Université de Montréal

Décodage de l'expression de gènes cryptiques

par
Sandrine Moreira

Centre Robert-Cedergren
Département de Biochimie et Médecine moléculaire
Faculté de médecine

Thèse présentée en vue de l'obtention du grade de
Ph.D. en Bio-informatique

évaluée par un jury composé des personnes suivantes:
Martine Raymond, Président-rapporteur
Gertraud Burger, directrice de recherche
Marcel Turcotte, co-directeur de recherche
Marlène Oeffinger, membre du jury
Purificacion Lopez-Garcia, examinateur externe
Pascale Legault, représentant du doyen de la FES

Dépôt Août 2016

Resumé

Pour certaines espèces, les nouvelles technologies de séquençage à haut débit et les pipelines automatiques d'annotation permettent actuellement de passer du tube Eppendorf au fichier genbank en un clic de souris, ou presque. D'autres organismes, en revanche, résistent farouchement au bio-informaticien le plus acharné en leur opposant une complexité génomique confondante. Les diplonémides en font partie. Ma thèse est centrée sur la découverte de **nouvelles stratégies d'encryptage de l'information génétique** chez ces eucaryotes, et l'**identification des processus moléculaires de décodage**.

Les diplonémides sont des **protistes marins** qui prospèrent à travers tous les océans de la planète. Ils se distinguent par une diversité d'espèces riche et inattendue. Mais la caractéristique la plus fascinante de ce groupe est leur **génome mitochondrial en morceaux** dont les gènes sont **encryptés**. Ils sont décodés au niveau ARN par trois processus: (i) **l'épissage en trans**, (ii) l'**édition par polyuridylation** à la jonction des fragments de gènes, et (iii) l'édition par substitution de **A-vers-I et C-vers-T**; une diversité de processus post-transcriptionnels exceptionnelle dans les mitochondries.

Par des **méthodes bio-informatiques**, j'ai reconstitué complètement le **transcriptome mitochondrial** à partir de données de séquences ARN à haut débit. Nous avons ainsi découvert six nouveaux gènes dont l'un présente des isoformes par épissage alternatif en trans, 216 positions éditées par polyuridylation sur 14 gènes (jusqu'à 29 uridines par position) et 114 positions éditées par déamination de A-vers-I et C-vers-T sur sept gènes (*nad4, nad7, rns, y1, y2, y3, y5*).

Afin d'identifier les composants de la **machinerie réalisant la maturation des ARNs mitochondriaux**, le génome nucléaire a été séquencé, puis je l'ai assemblé et annoté. Cette machinerie est probablement singulière et complexe car aucun signal en *cis* ni acteur en *trans* caractéristiques des machineries d'épissage connues n'a été trouvé. J'ai identifié plusieurs candidats prometteurs qui devront être validés expérimentalement: des **ARN ligases**, un nombre important de protéines de la famille des **PPR** impliquées dans l'édition des ARNs dans les organites de plantes, ainsi que plusieurs **déaminases**.

Durant ma thèse, nous avons mis en évidence de nouveaux types de maturation post-transcriptionnelle des ARNs dans la mitochondrie des diplonémides et identifié des candidats prometteurs de la machinerie. Ces composants, capables de lier précisément des fragments d'ARN et de les éditer pourraient trouver des **applications biotechnologique**. Au niveau évolutif, la caractérisation de nouvelles excentricités moléculaires de ce type nous donne une idée des processus de recrutement de gènes, de leur adaptation à de nouvelles fonctions, et de la mise en place de **machineries moléculaires complexes**.

Mots clés

Bio-informatique, génomique, édition d'ARN, épissage en trans, assemblage de génome, annotation

Abstract

Thanks to new high throughput sequencing technologies and automatic annotation pipelines, proceeding from an eppendorf tube to a genbank file can be achieved in a single mouse click or so, for some species. Others, however, fiercely resist bioinformaticians with their confounding genomic complexity. Diplonemids are one of them. My thesis is centered on the discovery of **new strategies for encrypting genetic information** in eukaryotes, and the identification of **molecular decoding processes.**

Diplonemids are a group of poorly studied **marine protists**. Unexpectedly, metagenomic studies have recently ranked this group as one of the most diverse in the oceans. Yet, their most distinctive feature is their **multipartite mitochondrial genome** with **genes in pieces**, and **encryption** by nucleotide deletions and substitutions. Genes are decrypted at the RNA level through three processes: (i) **trans-splicing**, (ii) **polyuridylation** at the junction of gene pieces and (iii) substitutions of **A-to-I and C-to-T**. Such a diverse arsenal of mitochondrial post-transcriptional processes is highly exceptional.

Using a **bioinformatics approach**, I have reconstructed the **mitochondrial transcriptome** from RNA-seq libraries. We have identified six new genes including one that presents alternative trans-splicing isoforms. In total, there are 216 uridines added in 14 genes with up to 29 U insertions, and 114 positions edited by deamination (A-to-I or C-to-T) among seven genes (*nad4, nad7, rns, y1, y2, y3, y5*).

In order to identify the **machinery that processes mitochondrial RNAs**, the nuclear genome has been sequenced. I have then assembled and annotated the genome. This machinery is probably unique and complex because no *cis* signal or *trans* actor typical for known splicing machineries have been found. I have identified promising protein candidates that are worth to be tested experimentally, notably **RNA ligases**, numerous members of the **PPR** family involved in plants RNA editing and **deaminases**.

During my thesis, we have identified new types of post-transcriptional RNA processing in diplonemid mitochondria and identified new promising candidates for the machinery. A system capable of joining precisely or editing RNAs could find biotechnological applications. From an evolutionary perspective, the discovery of new molecular systems gives insight into the process of gene recruitment, adaptation to new functions and establishment of complex molecular machineries.

Keywords

Bioinformatics, genomics, RNA editing, trans-splicing, genome assembly, genome annotation

Sommaire

Résumé ii

Mots clés iii

Abstract iv

Keywords v

Liste des figures xi

Liste des tableaux xiv

Abbréviations et acronymes xv

Remerciements xvi

Avant-propos xiii

Chapitre 1: Introduction 1

I. Les mitochondries, des organites baroques 2

1. L'origine des mitochondries est liée à l'histoire des eucaryotes 2
2. Structure du génome mitochondrial 5
3. Contenu génique des mitochondries 8
4. Protéome mitochondrial et fonctions 12

II. Expression de l'information génétique mitochondriale 14

1. Transcription 14
2. Introns et épissage 15
3. Edition 18
4. Polyadénylation, polyuridylation des ARNs mitochondriaux (mt-ARN) 24

III. L'analyse bio-informatique de données génomiques 25

1. Les données de séquençage ADN et ARN 25
2. Identification des sites mitochondriaux d'édition d'ARN 25
3. Recherche des candidats de la machinerie d'édition et de trans-splicing dans le génome nucléaire 29

IV. *Diplonema papillatum* 34

1. Les Euglenozoa 34
2. Ecologie des diplonémides 36
3. Expression des gènes mitochondriaux 38

Chapitre 2: Trans-splicing and RNA editing of LSU rRNA in *Diplonema* mitochondria 41

I. Abstract 42

II. Introduction 42

III. Materials and methods 43

1. Sequences deposited in public-domain databases 43
2. Strain, culture and extraction of mtRNA 44
3. Northern hybridization 44
4. CircRT-PCR and RT-PCR 45
5. Cloning and sequencing of amplicons 46
6. High-throughput RNA sequencing 46
7. RNA-Seq data analysis 46
8. RNA secondary structure modeling 47

IV. Results 48

1. Identification of mt-LSU rRNA and its gene in *Diplonema* 48
2. 2D structure modeling of mt-LSU rRNA 52
3. Deep sequencing and RT-PCR analysis of the rnl transcript population 54
4. Maturation intermediates of rnl transcripts 55
5. Antisense RNA covering module junction and editing site of mt-LSU rRNA 58

V. Discussion 59

1. Regulation of rnl gene expression in *Diplonema* mitochondria 59
2. Antisense transcripts 61
3. *Diplonema* mt-LSU rRNA is extraordinarily short and derived 62
4. Role of extensive U-'insertion' editing in mt-LSU rRNA from *Diplonema* 63

VI. Supplementary data 65

Chapitre 3: Novel modes of RNA editing in mitochondria 76

I. Abstract 77

II. Introduction 77

III. Materials and methods 82

1. Strains, culture, and DNA and RNA extraction 82
2. In vitro transcription and RNase cleavage of glyoxalated RNA 82
3. DNA library construction, sequencing, read processing and assembly 82
4. RNA-Seq library construction, sequencing and read processing 83
5. Mapping of Illumina reads to reference sequences and calculation of FPKM 83

6. Mitochondrial reference genome sequence 83
7. Transcript de novo assembly and function annotation 84
8. Mitochondrial transcriptome reconstruction and assignment of orphan modules 84
9. In silico identification of polymorphic genomic sites 84
10. *In silico* identification of RNA editing sites 84
11. Analysis of RNA processing intermediates and partially edited transcripts 85
12. Search for *cis* elements and RNA *trans*-factors that guide RNA editing 85
13. Analysis of Nad4 protein sequences 86
IV. Results 86
1. Assignment of modules to genes 86
2. Identification of the tentative gene for mt-SSU rRNA 87
3. Uncovering RNA editing events in mitochondrial transcripts 88
4. Catalog of nucleotide insertions and substitutions in the *D. papillatum* mito-transcriptome 88
5. Searches for potential cis-elements and trans-factors that guide RNA editing 89
6. Biochemically, A-to-I substitution RNA editing proceeds by deamination 90
7. RNA editing precedes trans-splicing and progresses stochastically within editing clusters 92
8. Exceptionally high overall RNA editing rate, but several incompletely edited sites 94
9. Crucial consequences of RNA editing for the tentative mt-SSU rRNA and Nad4 protein 94
10. Comparison of RNA editing in *nad4* across diplonemids 96
V. Discussion 96
1. RNA editing types and sites in *D. papillatum* mitochondria 96
2. How does the cellular machinery recognize RNA editing sites in *Diplonema* mitochondria? 98
3. The biological role of organellar RNA editing 98
4. Emergence and diversification of the two RNA editing types in diplonemid mitochondria 99
VI. Outlook
VII.Supplementary data 99
1. Supplementary File 102
2. Supplementary Methods 102
3. Supplementary Results and Discussion 112
Chapitre 4: Three-dimensional structure model and predicted ATP interaction rewiring of a deviant RNA ligase 2 130

I. Abstract **131**

1. Background 131
2. Results 131
3. Conclusion 131

II. Background **131**

III. Results **134**

1. HMM-based detection of a divergent RNA ligase 2 in *Diplonema* 134
2. DpRNL contains a nucleotidyltransferase domain typical for RNA ligases 2 135
3. The 3D model of apo-DpRNL possesses all structural features typical for RNA ligases 2 137
4. Molecular dynamics simulations confirm the stability of the DpRNL 3D model 138
5. Comparative 3D structure analysis of DpRNL with well characterized RNA ligases 2 139
6. Phylogeny of RNA ligases 2 142
7. DpRNL is predicted to interact with RNA in a T4-like fashion 142
8. Refinement of the DpRNL structural model by molecular dynamics simulations 143
9. Predicted interactions of DpRNL with adenine and ribose of ATP 145
10. The triphosphate tail of ATP engages in a rich network of stabilizing interactions 147

IV. Discussion **148**

1. How the postulated rewiring of ATP interactions in DpRNL may have evolved 149
2. The biological process involving DpRNL 150

V. Conclusion **150**

VI. Methods **151**

1. Identification of RNA ligase 2 151
2. Three-dimensional structure modeling 152
3. System preparation for molecular dynamics simulations 152
4. Molecular dynamics simulations 152
5. Exceptional residues 154
6. 3D model analyses 155
7. Expression 155
8. Phylogenetic reconstruction of RNA ligases 2 from Excavata 156

VII. Additional file 1 **157**

Chapitre 5: Résultats non publiés **168**

I. Assemblage et annotation du génome nucléaire **169**

1. Assemblage 169
2. Annotation 170
3. Recherche de gènes candidats 173
Discussion 174
I. Relation structure-fonction: y a-t-il un lien entre la fragmentation des gènes, l'édition par polyuridylation et la structure des protéines ? 175
1. Les sites de fragmentation des gènes sont-ils aléatoires ? 175
2. La polyuridylation est liée à la fragmentation des gènes 176
3. La polyuridylation des transcrits ne peut survenir que dans des régions correspondant à des zones spécifiques de la protéines 177
II. Des candidats pour le traitement post-transcriptionnels des ARNs mitochondriaux 179
1. Candidats pour l'édition par polyuridylation 179
2. Candidats pour l'édition par substitution C-vers-U, et A-vers-G 180
3. Candidats pour l'épissage en trans 181
Conclusion 183
Références 184
Appendices 199

