukaryotic Microbiology* **48**: 325-331.

Sung TY, Tseng CC & Hsieh MH (2010) The SLO1 PPR protein is required for RNA editing at multiple sites with similar upstream sequences in *Arabidopsis* mitochondria. *The Plant journal for cell and molecular biology*.

Takano H, Abe T, Sakurai R*, et al.* (2001) The complete DNA sequence of the mitochondrial genome of *Physarum polycephalum*. *Molecular & General Genetics* **264**: 539-545.

Takenaka M (2010) MEF9, an E-subclass pentatricopeptide repeat protein, is required for an RNA editing event in the nad7 transcript in mitochondria of Arabidopsis. *Plant Physiology and Biochemistry : PPB* **152**: 939-947.

Takenaka M, Zehrmann A, Verbitskiy D, Kugelmann M, Hartel B & Brennicke A (2012) Multiple organellar RNA editing factor (MORF) family proteins are required for RNA editing in mitochondria and plastids of plants. *Proceedings of the National Academy of Sciences of the United States of America* **109**: 5104-5109.

Talen JL, Sanders JP & Flavell RA (1974) Genetic complexity of mitochondrial DNA from *Euglena gracilis*. *Biochimica et Biophysica Acta* **374**: 129-135.

Thompson JD, Higgins DG & Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22**: 4673-4680.

Triemer RE & Ott DW (1990) Ultrastructure of *Diplonema ambulator* larsen & patterson (euglenozoa) and its relationship to *Isonema*. *European Journal of Protistology* **25**: 316-320.

Tsukihara T & Yoshikawa S (1998) Crystal structural studies of a membrane protein complex, cytochrome c oxidase from bovine heart. *Acta Crystallography A* **54**: 895-904.

Valadkhan S (2010) Role of the snRNAs in spliceosomal active site. *RNA Biology* **7**: 345-353.

Valadkhan S & Jaladat Y (2010) The spliceosomal proteome: at the heart of the largest cellular ribonucleoprotein machine. *Proteomics* **10**: 4128-4141.

Valles Y, Halanych KM & Boore JL (2008) Group II introns break new boundaries: presence in a bilaterian's genome. *PLoS One* **3**: e1488.

Verbitskiy D, Zehrmann A, Hartel B, Brennicke A & Takenaka M (2012) Two Related RNA-editing Proteins Target the Same Sites in Mitochondria of *Arabidopsis thaliana*. *The Journal of Biological Chemistry* **287**: 38064-38072.

Vlcek C, Marande W, Teijeiro S, Lukeš J & Burger G (2011) Systematically fragmented genes in a multipartite mitochondrial genome. *Nucleic Acids Research* **39**: 979-988.

von der Heyden S, Chao EE, Vickerman K & Cavalier-Smith T (2004) Ribosomal RNA phylogeny of bodonid and diplonemid flagellates and the evolution of euglenozoa. *Journal of Eukaryotic Microbiology* **51**: 402-416.

Wachtel C & Manley JL (2009) Splicing of mRNA precursors: the role of RNAs and proteins in catalysis. *Molecular BioSystems* **5**: 311-316.

Wahl MC, Will CL & Luhrmann R (2009) The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**: 701-718.

Waller RF & Jackson CJ (2009) Dinoflagellate mitochondrial genomes: stretching the rules of molecular biology. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology* **31**: 237-245.

Waller RF & Jackson CJ (2009) Dinoflagellate mitochondrial genomes: stretching the rules of molecular biology. *Bioessays* **31**: 237-245.

Ward BL, Anderson RS & Bendich AJ (1981) The mitochondrial genome is large and variable in a family of plants (Cucurbitaceae). *Cell* **25**: 793-803.

Waters LS & Storz G (2009) Regulatory RNAs in bacteria. *Cell* **136**: 615-628.

Wei W, Pelechano V, Jarvelin AI & Steinmetz LM (2011) Functional consequences of bidirectional promoters. *Trends in Genetics* **27**: 267-276.

Wei YH (1998) Mitochondrial DNA mutations and oxidative damage in aging and diseases: an emerging paradigm of gerontology and medicine. *Proceedings of the National Science Council, Republic of China. Part B, Life sciences* **22**: 55-67.

Wolstenholme DR (1992) Animal mitochondrial DNA: structure and evolution. *International Review of Cytology* **141**: 173-216.

Xu F, Morin C, Mitchell G, Ackerley C & Robinson BH (2004) The role of the LRPPRC (leucine-rich pentatricopeptide repeat cassette) gene in cytochrome oxidase assembly: mutation causes lowered levels of COX (cytochrome c oxidase) I and COX III mRNA. *The Biochemical Journal* **382**: 331-336.

Yagi Y, Tachikawa M, Noguchi H, Satoh S, Obokata J & Nakamura T (2013) Pentatricopeptide repeat proteins involved in plant organellar RNA editing. *RNA Biology* **10**.

Yasuhira S & Simpson L (1997) Phylogenetic affinity of mitochondria of *Euglena gracilis* and kinetoplastids using cytochrome oxidase I and *hsp60*. *Journal of Molecular Evolution* **44**: 341-347.

Yasuhira S & Simpson L (1997) Phylogenetic affinity of mitochondria of *Euglena gracilis* and kinetoplastids using cytochrome oxidase I and hsp60. *J Mol Evol* **44**: 341-347.

Zinshteyn B & Nishikura K (2009) Adenosine-to-inosine RNA editing. *Wiley Interdisciplinary Review Systems Biology and Medicine* **1**: 202-209.

Annexes

Table1: Amorces utilisées pour la RT-PCR et PCR

Amorces	Séquence
da21	AGTGGCCAGTACACCAGTA
ds1	CACTAATAGTCACAGAGTGTTGC
ds12	GGTTACTTCGTATGAGCACATC
re4	GTGCTACCATGCTATGGAGCA
re9	GGTATTGGTGCAATGGAGAG
dp56	CCAACCACTCCACTAGCAGCT
dp57	CCCTGGATGACGCAGGTAAC
dp59	CACGTTGTGGCCAACCAGTG
dp62	CCGTACAGGGTGCCAGTAACG
dp71	CTCCCGTAGCATGTCCTAGCG
dp72	CCATTAGCTCTACCGTACCTA
dp93	CTGGTGCTAGCTGAGACCTTAGG
dp114	CATCTGGTGTCCTACTCCGG
dp179	CCACCACATAGTATGCTATGAGC
dp180	CACCGGTTGGCGCTATGGTA
dp181	AGCTGTAACCCATCCACTAAC

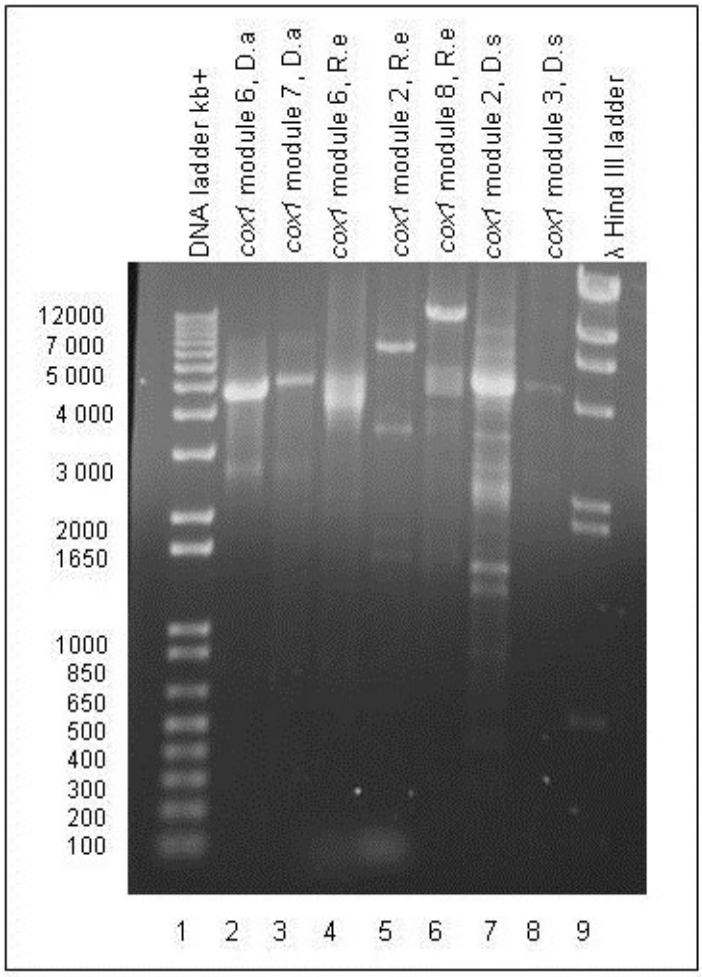
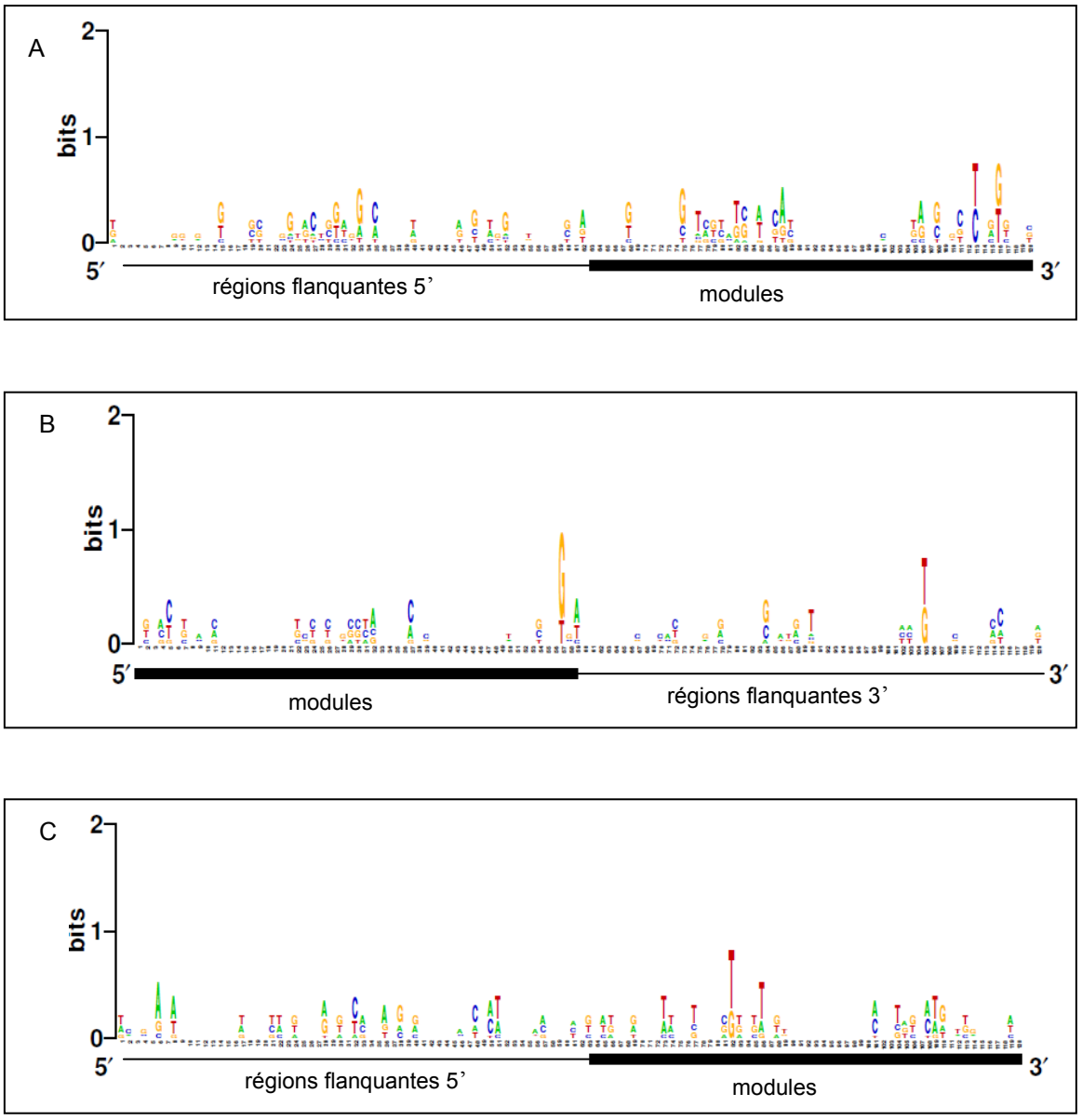
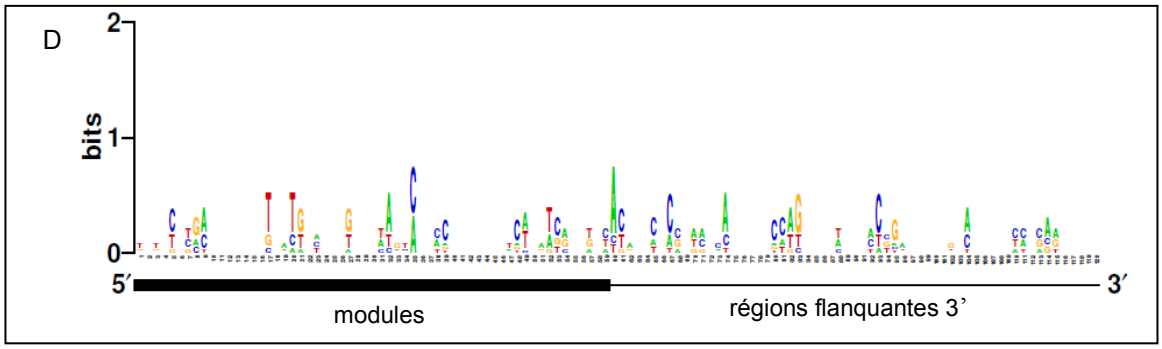
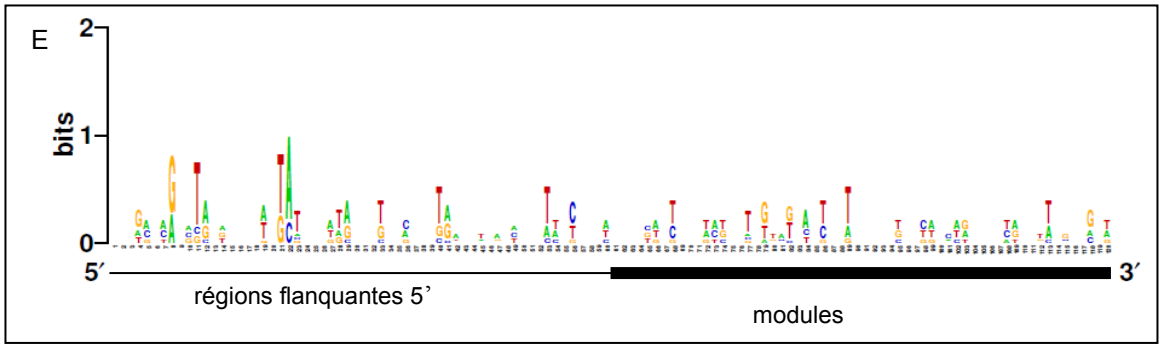


Figure 1. Chromosomes de *cox1* de diplonémides amplifiés par PCR sur gel d'agarose 0.8%. Les lignes 1 et 9 contiennent les marqueurs 1 kb+ et λ Hind III; 2 et 3 contiennent des chromosomes de *D. ambulator*; 4, 5 et 6 contiennent des chromosomes de *R. euleeides* et 7 et 8 contiennent des chromosomes de *D.* sp. 2.