Liste des figures

Chapitre 1 1

Figure 1. Schéma de l'origine évolutive des mitochondries. 3
Figure 2. Classification des eucaryotes. 4
Figure 3. Diversité de taille et de structure des génomes mitochondriaux. 6
Figure 4. Contenu génique des mitochondries pour une sélection d'espèces. 10
Figure 5. Cas d'épissage en trans classés par type de machinerie. 16
Figure 6. Séquence du Spliced Leader de *Diplonema papillatum* 17
Figure 7. Les cas d'édition d'ARN connus. 19
Figure 8. Edition chez les kinétoplastides. 21
Figure 9: Lectures simples, appariées et taille d'insert 25
Figure 10. Lectures, contigs et scaffolds 29
Figure 11. Phylogénie des eucaryotes mettant en évidence les Euglenozoa. 34
Figure12. Dessin artistique de Rhyncopus euleeides par Shona Teijero (Université de Montréal). 35
Figure13. Biosynthèse de la base J. 35
Figure 14. Spliced leader. 35
Figure 15. Distribution des diplonemides dans les océans. 37
Figure 16. Diversité et abondance des diplonemides. 38
Figure 17. Structure des chromosomes mitochondriaux de *D. papillatum* 39
Figure 18. Expression des gènes mitochondriaux (avant 2011). 40

Chapitre 2 41

Figure 1. Mitochondrial LSU rRNA of *Diplonema*. 50
Figure 2. Putative secondary structure of the mt-LSU rRNA (3′ moiety) from *Diplonema*. 53
Figure 3. Coverage of *Diplonema* mt-LSU rRNA by RNA-Seq data. 55
Figure 4. Maturation intermediates of rnl transcripts. 57
Figure 5. Maturation process of mt-LSU rRNA in *Diplonema* mitochondria. 61
Figure S1. Northern hybridization of mt-LSU rRNA from *Diplonema*. 69
Figure S2. U-tract length distribution of non-quality-trimmed RNA-Seq reads. 69
Figure S3. RT-PCR across the U-tract of mt-LSU rRNA and its antisense transcripts. 70
Figure S4. Coverage of RNA-Seq reads along reference transcript sequences of *rnl*. 72
Figure S5. Effect of quality clipping on reads that bridge the rnl module-1/module-2 junction. 73
Figure S6. Length distribution of poly-(A) tails downstream of rnl module 2 in RNA-Seq reads. 74
Figure S7. Potential 2° structure of the 5′ half of mt-LSU rRNA from *Diplonema*. 75

Chapitre 3 76

Figure 1. Substitution RNA editing of mitochondrial genes from *Diplonema papillatum*. 81
Figure 2. Demonstration of inosines in substitution 91
Figure 3. Partial substitution RNA editing in nad4-m1. 93
Figure 4. Multiple alignment of nad4-m1 genomic (A) and cDNA sequences (B) from diplonemids. 95
Figure S1. Mitochondrial chromosome structure and tentative mitochondrion-encoded SSU rRNA of *D. papillatum*. 121
Figure S2. Module composition and RNA editing of *D. papillatum* mitochondrial genes. 121
Figure S3. Search for potential *cis*-elements and RNA trans-factors that recognize RNA editing sites. 123
Figure S4. Optimization of the glyoxal/borate/RNase T1 assay for the detection of inosines in RNA. 125
Figure S5. The pattern of partial substitution RNA editing of nad4-m1 is essentially the same in poly(A) and total RNA. 126
Figure S6. Trans-spliced polyadenylated transcripts contain partially edited. 127
Figure S7. Predicted trans-membrane (TM) helices of diplonemid Nad4 proteins. 128
Figure S8. Multiple alignment of the N-terminal region of Nad4 from various eukaryotes. 129

Chapitre 4 130

Figure. 1 Delineation of the Nucleotidyltransferase domain. 136
Figure. 2 Structural alignment with DALI [283] of the *Diplonema* model (first line) and four structures: 1XDN (*Trypanosoma brucei*), 1S68 (Enterobacteriophage T4), 2VUG (*Pyrococcus abyssi*), and 3QWU (*Aquifex aeolicus*) 136
Figure. 3 Three-dimensional model of DpRNL inferred by I-Tasser. 138
Figure. 4 Protein properties mapped onto DpRNL. 142
Figure. 5 Catalytic pocket of DpRNL and TbREL1. 146
Figure. 6 Structurally and functionally equivalent residues in DpRNL, TbREL1 and T4RNL2. Residues on the same line are structural equivalents (at the same position in a structural alignment). 146
Figure S1. Catalytic steps of the ligation reaction performed by an RNA ligase 2 adapted from [281] 157
Figure S2. Alignment of the PF09414-Rnl2 HMM model with DpRNL. 158
Figure S3. RMSD (Root Mean Square Deviation). 158
Figure S4. Define Secondary Structure of Proteins (DSSP) calculated for DpRNL (replicate 15R0+35_1) with the tool VMD. 159
Figure S5. Hydrophobicity of residues represented by different shades of red according to the Eisenberg hydrophobicity scale [325] on *Pyrococcus abyssi* and *Diplonema* 3D representation. 161
Figure S6. Phylogenetic tree of RNA ligases 2 from Excavata. 161
Figure S7. RMSF (Root Mean Square Fluctuation) of DpRNL apo backbone, 50-ns simulation. Blue line: smoothed average. 162
Figure S8. Number of violations of distance restraints during the 15-ns of the four ATP-restrained simulation replicates. 163
Figure S9. Design of MD simulations performed. 164
Figure S10. Proposed model for ATP interactions rewiring in DpRNL. 165

Figure S11. Catalytic pocket for the deviant MD simulation 166

Chapitre 5 168

Figure 1. Procédure pour l'assemblage du génome nucléaire de *D. papillatum.* 170
Figure 2. Procédure d'annotation du génome nucléaire de *D. papillatum.* 172

Liste des tableaux

Chapitre 1 1

Tableau I. Nombre de protéines composant le protéome mitochondrial chez une sélection d'espèces. 13

Chapitre 2 41

Table 1. Non-encoded U-tract length of mt-LSU rRNA and its precursor transcripts 51
Table S1. Oligonucleotides used in this study 65
Table S2. Pre-processing of RNA-Seq reads 66
Table S3. Mapping of RNA-Seq reads to mt-LSU rRNA 66
Table S4. Mapping of RNA-Seq reads to the module-1/module-2 junction of mt-LSU rRNA 66
Table S5. RT-PCR across the U-tract of mt-LSU rRNA and its antisense transcripts 67
Table S6. Artifactual T-tract length variation 68

Chapitre 3 76

Table 1. Genes and RNA editing sites in *D. papillatum* mitochondria 80
Table S1. Oligonucleotides used in this study. 115
Table S2. Libraries used in this study. 115
Table S3. Genomic sequence variants in cassettes of *D. papillatum* mtDNA. 116
Table S4A. Intermediates of substitution RNA editing, end-processing, and trans-splicing. 119
Table S4B. Intermediates of U-appendage RNA editing, end-processing, and trans-splicing. 119
Table S5. Conservation statistics of Nad4. 120
Table S6. GenBank accession numbers of diplonemid mitochondrial sequences. 120

Chapitre 4 130

Table 1 Identification of the ligase family to which belongs DpRNL 135
Table S1. Comparison of DpRNL with close structural homologs of interest from the output of ITasser and by structural comparison of the refined model of DpRNL. 167
Table S2. Comparison of ATP interaction with DpRNL and TbREL1. 167

Chapitre 5 168

Tableau I. Statistiques des annotations syntaxiques et fonctionnelles de *D. papillatum.* 173
Tableau II. Protéines candidates impliquées dans le traitement post-transcriptionnel des ARNs mitochondriaux. 173

Abbréviations et acronymes

ARN: Acide Ribonucléique

GC : Guanine-Cytosine

kb: kilobase

mt-ARN: ARN mitochondrial

mt-LSU-ARN: ARN de la grande sous-unité ribosomique mitochondriale

mt-SSU-ARN: ARN de la petite sous-unité ribosomique mitochondriale

PPR: Protéine avec des motifs Pentatricopeptides

RMSD: Root Mean Square Deviation

SL: *Spliced Leader*, Leader d'épissage

Remerciements

Il n'y a pas de thèse heureuse. C'est du moins ce que j'avais observé autour de moi avant de me lancer dans cette aventure. Mon expérience a été toute autre.

Je le dois surement à cet étrange et excitant sujet d'étude qu'est *Diplonema papillatum*.

Je le dois aussi à ma directrice de recherche, Gertraud Burger, à mon co-directeur Marcel Turcotte, et aux membres de mon comité de thèse, Muriel Aubry et Marlene Oeffinger, qui m'ont guidé avec bienveillance et respect. Ils ont été des modèles inspirants chacun à leur manière.

Je le dois bien sûr aux membres de notre équipe, Matus Valach, Mohamed Aoulad-Aissa, Sophie Breton, Yifei Yan ; des chercheurs téméraires bien décidés à percer les mystères des diplonémides ! J'ai eu la chance d'encadrer des étudiants formidables: Annie Lebreton, Jean Coquet, Emmanuel Noutahi, Lili-Anh Le Minh, Stephan Mbatchou. Je garderai le souvenir d'une équipe solidaire et de personnes intelligentes et d'une grande humanité. Les étudiants et chercheurs du centre Robert-Cedergren, par leurs conseils, les moments de détentes, et les discussions que nous avons partagé auront été d'inestimables compagnons de route. J'espère n'oublier personne en citant Natacha Beck, Henner Brinkman, Rachid Daoud, Lise Forget, Franz Lang, Nicolas Lartilot, Simon Laurin-Lemay, Ioana Minoiu, Sahar Parto, Sebastian Pechmann, Hervé Philippe, Raphaël Pujol, Lila Salhi, Matt Sarrasin, Adrian Serohijos, Jean-François Theroux, Yun Zhu.

Je le dois aussi à Sylvain Foisy, Marie Pageau et Christian Baron qui m'ont fait confiance pour l'enseignement des laboratoires de Bio-informatique et m'ont permis de réaliser à quel point j'aimais enseigner.

Je le dois aux étudiants bio-informaticiens qui sont devenus des amis : Geneviève Galarneau, Claudia Kleinman, Julie Hussin, Julie Nocq, Olivier Tremblay-Savard. Une bio-informaticienne d'adoption, Elaine Meunier, a une place particulière.

Je le dois finalement à ma famille, qui m'a soutenue et encouragée dans mes projets, quels qu'ils soient, quand bien même ils me conduisent à 6000 km d'elle.

Et je le dois enfin à Philippe, Hélène et Raphaëlle qui, quotidiennement, m'entourent de tout leur amour.

Merci.

A mon père, pour ses inventions farfelues qui me manquent tant.

A ma mère, pour son amour des chiffres.

Avant-propos

Identifier une nouvelle machinerie moléculaire, c'est un peu comme résoudre un crime. On se retrouve avec un cadavre sur les bras (un ARN épissé par exemple) et on doit identifier l'arme du crime (comment ?), le meurtrier (qui ?) et éventuellement le mobile (pourquoi ?). Le «pourquoi» est une questions sans fin en biologie moléculaire dont la myriade de réponses possibles mène à encore plus de questions. Pourquoi des génomes et des gènes en morceaux ? Pourquoi l'épissage ? Pourquoi brouiller le message génétique en encryptant les gènes ? Pourquoi tant d'ADN non codant ? Les questions du «comment» et du «qui» sont beaucoup plus sures. Dans ce domaine, l'imagination évolutive ne connait quasi-aucune limite. Nommez une réaction enzymatique extravagante, il y a fort à parier qu'elle existe. Ma thèse est centrée sur ces deux questions: la **découverte de nouveaux processus de décodage de l'expression de l'information génétique** (comment?) et l'**identification de la machinerie moléculaire** en jeu (qui?). J'ai utilisé des méthodes **bio-informatiques** d'analyse du génome et du transcriptome pour répondre à ces questions.

Sur le podium des espèces les plus extravagantes, *Diplonema papillatum* occupe certainement une place de choix. Son génome mitochondrial contient non pas un chromosome comme pour la majorité des espèces connues, mais une **centaine de chromosomes circulaires,** chacun portant un **fragment de gène**. Ces fragments sont transcrits indépendamment puis raboutés par **épissage en trans** («trans-splicing») pour constituer un ARN fonctionnel. Cette procédure post-transcriptionnelle unique n'est pas le seul traitement inhabituel des ARNs. Nous avons découverts qu'elle s'accompagne de l'édition de l'ARN par ajout de **polymères d'uridine** à la jonction de certains fragments (polyuridylation) et de modification du message ARN par **substitution de bases**. J'ai identifié des **composants potentiels prometteurs** des machineries moléculaires permettant l'épissage en trans et l'édition. Leur implication reste à démontrer expérimentalement.

Le premier chapitre est consacré à l'**introduction** du sujet. Je décrirai dans un premier temps les mitochondries, leur origine et leurs étranges caractéristiques génomiques, résultat d'un chemin évolutif tortueux, qui en font aujourd'hui un organite définitivement baroque. Je décrirai ensuite ce que l'on sait de l'expression des gènes mitochondriaux, même si de nombreux aspects restent encore inconnus. La mitochondrie, tiraillée entre son origine procaryote et sa citoyenneté eucaryote, est le siège privilégié de déviations par rapport au schéma d'expression standard. J'en décrirai deux en détail: l'épissage en trans et l'édition. Je ferai ensuite un tour d'horizon des principales méthodes bio-informatiques que j'ai utilisées. J'aurais pris soin dans les premières parties d'éviter de parler des diplonemides pour leur consacrer la dernière partie de l'introduction.

Dans le chapitre 2, nous nous intéresserons à la **polyuridylation** avec le cas du gène codant pour l'ARN de la grande sous-unité ribosomique mitochondriale. Ce gène est composé de deux fragments transcrits indépendamment, puis liés par 26 uridines.

Dans le chapitre 3, nous verrons l'**intense édition** que subissent les ARNs mitochondriaux.

Dans le chapitre 4 nous nous intéresserons à un candidat décevant de la machinerie d'épissage en trans: l'**ARN ligase 2**.

Finalement, dans le chapitre 5, nous verrons d'autres candidats possibles qui restent à valider.

Le travail présenté dans ce mémoire a été publié dans les articles et revues suivantes :

Chapitre 1, section II.2

S. Moreira, S. Breton, and G. Burger. 2012. "Unscrambling Genetic Information at the RNA Level." Wiley Interdisciplinary Reviews: RNA 3(2):213–28.

Chapitre 2

M. Valach, S. Moreira, G. N. Kiethega, and G. Burger. 2014. "Trans-Splicing and RNA Editing of LSU rRNA in *Diplonema* Mitochondria.." Nucleic Acids Research 42(4):2660–72.

Chapitre 3

S. Moreira, M. Valach, M. Aoulad-Aissa, C. Otto, and G. Burger. 2016. "Novel Modes of RNA Editing in Mitochondria." Nucleic Acids Research 44(10):4907–19.

Chapitre 4

S. Moreira, E. Noutahi, G. Lamoureux, and G. Burger. 2015. "Three-Dimensional Structure Model and Predicted ATP Interaction Rewiring of a Deviant RNA Ligase 2." BMC structural biology 15(1):20.

Discussion

G. Burger, S. Moreira, and M. Valach. 2016 "Genes in Hiding." Trends in Genetics 32(9):553–565

Chapitre 1: Introduction

I.Les mitochondries, des organites baroques

"Our mitochondria are, in essence, domesticated bacteria"
John Archibald, 2014 [1]

Les mitochondries sont des organites cellulaires communément désignées comme les "usines à énergie" de la cellule. Tous les eucaryotes connus ont eu des mitochondries durant leur histoire évolutive. D'une origine commune, elles ont ensuite dérivé pour présenter une telle variété morphologique, génétique et fonctionnelle, qu'elles méritent pleinement leur qualificatif d'**organite baroque** [2].

1. L'origine des mitochondries est liée à l'histoire des eucaryotes

Les mitochondries sont d'origine bactérienne

Les mitochondries sont les **reliques** de l'intégration d'une **alpha-proteobactérie** dans une cellule hôte il y a plus d'un milliard d'années [3]-[6]. L'endosymbiose fut d'abord suspectée grâce à des observations morphologiques, puis étayée par l'identification d'un génome mitochondrial, d'un système de traduction distinct de celui du cytosol et par des caractéristiques structurales et biochimiques des ARNs et des protéines mitochondriales. La preuve définitive viendra des analyses de **phylogénie moléculaires des ARNs ribosomiques** mitochondriaux qui montrent leur origine bactérienne (et celle du génome qui les contient) [3]. Elles placent le symbiote dans la famille des alpha-protéobactéries dont le plus proche représentant moderne serait un ***rickettsiale*** [5], [7]. La bactérie n'a été ni digérée, ni éliminée par l'hôte mais conservée sous forme de symbiote. La possibilité pour l'hôte de produire plus d'énergie pour le développement des coûteuses innovations eucaryotes serait un élément clé de la pérennité de cette symbiose [8].

La cellule hôte et le processus d'endosymbiose ne sont pas encore élucidés

La nature de la cellule hôte et le processus d'établissement de la symbiose sont encore des sujets très discutés [8], [9]. Plusieurs scénarios que l'on peut regrouper en deux grandes catégories ont été proposés (figure 1). L'**hypothèse archezoan** suppose que l'hôte serait une cellule proto-eucaryote qui présenterait déjà certains traits complexes des eucaryotes comme le noyau et la possibilité de phagocyter des proies, une caractéristique lui ayant permis d'ingérer la bactérie. En corollaire de ce scénario, il pourrait exister des descendants de cette cellule proto-eucaryote, qui n'auraient jamais eu de mitochondrie. L'**hypothèse par fusion** suppose au contraire, que l'acquisition des caractéristiques des cellules eucaryotes est contemporaine (ou successive) à la fusion de la bactérie avec un autre procaryote, probablement une archée. Selon cette théorie, on ne pourrait donc jamais trouver d'eucaryote antérieur à la fusion, puisqu'elle est l'élément fondateur de la lignée. Toutes les cellules eucaryotes connues ayant eu des mitochondries, et toutes les mitochondries étant monophylétiques, il est probable que l'endosymbiose ait été un évènement fondateur dans l'établissement du règne des eucaryotes [10].

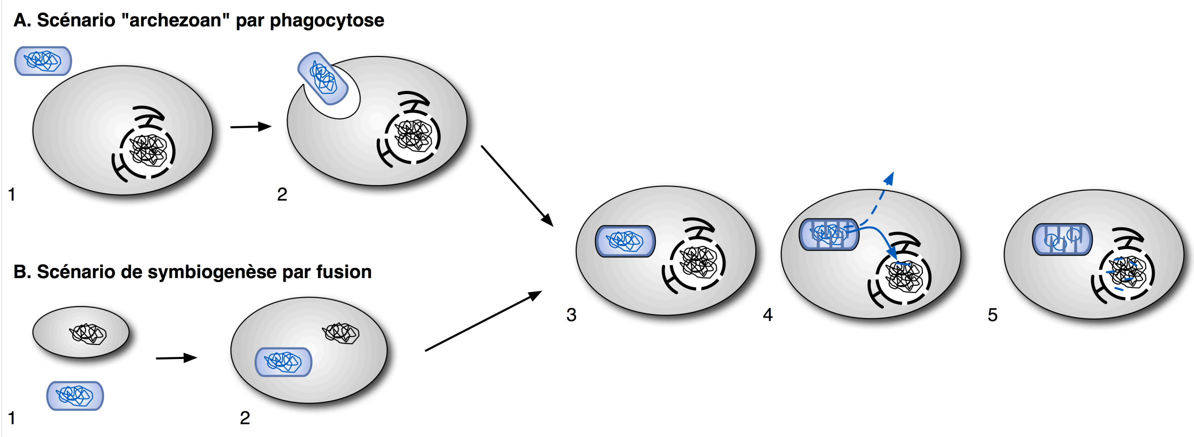

A. Scénario "archezoan" par phagocytose
1
2
B. Scénario de symbiogenèse par fusion
1
2
3
4
5

Figure 1. Schéma de l'origine évolutive des mitochondries. Deux principaux scénarios pour l'origine des mitochondries peuvent être envisagés. **A.** Dans le scénario dit "archezoan", la cellule hôte, probablement d'origine archée avait déjà de nombreuses caractéristiques typiques des eucaryotes, en particulier un noyau et la phagotrophie. Cette cellule hôte proto-eucaryote représentée en gris a phagocyté une alpha-protéobactérie en bleu. **B.** Dans le scénario par fusion, la bactérie et une archée fusionnent puis, secondairement, cette cellule développe les traits complexes des eucaryotes. L'endosymbiote perd des gènes dont certains sont éventuellement relocalisés dans le noyau de l'hôte.

La lignée des eucaryotes

Le domaine des eucaryotes montre une grande diversité. La classification taxonomique proposée par le comité international de protistologie [11] distingue cinq super-groupes illustrés sur la figure 2. Les Amorphea regroupent les **Opisthokonta** parmi lesquels on trouve les animaux et les champignons, et les **Amoebozoa**. Les diplonémides, notre groupe d'intérêt, sont dans le super-groupe des **Excavata**. Les **Archaeplastida** contiennent essentiellement des organismes photosynthétiques, ce qui comprend les plantes au sens large. Finalement, le super-groupe SAR regroupe les **Stramenopila, Alveolata** et **Rhizaria**. Certains groupes comme les Haptophyta et les Cryptophyta ont une place encore controversée dans cette classification. Dans ce manuscrit, le terme de "**protiste**" désigne l'ensemble des eucaryotes qui ne sont ni des plantes, ni des métazoaires, tandis que les "**microbes**" désignent les organismes unicellulaires, qu'ils soient eucaryotes ou procaryotes.

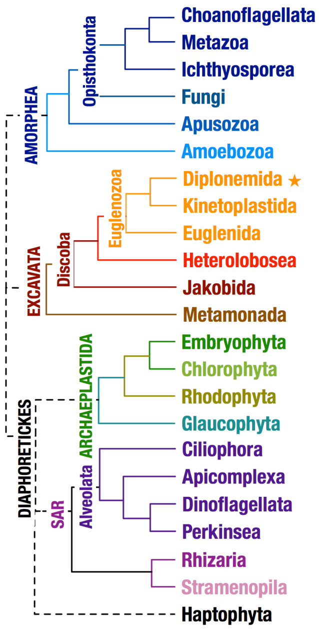


Figure 2. Classification des eucaryotes. Les diplonémides, notre groupe d'intérêt, sont indiqués avec une étoile. Nomenclature des super-groupes selon le comité international de protistologie [11]

Certains eucaryotes ont une mitochondrie très dérivée, ou absente

Plusieurs eucaryotes (comme *Trichomonas vaginalis*, *Giardia lamblia*, *Entamoeba histolytica*) ont été suspectés de ne pas avoir de mitochondrie car ils en ont une version très dérivée, difficilement reconnaissable, nommée MRO pour ***Mitochondrion-Related Organelles*** [12]. Sous ce terme sont regroupés l'**hydrogenosome** qui génère de l'ATP de façon anaérobie et utilise l'hydrogène au lieu de l'oxygène comme accepteur final d'électron, et les **mitosomes** qui ont complètement perdu leur rôle de production d'énergie [13], [14]. Les mitochondries dérivées conservent cependant leur rôle de genèse des centres Fer-Soufre via la machinerie ISC (*Iron-Sulfur Cluster machinery*). Certaines ont complètement perdu leur génome mitochondrial comme chez *Trichomonas vaginalis* et *Entamoeba histolytica*.

Un seul eucaryote a été découvert **sans mitochondrie**, l'oxymonade ***Monocercomonoides sp.***, mais son absence résulte d'une perte de l'organite et n'est pas un caractère ancestral [15]. Cet organisme, isolé de l'intestin d'un petit rongeur, fait partie du groupe des métamonades, groupe basal des Excavata, dont les membres anaérobies ou microaérobies ont des mitochondries dérivées. Le séquençage de l'ADN total n'a révélé aucune protéine typiquement mitochondriale comme la machinerie ISC. Cependant des homologues de la machinerie bactérienne de synthèse des centres Fer-Soufre (SUF) ont été identifiés. Les auteurs supposent que l'acquisition par transfert horizontal de la machinerie SUF aurait permis la perte secondaire de l'organite. Une analyse par microscopie électronique serait une preuve supplémentaire de l'absence de mitochondrie.