Figure 2. Logos de la conservation des séquences dans les régions 5’ et 3’ des neufs modules de *cox1* chez les diplonémides. A, B, chez *D. papillatum*; C, D, chez *D.* sp. 2; E, F, chez *R. euleeides*.




D
bits
2
1
0
5′
3′
modules
régions flanquantes 3’


E
bits
2
1
0
5′
3′
régions flanquantes 5’
modules

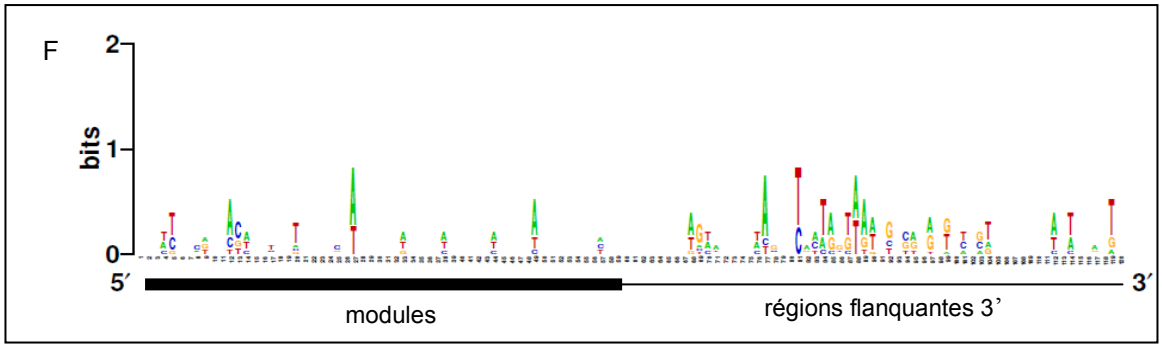
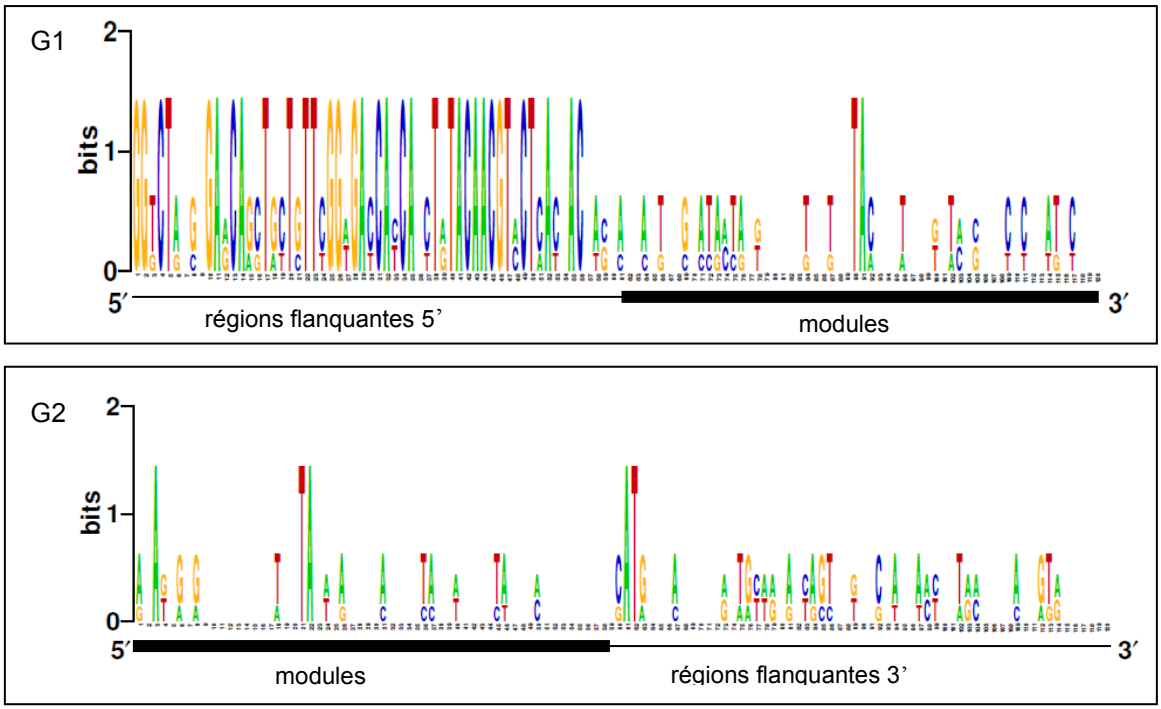
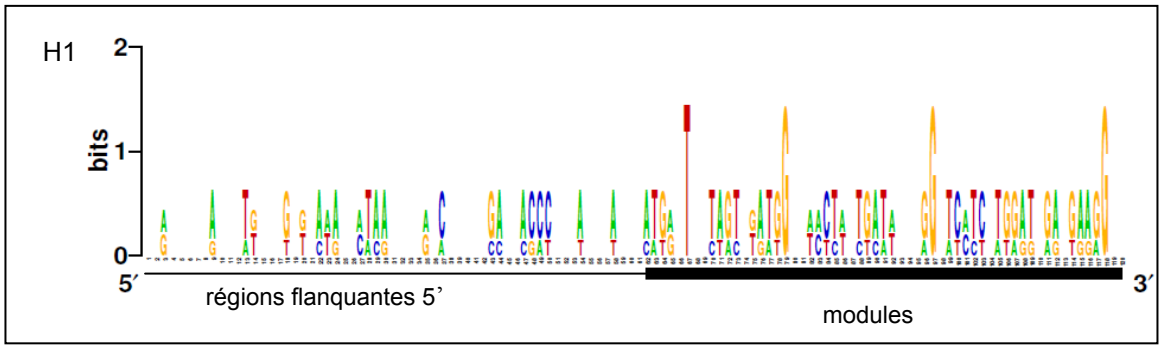
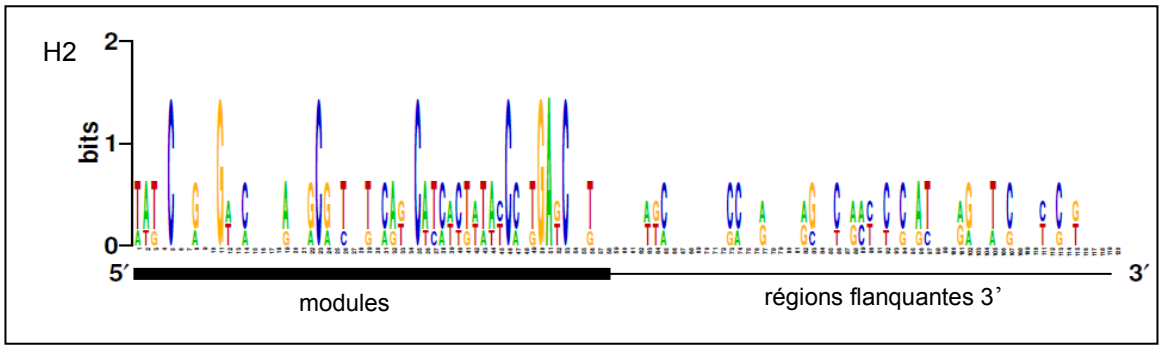
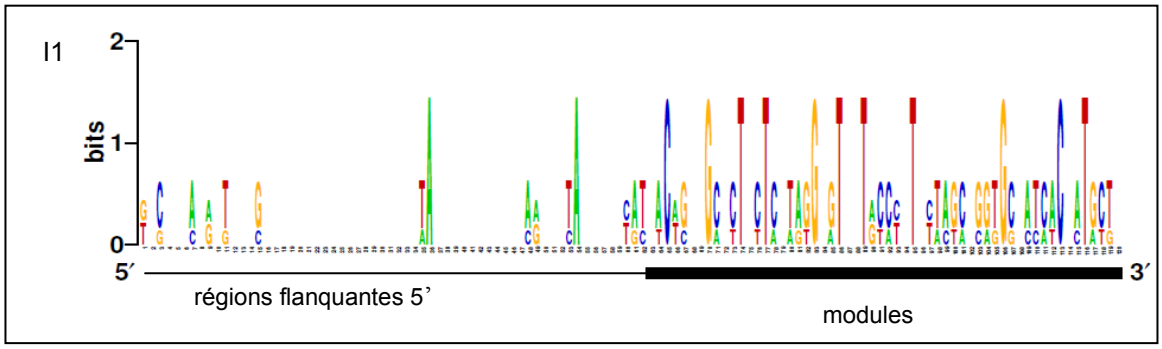