I.2. Structure du génome mitochondrial

La préconception de la structure circulaire des chromosomes mitochondriaux est le résultat de plusieurs facteurs confondants [16]. L'origine procaryote, d'abord, suggère un génome circulaire. L'obtention de molécules linéaires, ensuite, était attribuée à des artefacts techniques, d'autant que des molécules circulaires avaient été isolées avec succès chez les mammifères. La cartographie par fragments de restriction, enfin, dont les résultats cohérents avec un génome circulaire, peuvent en fait cacher d'autres structures comme des concatémères linéaires, difficiles à identifier expérimentalement [17], [18].

Des approches comme la microscopie électronique permettent de visualiser des molécules d'ADN circulaires ou des structures plus complexes comme le réseau intriqué du maxi-cercle et des mini-cercles chez les kinetoplastides ou les structures ramifiées de certains champignons. L'électrophorèse à champ pulsé permet de séparer les différents types de molécules, d'estimer leur taille et de déterminer leur super-enroulement [17], [18]. Finalement, le séquençage en masse de génomes mitochondriaux de protistes par des initiatives comme ***Organelle Genome Megasequencing Program*** (**OGMP**) ont permis de mieux apprécier la diversité génomique des mitochondries [19].

La taille des génomes mitochondriaux est généralement comprise entre 15 et 60 kb [2], avec de grandes variations autour de ces valeurs moyennes. La figure 3 illustre la diversité de taille et d'architecture des génomes mitochondriaux par quelques exemples choisis dans l'arbre des eucaryotes.

L'orthodoxe génome mitochondrial des mammifères et les variations chez les animaux

Tous les **mammifères** ont un génome circulaire d'environ 16 kb contenant 37 gènes [20]. Les autres **animaux** s'écartent parfois de cette uniformité. C'est le cas du génome mitochondrial des poux "suceurs" (*Sucking lice*) qui contient ces mêmes 37 gènes mais est fragmenté en plusieurs mini-chromosomes circulaires. Par exemple, le pou de tête *Pediculus* *humanus* a 18 mini-chromosomes d'une taille de 3 à 4 kb qui contiennent un à trois gènes et une région de contrôle [21], [22]. Le nématode *Globodera pallida* possède plusieurs chromosomes circulaires mosaïques contenant des fragments non fonctionnels de gènes, sans doute le résultat de recombinaisons [23]. Parmi les éponges et les cnidaires, certains ont un génome mitochondrial composé d'une ou plusieurs molécules linéaires [24].

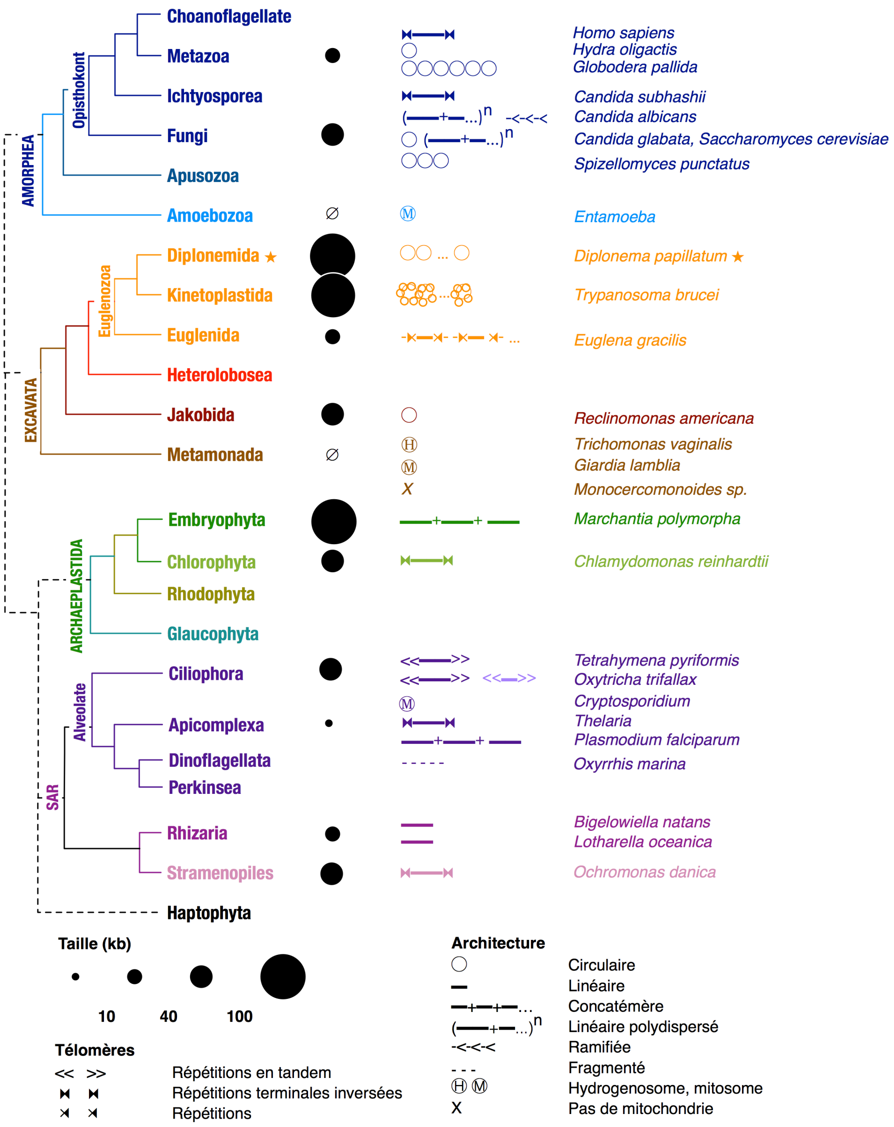


Figure3. Diversité de taille et de structure des génomes mitochondriaux. Les plasmides sont indiqués avec une couleur plus pâle.

La diversité des génomes mitochondriaux de champignons

L'analyse de la structure du génome de la levure *Saccharomyces cerevisiae* par électrophorèse à champ pulsé a montré des molécules circulaires correspondant à la taille attendue du génome et un assortiment de molécules linéaires de tailles variables correspondant à des **permutations circulaires** du génome [25]. Cet assortiment de molécules linéaires appelé **ADN linéraire polydispersé** est une architecture fréquemment retrouvée chez les champignons, par exemple chez *Candida glabatra* [25], [26], et serait la conséquence du type de réplication par cercle roulant. Les molécules sont en réplication continuelle et en recombinaison intense tout au long du cycle cellulaire [25]. Il a été proposé que la molécule circulaire correspond au génome hérité ou **génome maître**. *Candida albicans* possède une architecture un peu differente, essentiellement sans molécule circulaire, mais avec de l'ADN linéaire polydispersé, et une autre **structure très ramifiée** qui pourrait correspondre à des molécules en cours de réplication selon un processus dit amorcé par recombinaison (*recombination-driven replication*) [27]. Une troisième espèce de *Candida* n'a que des monomères linéaires terminés par des répétitions inversées, ce qui montre la grande diversité d'architecture, même au sein d'espèces proches [28], [29].

Les gigantesques génomes mitochondriaux des plantes

Chez les plantes, on trouve à la fois des molécules linéaires sous forme d'ADN polydispersé comme chez *Marchantia polymorpha* [30] et également de nombreuses molécules circulaires comme dans le gigantesque génome de 11.3 Mb de la plante à fleur *Silene conica,* réparti sur 128 chromosomes circulaires [31]. Malgré leur grande taille, les génomes de plante ne sont proportionnellement pas plus riches en gènes que les autres espèces, mais contiennent plus d'introns et de séquences répétées [32]. Les génomes mitochondriaux de plantes sont en interconversion constante entre plusieurs alternatives structurales, probablement à cause d'un taux de recombinaison élevé et de la présence de répétitions [32].

L'extraordinaire diversité et complexité des génomes de protistes

Chez les protistes, on trouve d’autres architectures en plus de celles précédemment décrites, peut être encore plus surprenantes. Les protistes constituent le groupe ayant la plus grande variation en taille du génome mitochondrial: de 6 kb chez *Plasmodium falciparum* [33] à plusieurs mégabases pour le complexe réseau mitochondrial des kinetoplastides, le groupe frère de notre organisme d'intérêt. Bien que la biodiversité des protistes soit difficile à estimer et soit encore un sujet très débattu [34], [35], il est probable que l'extraordinaire diversité de structure des génomes mitochondriaux ne soit que le reflet d'une biodiversité bien plus riche qu'attendue.

Chez les **dinoflagellés** par exemple, la structure du génome n'a pu être résolue car le séquençage révèle un large assortiment de molécules linéaires, mosaïques composées de gènes entiers et fragmentés dont la logique reste obscure [36]-[38].

Le groupe des **euglénozoa** dans lequel se trouve *Diplonema papillatum*, notre espèce d'intérêt, est sans doute celui qui contient les plus étonnantes structures [39]. Dans le groupe des **euglénides** à la base des euglénozoa, *Euglena gracilis* a un génome mitochondrial composé d'un ensemble de molécules linéaires de 5 à 8 kb contenant généralement un seul gène fonctionnel, et des fragments de gènes. Les chromosomes sont flanquées de régions répétées qui se ressemblent entre chromosomes et contiennent des séquences palindromiques [40], [41]. Les génomes mitochondriaux les plus étudiés des Euglenozoa sont ceux des parasites humains du groupe des **kinétoplastides**. Ils doivent leur nom à leur génome mitochondrial réticulé appelé kinetoplaste, lui même nommé ainsi car il forme un organite («-plaste») localisé à la base des flagelles permettant le mouvement («kineto-»). Le kinetoplaste est composé d'un grand chromosome de 20 à 50 kb appelé maxi-cercle, et de milliers de petits chromosomes ou mini-cercles qui peuvent être enchainés les uns aux autres en un complexe et élégant réseau [42]. Le maxi-cercle porte des gènes protéiques et ARNs dont les ARNm nécessitent une intense édition par addition ou délétion d'uridines pour être fonctionnels. Les mini-cercles codent pour des ARNs guides spécifiant la séquence correcte des portions de gènes édités [42]. Le troisième groupe taxonomique, les diplonémides, arbore aussi un génome fragmenté, mais d'une façon encore plus extrême car les gènes sont eux-mêmes en morceaux et portés par des chromosomes différents [43]-[45]. J'en ferai une description plus détaillée dans la dernière section de l'introduction.

3. Contenu génique des mitochondries

Le génome de l'endosymbiote fondateur a considérablement réduit, soit par perte définitive de gènes devenus fonctionnellement inutiles ou redondants, soit par re-localisation des gènes dans le génome nucléaire de l'hôte. Les nombreux pseudogènes mitochondriaux observés dans les génomes nucléaires attestent de ces transferts. Rares sont les exemples de transfert de gènes nucléaires vers la mitochondrie. Ainsi, le contenu génique actuel des mitochondries est essentiellement le reliquat de cette fuite génique [46], [47].

Nombre et fonction des gènes mitochondriaux

Le contenu génique des mitochondries varie d'un ordre de magnitude entre les minuscules génomes des Apicomplexa et des dinoflagellés contennant deux ou trois gènes protéiques, et les génomes des jakobides réputés les plus "ancestraux" des génomes mitochondriaux [48], [49]. Le jakobide *Andalucia godoyi* avec ses 100 gènes est l'espèce la plus riche en gènes [49]. Nous avons vu que la taille du génome n'est pas corrélée au contenu en gènes. Ainsi le gigantesque génome de la plante à fleur *Silene conica* contient seulement 25 gènes protéiques sur plus de 11 Mb de génome car 40% de son génome est composé de séquences répétées [31].

Le génome mitochondrial encode systématiquement les gènes des ARNs de la petite et grande sous-unité ribosomique (*rns*, *rnl*), et les gènes protéiques de la coenzyme Q-cytochrome c réductase (*cob*) et de la sous-unités 1 de la cytochrome c oxidase (*cox1*). Entre ce jeu minimal et le contenu maximal du génome des jakobides, les génome mitochondriaux sont constitués d'un assortiment variable de gènes. On trouve fréquemment des gènes de la chaîne de transport des électrons, des gènes de l'ATP synthase, d'ARN de transfert et de protéines ribosomiques (Figure 4).

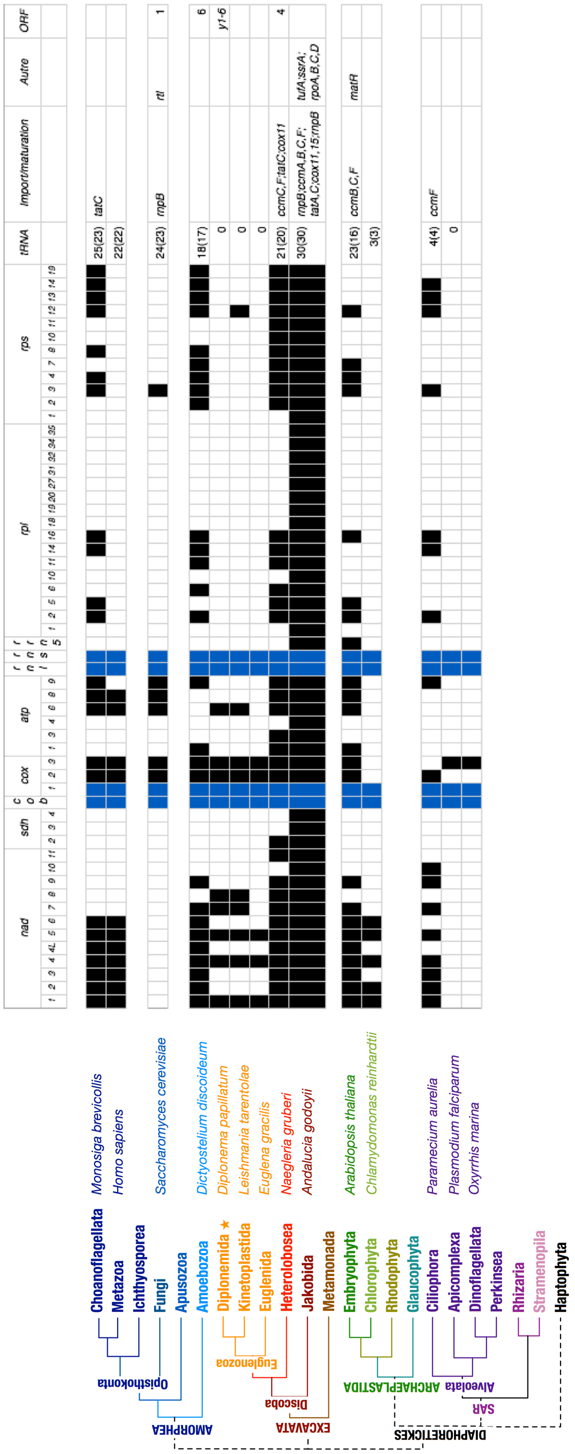


Figure 4. Contenu génique des mitochondries pour une sélection d'espèces. *atp*: ATP synthase, *nad*: NAD deshydrogénase, *cob*: Cytochrome b, *cox*: Cytochrome oxydase, *rns*: petite sous-unité ribosomique, *rnl*: grande sous-unité ribosomique, *trn*: ARN de transfert. D'après http://amoebidia.bcm.umontreal.ca/pg-gobase/searches/compilations.php

Structures géniques exceptionnelles

Chevauchement de gènes

L'apparente orthodoxie du génome mitochondrial humain cache quelques bizarreries. En particulier, les 46 dernières paires de base du gène *atp8* chevauchent le début du gène *atp6*, ces gènes codant pour deux sous-unités du complexe V de l'ATP synthase. De même, *nad4L* et *nad4*, deux gènes codant pour des sous-unités du complexe I se chevauchent sur sept paires de base [50].

Fission de gènes

Les gènes sont généralement transférés en entier vers le noyau, mais des cas de transfert partiel existent [47]. Chez le chlorophyte *Chlamydomonas reinhardtii* le gène *cox2* a été scindé en deux gènes *cox2a* et *cox2b*, tous deux transférés vers le noyau [51]. La même situation est retrouvée chez les Apicomplexa comme *Plasmodium falciparum*, deux évènements qui malgré leur rareté se sont révélés indépendants [52]. Chez le chlorophyte *Scenedesmus*, le transfert est dans un état intermédiaire: *cox2a* est encore dans la mitochondrie tandis que *cox2b* est déjà transféré vers le noyau [52]. Le gène mitochondrial *nad1* chez les ciliés *Tetrahymena pyriformis* et *Paramecium aurelia* est scindé en deux morceaux qui ne sont pas épissés en trans [53]. Dans tous ces exemples, des analyses protéomiques seraient nécessaires afin de vérifier si les fragments sont assemblés au niveau protéique, comme celà existe avec les intéines [54].

Les **gènes ARNs** peuvent également être morcelés. Un cas intéressant est la fragmentation sans épissage des gènes d'**ARN ribosomiques** que l'on trouve très communément chez les alvéolés (ciliés, dinoflagellés, apicomplexans) [36], [53], [55], [56] et chez d'autres espèces comme *Euglena gracilis* [40], *Chlamydomonas reinhardtii* [57]. Reconstituer la séquence complète de l'ARN ribosomique est un défi, car les fragments sont courts, les gènes sont très divergents et présentent souvent un biais important de composition (taux de G et C bas).

Fusion de gènes

Chez les Amoebozoa *Dictyostelium discoideum et Acanthamoeba castellanii*, les gènes *cox1* et *cox2* ont été trouvés fusionnés [58]. Dans le cas d'*Acanthamoeba*, l'analyse par Western Blot a cependant montré que la protéine Cox2 n'était pas produite sous forme de protéine de fusion [58].

Chez le dinoflagellé *Oxyrrhis marina*, trois gènes protéiques seulement sont encodés par le génome dont deux, *cob* et *cox3*, sont fusionnés [36]. Bien que les deux gènes soient monocistroniques, c'est-à-dire transcrits sous la forme d'une seule molécule, il n'est pas encore démontré qu'une protéine de fusion soit produite et fonctionnelle dans les deux complexes différents de la chaine respiratoire dans lesquels ils interviennent.

4. Protéome mitochondrial et fonctions

Alors qu'une poignée de gènes seulement est présente sur le génome mitochondrial, la mitochondrie participe à un large éventail de fonctions cellulaires. Ces fonctions sont assurées essentiellement par des protéines encodées par le génome nucléaire puis importées secondairement dans la mitochondrie [59].

Le protéome mitochondrial : un coeur commun et une mosaïque de protéines variées

La composition du protéome mitochondrial peut être inférée par une combinaison d'approches expérimentales comme la spectrométrie de masse en tandem (MS/MS) sur des extraits purs de protéines mitochondriales, et de prédictions bio-informatiques basées sur la reconnaissance de séquences d'adressage mitochondrial ou de biais de composition des protéines. Le protéome mitochondrial est composé en moyenne d'environ 1000 protéines, mais des espèces peuvent présenter un protéome bien plus réduit telle *Tetrahymena thermophila* avec 573 protéines prédites par MS/MS [60]. Chez l'homme et la souris, 1158 protéines composent le protéome mitochondrial et 96% sont communes aux deux ensembles [61]. Pris individuellement, les protéomes sont des assemblages éclectiques composés de 10 à 20% seulement de protéines reliées aux alpha-protéobactéries et composant le coeur commun, 20% environ sont d'origine procaryotique, 40% d'origine eucaryote et 25% sont uniques à l'organisme en question [59]. Ces derniers 25% de proteines uniques peuvent cependant être trop dérivés pour être reconnaissables.

Tableau I. Nombre de protéines composant le protéome mitochondrial chez une sélection d'espèces.

Espèce	**Nombre de protéines mitochondriales prédites**	**Référence**
Acanthamoeba castellanii	1033	Gawryluk, 2014 [62]
Arabidopsis thaliana	843	Lee, 2013 [63]
Homo sapiens	1158	Calvo, 2016 [61]
Mus musculus	1158	Calvo, 2016 [61]
Saccharomyces cerevisiae	851	Reinders, 2006 [64]
Tetrahymena thermophila	573	Smith, 2007 [60]
Trypanosoma brucei	1008	Panigrahi, 2009 [65]

Les mitochondries ont un large panel de fonctions

Les fonctions cellulaires assurées par les mitochondries ont été étudiées expérimentalement chez les organismes modèles. Pour les autres espèces les fonctions mitochondriales peuvent être inférées par analyse des protéomes. La comparaison révèle des fonctions partagées par la plupart des eucaryotes [59], [66]. Parmi celles-ci, on retrouve les fonctions liées aux processus informationnels (réplication, transcription et traduction) et à la maintenance des mitochondries (fission). Le rôle principal des mitochondries est la production d'énergie sous forme d'ATP par phosphorylation oxydative grâce aux complexes protéiques de la chaîne respiratoire [67]. Elles jouent un rôle central dans la biosynthèse des protéines ayant un centre Fer-Soufre, et la biosynthèse de l'hème. Les mitochondries participent aussi à l'équilibre ionique de la cellule en régulant les stocks de fer et de calcium cellulaire [68]. Pour assurer l'import et l'export de composants cellulaires (protéines, ARN, métabolites) la mitochondrie possède un arsenal élaboré de transporteurs cellulaires.

II.Expression de l'information génétique mitochondriale

1. Transcription

Initiation de la transcription

Nous avons vu auparavant que le génome mitochondrial des **mammifères** est circulaire, petit et très dense en gène. La plus longue région intergénique appelée boucle de déplaçement ou *D-loop* sert de région de contrôle pour le démarrage de la réplication et de la transcription. Un des brins du génome mitochondrial des métazoaires contient les gènes protéiques et les ARN ribosomiques (28 gènes sur 37) et est appelé brin lourd (*heavy*), les autres gènes, des gènes d'ARN de transfert, sont codés sur le brin dit léger (*light*). Dans la *D-loop* deux promoteurs transcrivent les ARN ribosomiques et l'ensemble du brin lourd tandis qu'un troisième promoteur adjacent aux deux autres transcrit le brin léger. Les gènes sont transcrits sous forme de longues molécules polycistroniques [69].

Chez les **levures** comme chez les **plantes**, la transcription débute à plusieurs endroits mais des ARN polycistroniques sont également produits [70]. Les promoteurs mitochondriaux chez les levures sont des motifs de 9 pb. Le génome mitochondrial de *Saccharomyces cerevisiae* contient 28 promoteurs tandis que celui *Schizosaccharomyces pombe*, plus petit (19 kb) n'a que deux promoteurs [71], [72]. Chez les plantes, on trouve aussi des ARN monocistroniques [73]. La présence de plusieurs promoteurs pourrait être liée à la plus grande taille des génomes. Pourtant, avec un génome mitochondrial de 56 kb beaucoup plus grand que celui des mammifères, l'**amibe** *Dictyostelium discoideum* a une seule région de démarrage de la transcription [74]. La présence de plusieurs promoteurs pourrait également assurer la pérénité de la transcription malgré des réarrangements chromosomiques comme cela a été proposé pour les plantes [73].

ARN polymérase

La transcription est réalisée quasi-ubiquitairement par une machinerie composée essentiellement d'une ARN polymérase, d'un facteur d'initiation de la transcription et d'un facteur de terminaison de la transcription, tous encodés par le génome nucléaire [69], [72], [73]. L'**ARN polymérase** est homologue de celle des **phages T3 et T7** [75]. Contrairement à celle des phages, l'ARN polymérase mitochondriale nécessite des facteurs additionnels d'initiation de la transcription pour se fixer spécifiquement aux promoteurs. Une exception notable a été découverte chez les jakobides comme *Reclinomonas americana* et *Andalucia godoyi.*

Leur génome mitochondrial, par ailleurs le plus riche en gène, encode quatre sous-unités d'une **ARN polymérase bactérienne** [48], [49]. Le remplacement de l'ARN polymérase bactérienne par une ARN polymérase phagique est un exemple d'un transfert horizontal subséquent à l'endosymbiose [59].

Transcrits polycistroniques

Dans les mitochondries des mammifères, les transcrits polycistroniques sont clivés co-transcriptionnellement par des endonucléases. Cette étape est réalisée par excision des ARN de transfert qui flanquent chaque gène, un système de maturation appelé "**ponctuation par ARNt**" [76]. En 5', la coupure est réalisée par la RNase P, et en 3' par la RNase Z [77], [78]. Chez les levures, les transcrits sont également séparés par l'excision des ARN de transfert ou par le début de la transcription. Les extrémités 3' sont en plus clivés au niveau d'un motif riche en C chez *Schizosaccharomyces pombe* et d'un dodécamer chez *S. cerevisiae* [71], [79], [80].