F
bits
2
1
0
5′
3′
modules
régions flanquantes 3’

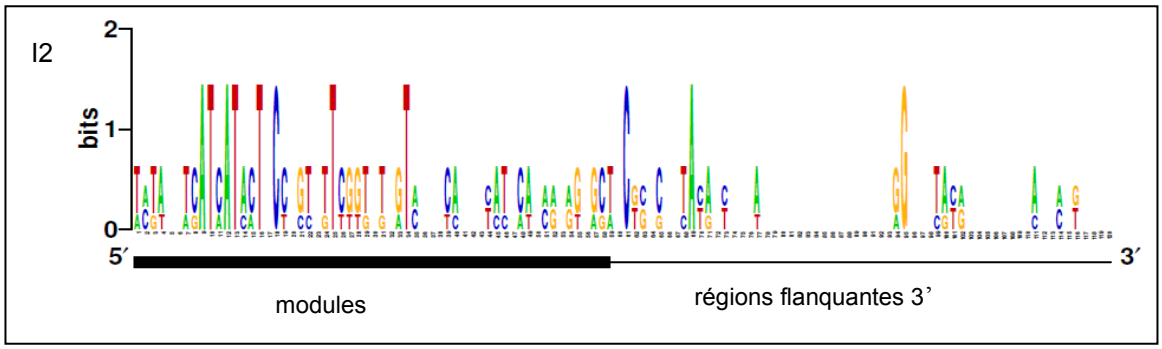
Figure 3. Logos de la conservation des séquences dans les régions 5’ et 3’ de chaque module de *cox1* chez les diplonémides. Les figures G1, H1, I1, J1, K1, L1, M1, N1: montrent la conservation des séquences dans les régions 5’ des modules 1, 3, 4, 5,6, 7, 8, 9 respectivement; G2, H2, I2, J2, K2, L2, M2, N2 montrent la conservation des séquences dans les régions 3’ du module 1, 3, 4, 5, 6, 7, 8, 9 respectivement.

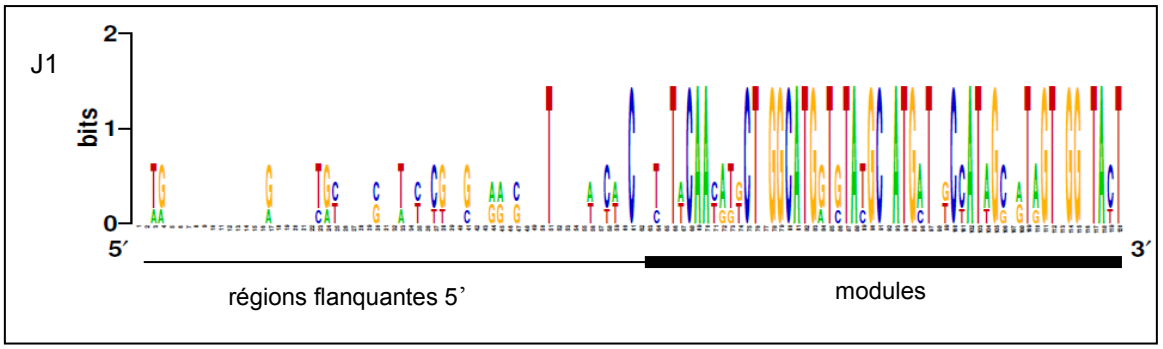


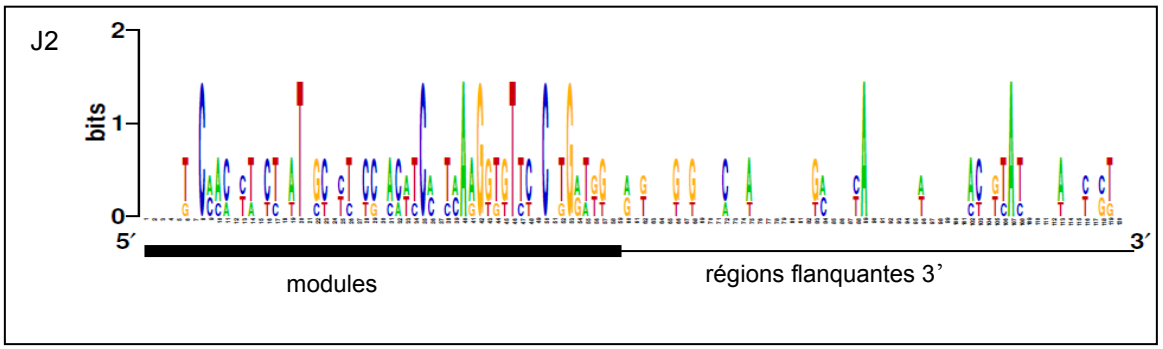

H1
bits
2
1
0
5′
3′
régions flanquantes 5’
modules

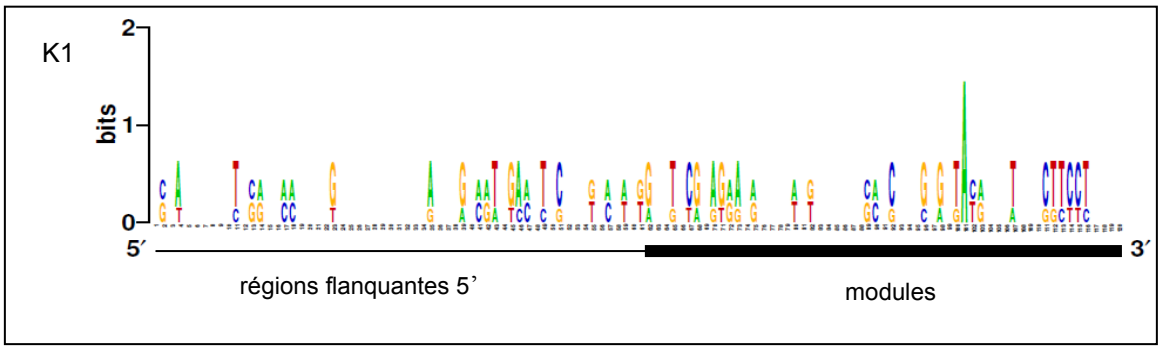

H2
bits
2
1
0
5′
3′
modules
régions flanquantes 3’