2. Introns et épissage

Les introns sont des séquences interrompant la portion codante d'un gène et excisées post-transcriptionnellement. Le type d'épissage le plus répandu dans le noyau est l'**épissage "spliceosomal"** réalisé par l'énorme machinerie ribonucléoprotéique qu'est le spliceosome. D'autres types existent dont le mécanisme est élucidé: les introns de groupe I (noyau, organites), de groupe II (organites) et les introns ARNt ou archée (noyau). Enfin, des cas non élucidés chez les dinoflagellés et chez les diplonémides sont rapportés [81]. À chaque type d'intron correspond une machinerie spécifique. Je les passerai en revue dans cette section en soulignant ceux que l'on retrouve dans les mitochondries.

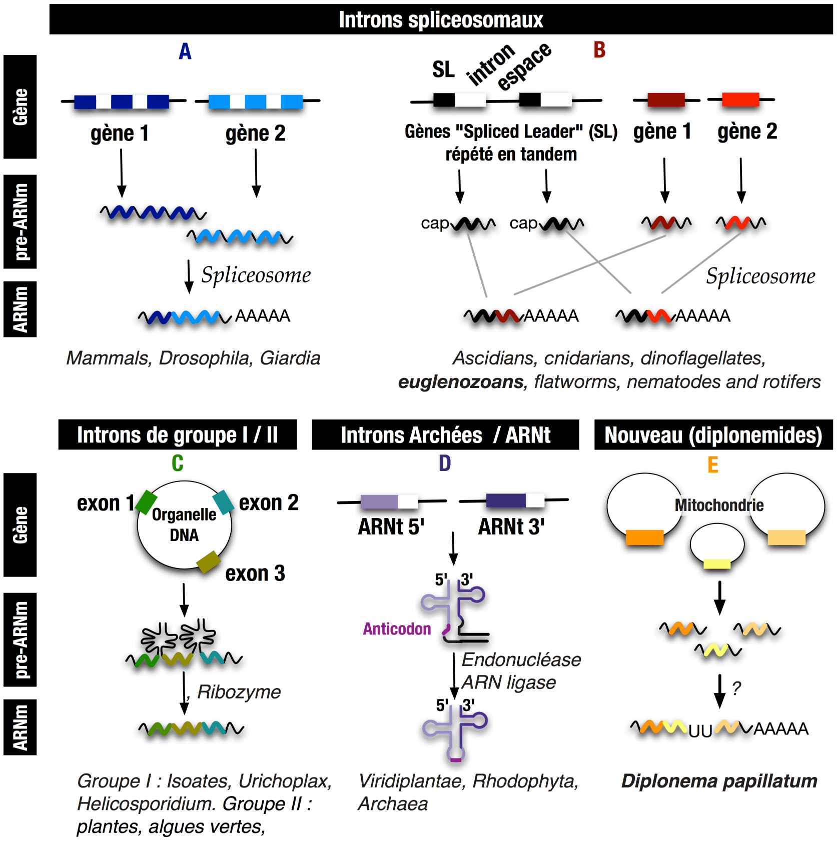


Figure 5. Cas d'épissage en trans classés par type de machinerie.

L'épissage en trans, universel mais rare

Dans la grande majorité des cas, les exons d'un gène sont contigus sur un chromosome et co-transcrits. Il existe toutefois des cas où des exons transcrits sur des molécules différentes sont épissés, un phénomène appelé **épissage en trans**. Pour tous les types d'intron connus, il existe des cas d'épissage en trans (figure 5). Il semble qu'une fois la machinerie d'épissage installée, son fonctionnement en trans plutôt qu'en cis ne soit pas un obstacle majeur. Néanmoins, les cas d'épissage en trans sont beaucoup plus rares que ceux en cis [82].

Les classiques introns spliceosomaux et le cas des séquences *leader* épissées en trans

Le **spliceosome** est l'une des plus grosse machineries cellulaire. Près de 200 composants ARN et protéiques composent le spliceosome chez l'homme, soit ~1% des gènes. Les introns spliceosomaux ne sont pas trouvés dans les organites. Il est possible que la nécessité d'importer une telle machinerie cellulaire soit un frein à la dissémination des introns spliceosomaux dans les organites. Un type particulier d'épissage en trans d'introns spliceosomaux est une séquence *leader* épissée en trans ou *Spliced-Leader trans-splicing* (figure

5B). Les ***spliced-leaders*** (SL) sont de petits gènes nucléaires répétés en tandem, retrouvés à travers l'arbre eucaryote: chez des animaux comme le nématode *Caenorhabditis elegans*, chez les trois groupes d'euglénozoaires, et chez les dinoflagellés. Ils sont composés d'un mini-exon de 16 à 50 pb et d'un intron. Après transcription, le mini-exon est épissé en trans en amont des gènes, apportant parfois le signal de démarrage de la traduction. La totalité des transcrits peuvent être précédés de cette séquence comme chez *Trypanosoma brucei* ou aussi peu que 1% chez le nématode *Trichinella spiralis.* La portion de type intronique en 5' des gènes porte le nom d'*outron*. Les SL peuvent également avoir pour fonction de scinder les ARNs polycistroniques par épissage entre chaque cistron (région codante). La séquence du SL des diplonémides est identifiée et mesure 39 pb (figure 6) [82].

```
AACCAACGATTTAAAAGCTACAGTTTCTGTACTTTATTG
```

Figure 6. Séquence du Spliced Leader de *Diplonema papillatum*

Les introns catalytiques de groupe I et II

Les seuls types d'intron caractérisés dans les organites sont les introns de groupe I et II. Ces introns sont surtout abondants dans les organites de champignons, plantes et chez certains protistes. On les trouve dans les gènes d'ARNr, d'ARNt ou de protéine. Les introns de groupe II ne sont jamais présents dans le noyau des eucaryotes alors que l'on trouve des introns de groupe I dans les ARNr nucléaires. Ces deux types d'introns sont des éléments génétiques mobiles. Ils adoptent une structure secondaire caractéristique de leur groupe et sont tous deux des ribozymes auto-catalytiques bien que des protéines accessoires soient nécessaire pour leur excision. Malgré ces similitudes, leur structure secondaire et les étapes de la catalyse sont très différents et justifient de les séparer en deux groupes distincts [81].

Les introns ARNt et leur machinerie exclusivement protéique

Les introns ARNt ou d'archées sont les seuls épissés par une machinerie exclusivement protéique. Nécessitant seulement une endonucléase et une ARN ligase, ils ne sont pourtant trouvés que chez les archées et dans les noyaux eucaryotes, et n'ont encore jamais été mis en évidence dans les organites. Malgré leur nom, ils ne se trouvent ni exclusivement dans les gènes d'ARNt ni exclusivement chez les archées.

Les cas d'épissage non élucidés

Pour quelques cas, le mécanisme d'épissage reste inconnu. Nous verrons le cas des diplonémides dans la dernière section de l'introduction. L'autre cas non élucidé d'épissage en trans est celui du gène *cox3* scindé en deux fragments chez plusieurs dinoflagellés [83]. Les transcrits des deux fragments sont polyadénylés et plusieurs adénines (jusqu'à 10 chez *Symbiodinium* sp.) sont retenues dans le transcrit mature à la jonction entre les deux moitiés.

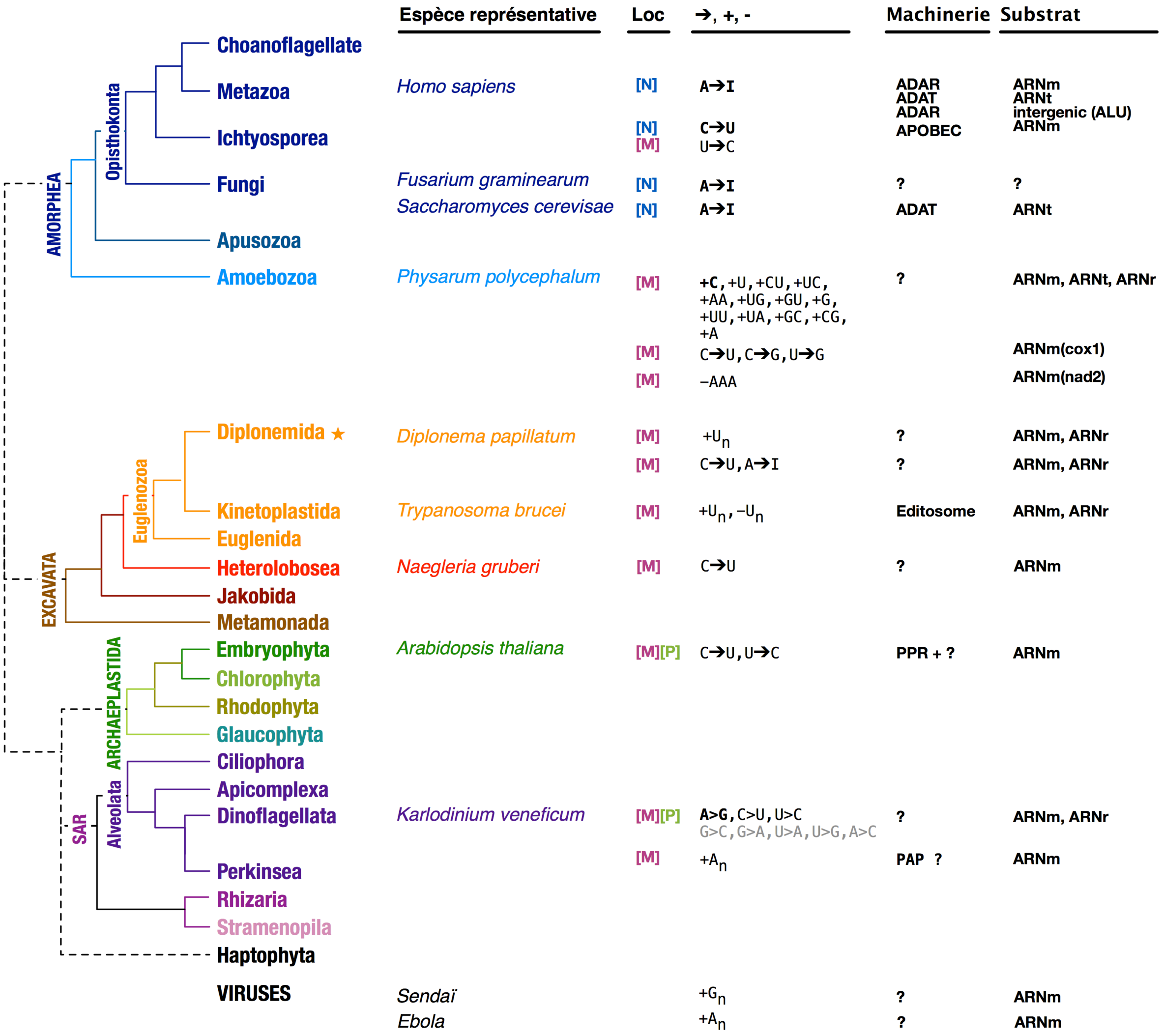
3. Edition

Dit simplement, l'édition est la modification de la séquence codante d'un gène au niveau ARN. Cette définition recouvre une panoplie de modifications que l'on peut classer en deux catégories: les modifications ponctuelles qui changent un nucléotide en un autre, ci-après nommée **édition par substitution**, et l'insertion ou la délétion d'un nombre variable de nucléotides que l'on nommera **édition par indel**. Un cas particulier concerne les modifications des nucléotides des ARN de transfert et ribosomiques, qui sont plutôt considérés comme des processus de maturation post-transcriptionnels et qui seront décrits succinctement.

L'édition peut être d'étendue variable, de quelques nucléotides à une centaine de positions, et concerner tous les ARNs ou seulement quelques-uns. C'est un phénomène exclusivement eucaryote, si l'on exclu les modifications des ARN structuraux, avec une distribution phylogénétique sporadique. D'autre part, bien que l'édition d'ARN existe dans le noyau, elle est cependant plus répandue et diversifiée dans les organites où elle a été mise en évidence pour la première fois chez les trypanosomes [84]. Nous verrons dans un premier temps les cas d'édition dans les organites puis ceux observés dans les noyaux (figure 7).