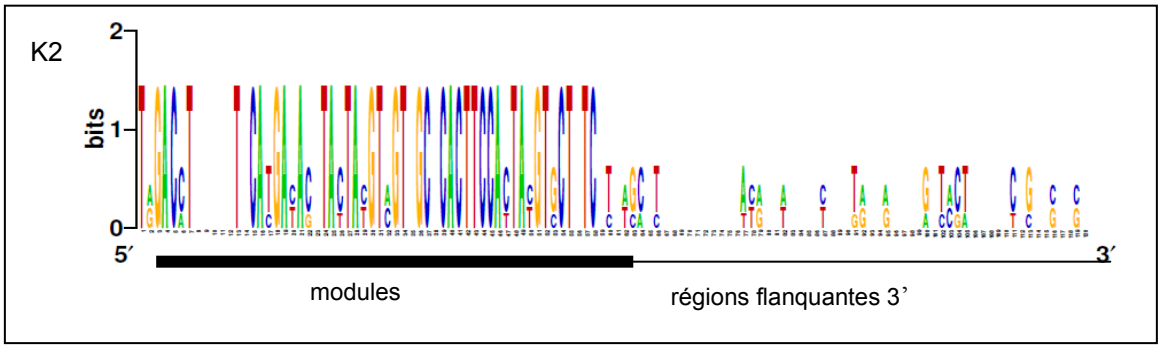

I1
bits
2
1
0
5′
3′
régions flanquantes 5’
modules

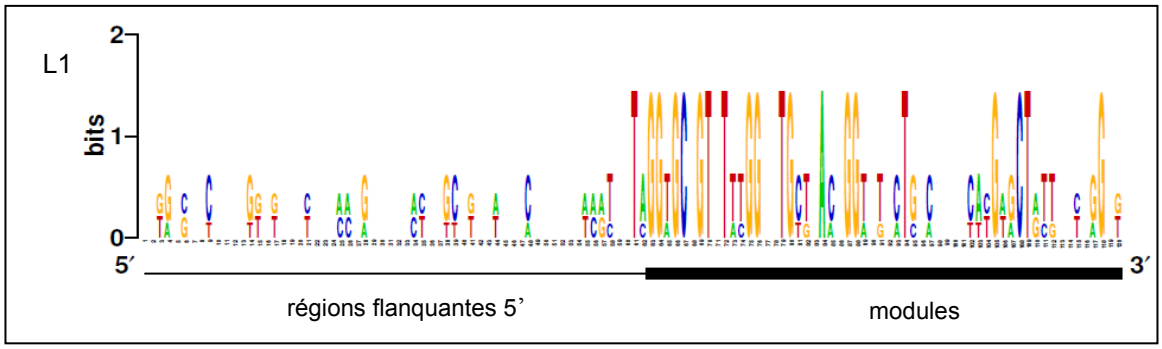

I2
bits
2
1
0
5′
3′
modules
régions flanquantes 3’

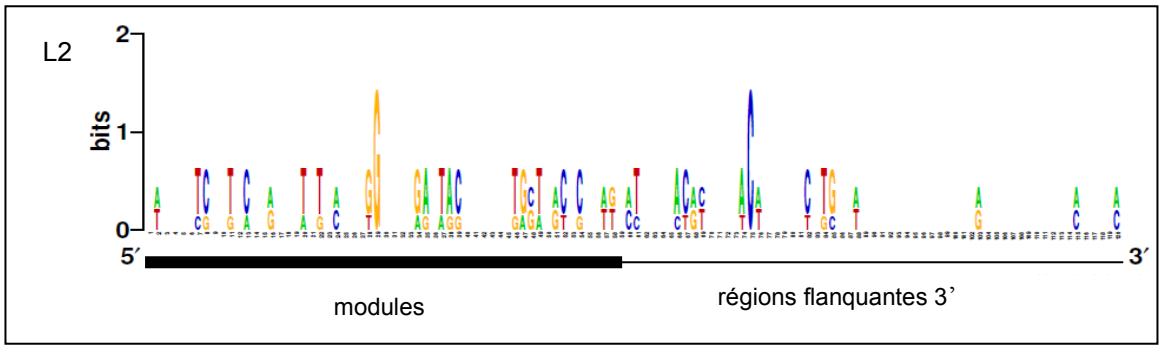

J1
bits
2
1
0
5′
3′
régions flanquantes 5’
modules

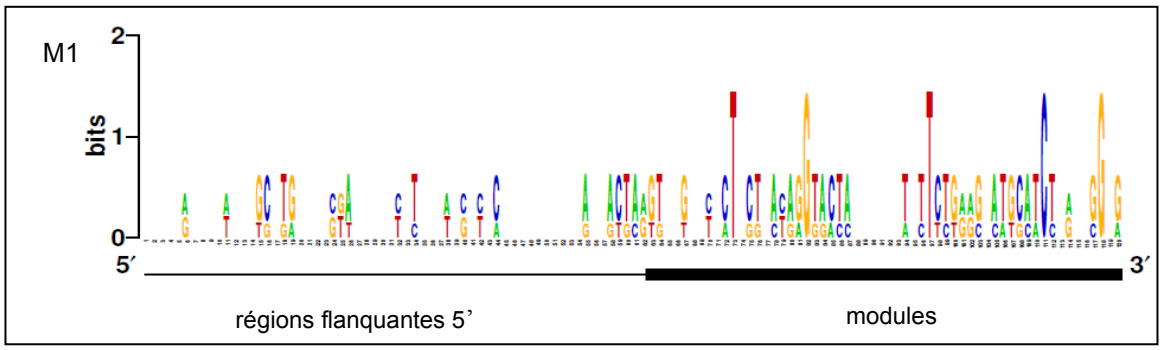

J2
bits
2
1
0
5′
3′
modules
régions flanquantes 3’