Il est important de souligner que pour **mettre en évidence des évènements d'édition** parfois rares, il faut comparer la séquence génomique avec une couverture suffisante de séquences de transcrits ; une analyse que l'on pouvait difficilement réaliser avant l'avènement du séquençage ARN à haut débit. Ainsi leur apparente rareté ou absence dans certaines taxons est peut être d'avantage dû à un manque d'investigation qu'une réelle preuve de non-existence. Un autre point à considérer est que l'édition est parfois régulée par les conditions environnementales [85] ou le stade de développement comme chez le champignon *Fusarium graminearum* [86], ce qui complique encore leur identification. Enfin, les efforts de séquençage sont biaisés vers certains groupes taxonomiques (métazoaires, champignons, plantes terrestres)

[87]. Ainsi, il est probable que l'avenir nous réserve encore de belles découvertes dans le domaine de l'édition d'ARN.



	Espèce représentative	Loc	→, +, -	Machinerie	Substrat
Choanoflagellate					
Metazoa	*Homo sapiens*	[N]	A→I	ADAR ADAT ADAR	ARNm ARNt intergenic (ALU)
		[N]	C→U	APOBEC	ARNm
Ichtyosporea		[M]	U→C		
Fungi	*Fusarium graminearum*	[N]	A→I	?	?
	Saccharomyces cerevisae	[N]	A→I	ADAT	ARNt
Apusozoa					
Amoebozoa	*Physarum polycephalum*	[M]	+C,+U,+CU,+UC, +AA,+UG,+GU,+G, +UU,+UA,+GC,+CG, +A	?	ARNm, ARNt, ARNr
		[M]	C→U,C→G,U→G		ARNm(cox1)
		[M]	−AAA		ARNm(nad2)
Diplonemida ★	*Diplonema papillatum*	[M]	$+U_n$	?	ARNm, ARNr
		[M]	C→U,A→I	?	ARNm, ARNr
Kinetoplastida	*Trypanosoma brucei*	[M]	$+U_n$,$-U_n$	Editosome	ARNm, ARNr
Euglenida					
Heterolobosea	*Naegleria gruberi*	[M]	C→U	?	ARNm
Jakobida					
Metamonada					
Embryophyta	*Arabidopsis thaliana*	[M][P]	C→U,U→C	PPR + ?	ARNm
Chlorophyta					
Rhodophyta					
Glaucophyta					
Ciliophora					
Apicomplexa					
Dinoflagellata	*Karlodinium veneficum*	[M][P]	A>G,C>U,U>C G>C,G>A,U>A,U>G,A>C	?	ARNm, ARNr
		[M]	$+A_n$	PAP ?	ARNm
Perkinsea					
Rhizaria					
Stramenopila					
Haptophyta					
VIRUSES	*Sendaï*		$+G_n$	?	ARNm
	Ebola		$+A_n$	?	ARNm

Figure 7. Les cas d'édition d'ARN connus. Loc, localisation; N, noyau; M, mitochondrie; P, plastide; ->, édition par substitution; +, édition par addition; -, édition par délétion; ?, machinerie partiellement ou totalement inconnue

Aux limites de l'édition: les modifications post-transcriptionnelles des ARNt et des ARNr

Pour les **ARN de transfert**, une queue CCA est ajoutée à l'extrémité 3' et un 'G' est ajouté en 5' des ARNt^{His}, lorsqu'ils ne sont pas encodés par le génome. On dénombre environ 85 autres modifications chimiques des nucléotides, dont la majorité se trouve aux positions 34 et 37, dans la boucle de l'anticodon [88]-[90]. Chez la levure, on retrouve 25 de ces modifications [91]. La déamination de l'adénosine 34 en Inosine par des Adénosines

Déaminases (ADAT) qui permet la reconnaissance flottante de différents codons ("Wobble position"), est à la limite entre ce qui est considéré comme de l'édition et de la maturation post-transcriptionnelle [92]. Les modifications des nucléotides dans les **ARN ribosomiques** sont principalement de trois types: la conversion d'uridine en pseudouridine, la méthylation du ribose en 2' et la méthylation des bases à diverses positions. Chez la levure *Saccharomyces cerevisiae*, on observe respectivement 44, 55 et 10 occurrences des modification pré-citées [93].

L'édition dans les organites

C-vers-U et U-vers-C dans les mitochondrie et les plastides des plantes

Ces deux types d'édition concernent plusieurs centaines de sites dans les deux organites des plantes [94]. L'édition de C-vers-U procède par déamination ; le mécanisme de l'édition inverse pourrait être une déamination ou une transamination. Si l'enzyme catalysant l'édition n'est pas encore identifiée, on sait néanmoins que les sites sont spécifiés par des protéines de la grande famille des **pentatricopeptides** (PPR). Ces protéines sont impliquées dans une large gamme de fonctions liées au traitement des ARNs et sont très abondantes chez les plantes (plus de 400 chez *Arabidopsis thaliana*) [95]. Les PPRs sont composées de répétitions en tandem de courts motifs de 35 acides aminés environ qui adoptent une structure secondaire en hélice-coude-hélice dont les acides aminés 2, 4 et 35 établissent des interactions avec une base sur l'ARN [96]. Bien que dégénérée, cette correspondance a permis de prédire, dans certains cas, le motif ciblé sur l'ARNm [97]. Chez les plantes, les PPR impliquées dans l'édition d'ARNm sont composées par la succession de trois types de motifs, le type classique P suivi d'un motif long L, puis un motif court S. Elles contiennent en position C-terminale des domaines appelés E1 et E2 qui sont des versions dégénérées de motifs PPR et du domaine DYW respectivement [96].

Des substitutions variées dans les mitochondrie et les plastides des dinoflagellés

Pratiquement tous les types de substitutions possibles sont observés dans les organites des dinoflagellés, avec une prévalence de l'édition de A-vers-G, présente même chez les espèces où l'édition est la plus rare. Le nombre de sites édités peut être très élevé, atteignant 6% des positions pour le gène *cox3* de *Karlodinium*. Les positions éditées sont groupées et certains sites sont très conservés entre dinoflagellés [38]. Le mécanisme et les machineries sont inconnus. La présence dans le génome de fragments de gène pourrait permettre la transcription d'ARNs guides comme c'est le cas chez les kinetoplastides. Ces ARNs n'ont cependant pas encore été trouvés [98]. Parmi les Alveolata, on ne détecte pas d'édition dans les Apicomplexans et les Ciliés [38]. Ces modifications ne sont pas observées chez les dinoflagellés les plus basaux comme *Oxhirris marina*, suggérant une acquisition postérieure à leur divergence.

Dans plusieurs taxa, le gène *cox3* est divisé en deux parties, transcrites indépendamment, polyadénylées, puis épissées en trans (voir partie II.2) en retenant quelques adénosines entre les deux fragments [83]. Ce phénomène est absent pour le gène fusionné *cob-cox3* d'*Oxhirris marina*. Chez ce dernier, on observe une polyuridylation en 5' des transcrits [36].

+U/-U dans les mitochondries des kinétoplastides

Chez les kinetoplastides, l'intense édition des ARNs par ajout ou délétion d'uridine est l'une des mieux étudiée, et le premier type d'édition identifié [84] (figure 8).

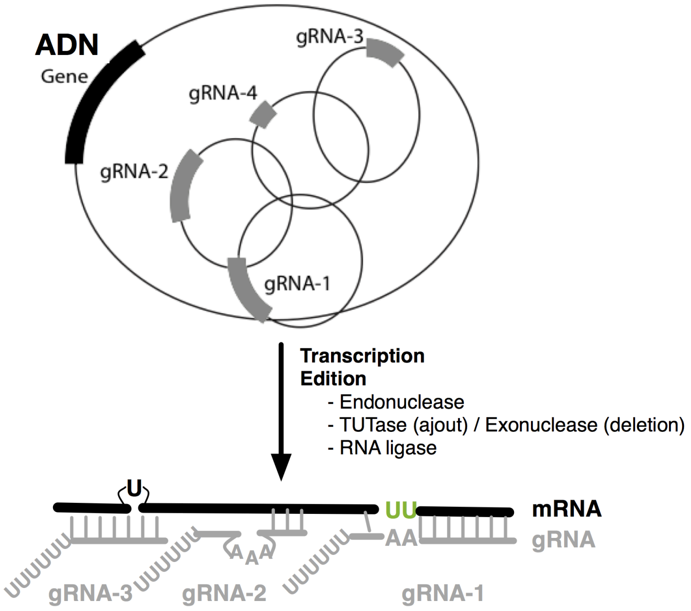


Figure 8. Edition chez les kinétoplastides. Le chromosome maxi-cercle en noir code pour les gènes ARN et protéiques. Les chromosomes mini-cercles en gris codent pour les ARN guides (notés gRNA) qui s'apparient avec la région à éditer dans l'ARNm et spécifient le nombre d'uridines à ajouter ou supprimer. Le coeur de la machinerie d'édition, l'éditosome, est composé d'une endonucléase (qui coupe le transcrit), d'une TUTase (pour l'ajout d'uridine) ou d'une exonucléase (pour la suppression d'uridine) et d'une ARN ligase (qui répare la coupure après édition).

L'édition, dirigée par des ARN guides anti-sens, procède de l'extrémité 3' vers l'extrémité 5' du transcrit. Les ARNs guides s'hybrident en 3' de la portion d'ARN à éditer par une région complémentaire. Lorsqu'une position non-appariée est atteinte, une endonucléase coupe l'ARN. Dans le cas d'insertion d'uridines, une **Terminal-Uridyl-Transferase** (TUTase), ajoute autant d'uridines que spécifié par l'ARN guide. Dans les cas moins fréquents d'édition par suppression, c'est une **exonucléase** qui les excise. La brèche est ensuite fermée par une **ARN ligase**. Le degré d'édition est variable selon les ARNs et les espèces. Chez *Trypanosoma brucei*, certains ARNs ne sont pas du tout édités (*nad1*, *nad4*, *nad5* et *cox1*) alors que d'autres le sont intensivement comme *nad7* avec 553 insertions et 89 suppressions. Les ARN ribosomiques sont simplement poly-uridylés en 3'[99].

Edition par substitution et insertion chez les myxomycetes

L'édition chez les myxomycètes (*slime-molds*) comme *Physarum polycephalum* est sans doute la plus diversifiée. L'insertion de nucléotide est prépondérante, en particulier l'insertion de C qui représente presque 95% des cas d'édition, mais on observe aussi l'insertion de A, U, G et de dinucléotides. Dans un système de transcription *in vitro*, diminuer la concentration du nucléotide suivant le point d'insertion (ce qui ralentit la polymérase) aboutit à un taux d'édition plus important et montre que l'édition par insertion procède par un mécanisme co-transcriptionnel. Le nucléotide additionnel est ajouté à l'extrémité 3' du transcrit naissant [100]. À l'opposé, l'édition par substitution est beaucoup plus rare (6 sites chez *Physarum polycephalum*) et est post-trancriptionnelle [101].

L'édition dans les noyaux

C-vers-U, APOBEC

Seulement deux types d'édition sont connus dans les noyaux. Un premier type, rare, est la déamination de Cytidine en Uridine (C-vers-U). Identifiée originellement dans le gène de l'apolipoprotéine B chez les mammifères [102], l'édition conduit à une protéine tronquée par la conversion d'un codon glutamine (CAA) en codon STOP (UAA). Le gène *apoB* donnera son nom à l'enzyme catalysant la réaction, APOBEC [103]. Le site à éditer est préférentiellement situé dans une région riche en A et U et une séquence de 11 nt située ~5 nucléotides en aval (***Mooring sequence***) permet la fixation de la protéine ACF (*APOBEC1 complementation factor*) qui recrute APOBEC1. D'autres enzymes de la famille APOBEC ont été ultérieurement identifiées mais celles-ci semblent limitées aux vertébrés et éditent préférentiellement l'ADN double brin [104]. Les enzymes de la famille APOBEC appartiennent à la grande famille des Déaminases zinc-dépendantes qui comprennent des enzymes impliquées dans le métabolisme des purines et des pyrimidines, et les adénosines déaminases agissant sur les ARNt (ADAT) ou les autres ARNs (ADAR) [105].