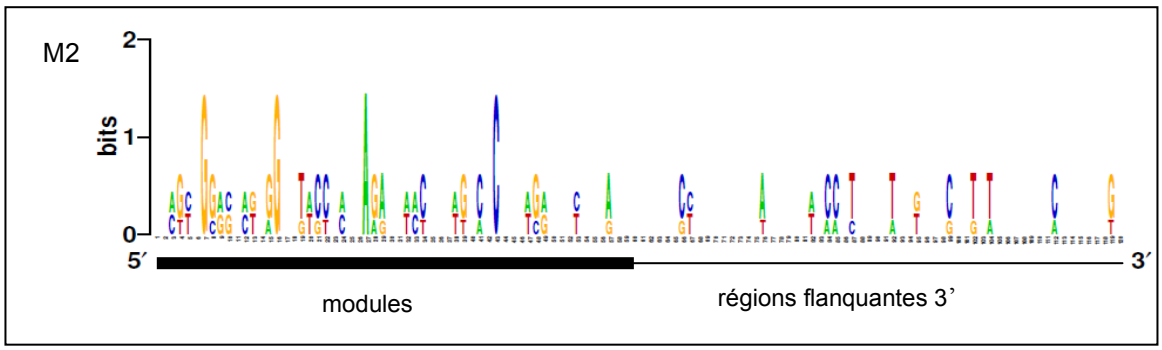

K1
bits
2
1
0
5′
3′
régions flanquantes 5’
modules

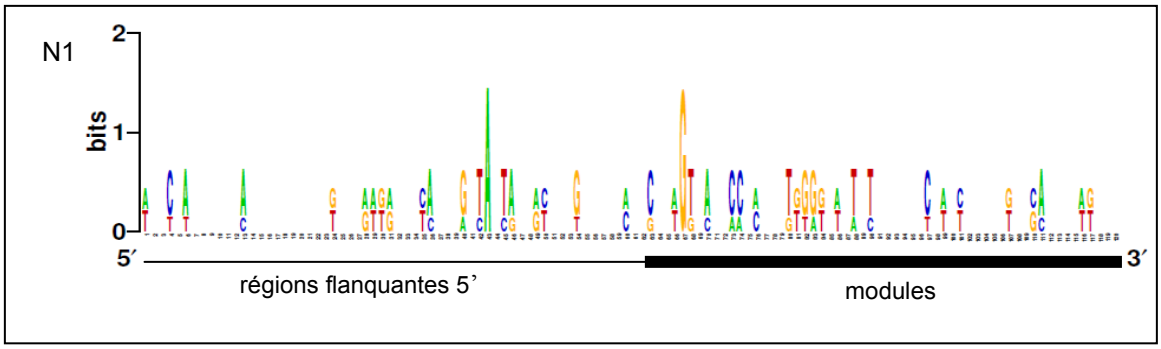

K2
bits
2
1
0
5′
3′
modules
régions flanquantes 3’

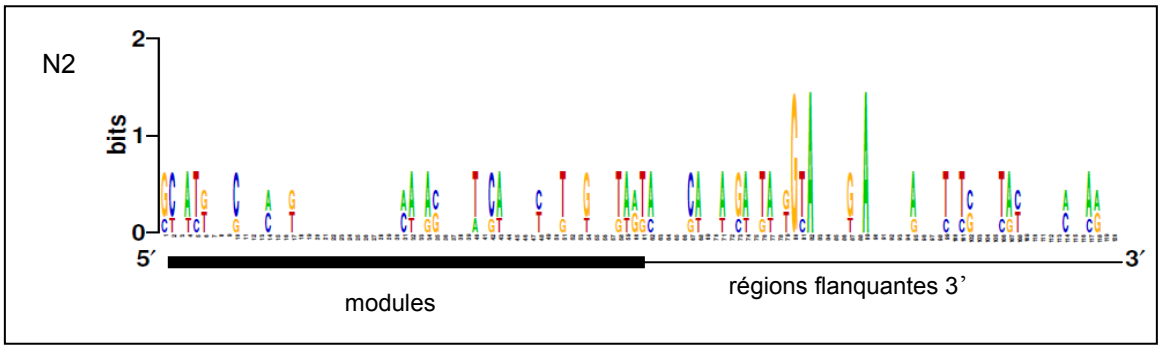

L1
bits
2
1
0
5′
3′
régions flanquantes 5’
modules


L2
bits
2
1
0
5′
3′
modules
régions flanquantes 3’


M1
bits
2
1
0
5′
3′
régions flanquantes 5’
modules


M2
bits
2
1
0
5′
3′
modules
régions flanquantes 3’


N1
bits
2
1
0
5′
3′
régions flanquantes 5’
modules


N2
bits
2
1
0
5′
3′
modules
régions flanquantes 3’

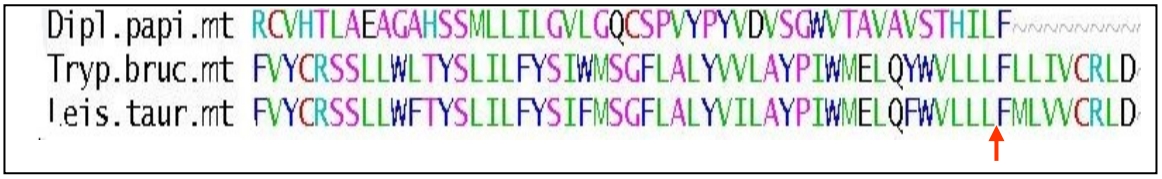


Figure 4. Alignement de séquences protéiques déduites des gènes *cob* chez trois euglénozoaires. La flèche rouge indique la phénylalanine conservée.