A-vers-I, ADAR

Il s'agit du type d'édition le plus répandu dans les noyaux des animaux. Les enzymes **ADAR** (*Adenosine Deaminase Acting on RNA*) déaminent l'Adénosine en Inosine (A-vers-I) qui est interprétée comme une Guanosine lors de la traduction ou durant le séquençage. Les adénosines déaminases agissant sur les ARNt (ADAT) qui changent la position flottante (*wobble*) de l'anti-codon appartiennent à cette même famille [90]. Les enzymes de la famille ADAR agissent préférentiellement sur des ARNs double brin tandis que les ADAT ciblent les boucles des épingles à cheveux comme c'est le cas pour la boucle de l'anti-codon de l'ARN. Alors que les ADAT sont largement répandues chez les eucaryotes, les ADAR sont une innovation des métazoaires. Elles seraient survenues par duplication d'un gène ADAT puis recrutement par l'un des paralogues d'un domaine de liaison à l'ARN double brin chez l'ancêtre des métazoaires. La famille s'est ensuite diversifiée par duplication des domaines de liaison à l'ARN et/ou recrutement d'un nouveau domaine de liaison à l'ADN en conformation Z [106].

L'édition chez les virus

Le virus de l'hépatite D a un génome ARN négatif codant pour une seule protéine. Sur l'ARNm dérivé, il utilise l'ADAR1 de son hôte, l'homme, pour déaminer un **A en I** dans le codon UAG, ce qui change le codon STOP pour le tryptophane et permet de produire une protéine plus longue. La protéine longue est impliquée dans l'assemblage des protéines virales tandis que la version courte est impliquée dans la réplication du virus [107]. L'**addition co-transcriptionnelle de un à six G** chez les paramyxovirus (comme le virus de la rougeole ou de Sendaï) et de un A chez le virus Ebola par bégaiement (*stuttering*) de l'ARN polymérase provoque un changement de cadre de lecture et la production de différentes protéines. Le site d'insertion est déterminé par une séquence en amont de type U_nC_n où le C est relu plusieurs fois par l'ARN polymérase. Ce mécanisme est également utilisé par les virus pour synthétiser la queue poly-A des transcrits [108]-[110].

Autres cas dont la machinerie n'est pas identifiée

L'édition massive de **A-vers-G?** dans le noyau d'un **champignon filamenteux** pendant la reproduction sexuée est une première en dehors des métazoaires. Ce champignon ne possède pas d'ADAR mais trois paralogues d'ADAT qui ne sont cependant pas exprimés significativement durant la reproduction sexuée. Notons que ces modifications sont rapportés par les auteurs comme des éditions de A-vers-I bien que la démonstration de la présence d'inosine à la place de la guanosine détectée par le séquenceur n'ai pas été faite [86].

Deux sites d'édition de **C-vers-U** dans les gènes *cox1* et *cox3* ont été répertoriés dans les mitochondries de l'heterolobosean ***Naegleria gruberi*** [111].

4. Polyadénylation, polyuridylation des ARNs mitochondriaux (mt-ARN)

Dans le noyau des **eucaryotes**, la polyadénylation joue un rôle **stabilisateur** des ARN messagers et stimulateur de la transcription. Chez les **bactéries**, il s'agit plutôt d'un **signal de dégradation** [112].

Les transcrits mitochondriaux ne sont généralement pas **polyadénylés**. Pourtant chez les mammifères, la plupart des transcrits sont polyadénylés par une poly(A) polymerase mitochondriale spécifique, ce qui reconstitue fréquemment le codon stop. Dans les mitochondries humaines, un raccourcissement de la queue poly-A est associée à une durée de vie plus courte des mt-ARNs [113]. A l'inverse, la queue poly-A des mt-ARN serait plutôt un signal de dégradation chez les plantes. Chez la levure, le rôle stabilisateur serait joué par un dodécamère présent à la fin de tous les mt-ARNm [112]. Dans les mitochondries des trypanosomes, la queue poly-A joue des rôles différents: signal de dégradation lorsque le transcrit n'est pas édité mais stabilisatrice et de taille proportionnelle au degré d'édition dès que le transcrit commence à être édité [114].

Certains transcrits mitochondriaux ont une **queue poly-U en 3'**. C'est le cas des ARNs ribosomiques et des ARN guides des kinetoplastides [115]. On retrouve également une queue poly-U en 3' des transcrits du dinoflagellé *Lingulodinium* [116]. Chez celui-ci, la polyuridylation n'est pas un signal de dégradation car elle concerne la majorité des transcrits mais le poly-U n'est pas traduit car il est en aval du codon stop [116]. Le dinoflagellé *Oxhirris marina*, a la particularité de présenter une **queue poly-U en 5'** de ses deux transcrits mitochondriaux [36].

III.L'analyse bio-informatique de données génomiques

L'objectif de cette section n'est pas tant de faire une revue méthodologique exhaustive que d'introduire le vocabulaire, les concepts et les défis liés aux analyses bio-informatiques en génomique que j'ai réalisées durant ma thèse. Je décrirai les méthodes utilisées pour l'identification des sites édités (utilisées dans les analyses des chapitres 2 et 3) et la caractérisation des candidats de la machinerie d'édition et d'épissage en trans (chapitres 4 et 5).

1. Les données de séquençage ADN et ARN

Depuis le tout premier organisme autonome séquencé, la bactérie *Haemophilus influenzae* en 1995 [117], les techniques de séquençage de génomes entiers ont fait des progrès remarquables. Les stratégies de **nouvelle génération** (NGS, "New Generation Sequencing") permettent une parallélisation du séquençage et autorisent ainsi l'application de la stratégie par fragmentation aléatoire à des génomes eucaryotes entiers, à des transcriptomes (Séquençage ARN), et même à des écosystèmes [118], [119].

A l'issue du séquençage on obtient en fonction du protocole utilisé des **lectures simples** contiguës (*single reads*) ou des **lectures par paires** (*paired-end reads* ou *mate-paired reads*) (figure 9). Ces dernières sont obtenues en séquençant les deux extrémités de fragments d'ADN, généralement de plusieurs kilobases. La grande couverture de séquençage et la fragmentation aléatoire permettent en théorie d'obtenir la séquence complète.

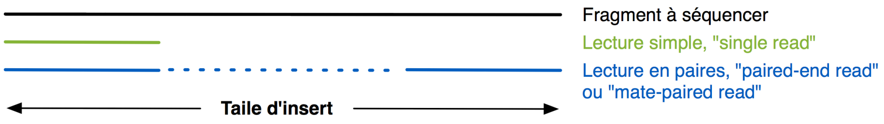

Fragment à séquencer
Lecture simple, "single read"
Lecture en paires, "paired-end read"
ou "mate-paired read"
Taille d'insert

Figure 9: Lectures simples, appariées et taille d'insert

2. Identification des sites mitochondriaux d'édition d'ARN

Classiquement, les sites édités sont identifiés en comparant la séquence d'une région génomique (éventuellement, un gène) à celle de son transcrit (un ARN messager dans le cas d'un gène), puis en identifiant les différences de séquence, parfois nommées *RNA-DNA Differences* (RDD) [120]. D'autres méthodes commencent à émerger qui utilisent uniquement l'information du transcriptome, mais elles nécessitent un critère pour différencier les polymorphismes allèliques des sites d'édition. Par exemple, l'outil GIREMI utilise le déséquilibre de liaison entre les sites édités et les SNP [121]. Comme nous avions également une librairie des séquences génomiques mitochondriales, nous avons préféré l'approche RDD

Par certains aspects, rechercher les RDD est similaire à identifier des polymorphismes allèliques et utilise une approche et des méthodes communes. Dans un premier temps, les lectures ARN sont alignées sur le génome (*RNA-seq mapping*), puis les variants (les mésappariements) sont identifés par appel de variant (*variant calling*), et enfin les variants correspondant à des artéfacts ou à des polymorphismes sur l'ADN sont exclus. Ces étapes seront détaillées dans les sections suivantes.

Alignement sur le génome des lectures ADN et ARN (*RNA-seq mapping*)

L'étape d'alignement est cruciale car elle doit être suffisamment sensible pour ne pas exclure les lectures avec mésappariment (correspondant aux SNPs ou aux sites édités), mais aussi spécifique pour aligner sur une région uniquement les lectures provenant de celle-ci. L'outil Bowtie permet d'aligner les lectures de manière globale ou de manière locale, c'est-à-dire en autorisant que les extrémités ne s'alignent pas [122]. Les outils de la suite TopHat [123] et STAR [124] permettent d'aligner des lectures chevauchant les jonctions d'exons en autorisant des brèches (*gap*). Des outils d'alignement spécifiquement adaptés à la détection d'évènements d'édition appliquent des filtres très stricts sur la qualité de l'alignement afin de réduire les faux positifs éventuels [125].

Appel de variant (*Variant calling*)

Cette étape permet d'identifier les SNPs (variants alléliques sur l'ADN) et les sites potentiels d'édition (variants sur l'ARN). Les outils FreeBayes, [126], GATK [127], et samtools mpileup [128] permettent d'identifier des variations par rapport au génome de référence. Ils diffèrent quant aux lectures utilisées pour trouver les variants (GATK écarte les lectures de mauvaise qualité et dupliquées), quant aux critères distinguant les vrais des faux polymorphismes (FreeBayes et GATK utilisent une stratégie d'apprentissage sur les données tandis que samtools utilise des critères fixés) et quant à la ploïdie (GATK et FreeBayes acceptent des ploïdies supérieures à 2).

Filtrage des faux sites d'édition

La difficulté de détection de vrais sites d'édition tient tant à leurs propriétés intrinsèques qu'à des biais techniques.

Par nature, les sites d'édition d'un ARN peuvent être rares, et donc difficiles à détecter, si le gène dont est issu l'ARN est faiblement exprimé ou si le degré d'édition est faible (c'est-à-dire que seule une partie des ARNs produits sont édités). Ceci contraste avec les variants ADN pour lesquels on s'attend à une couverture uniforme sur tout le génome et dont les proportions dans l'échantillon sont attendues (1/n pour une espèce de ploïdie n). Ces difficultés sont surmontées, en partie, par l'utilisation de méthodes d'appel de variant dont les modèles statistiques acceptent des ploïdies et des fréquences variables. C'est le cas pour GATK [127] et FreeBayes [126], par exemple.

Techniquement, le séquençage haut débit a des caractéristiques qui compliquent encore l'analyse. On exclura de l'analyse (i) les lectures qui s'alignent de façon ambigüe sur plusieurs loci du génome, (ii) les lectures dupliquées provenant d'une amplification artificielle lors de la construction de la librairie, (iii) les lectures pairées s'alignant de manière discordante (dans une mauvaise orientation), (iv) les lectures alignées avec une mauvaise qualité, et (v) les lectures s'alignant en sens inverses pour les librairies brin-spécifique. Une fois l'appel de variant effectué, il faudra également exclure (i) les sites présents dans une région de faible couverture, (ii) les sites présentant un biais de distribution des lectures dans l'une des orientations pour les librairies bi-directionnelles, (iii) les sites avec une mauvaise qualité de séquençage, (iv) les sites dans les régions homopolymériques, (v) les sites proches des jonctions introniques, et (vi) les sites proches des extrémités des lectures, souvent de plus mauvaise qualité. Les cas (i) à (iv) sont généralement pris en charge par les outils d'appel de variants. Certains outils, comme REDitools[128], sont développés spécifiquement pour la détection d'évènements d'édition et prennent en charge l'ensemble de ces critères.

Défis spécifiques à *D. papillatum*.

Les sites d'édition de l'ARN chez *Diplonema* possèdent des caractéristiques spécifiques qui les rendent difficile à identifier. Il est important de noter que la découverte des régions éditées s'est faite fortuitement lors de l'inspection visuelle des alignements ARN sur les cassettes et les transcrits. Pour leur analyse exhaustive, nous avons cependant utilisé une approche plus rigoureuse et reproductible qui se base sur le protocole décrit ci-dessus mais a nécessité des adaptations spécifiques pour *Diplonema*.

L'épissage en trans.

Les logiciels classiques d'alignement de séquence sont parfois capables de détecter des évênements d'épissage en trans qui peuvent survenir naturellement ou sont le signe de remaniements chromosomiques dans les lignées cancéreuses [129]. Cependant, ces évènements sont rares et les intermédiaires de transcription ne possèdent pas de régions flanquantes non codantes comme chez *Diplonema*. Pour s'affranchir du problème de l'épissage en trans, nous avons donc utilisé comme référence les **transcrits mitochondriaux** plutôt que le génome pour aligner les lectures ADN et ARN. Afin de ne pas écarter trop de lectures issues de transcrits intermédiaires (avec des régions flanquantes), et ainsi obtenir une couverture suffisante des jonctions entre modules, les lectures ont été alignées de manière locale, c'est-à-dire en autorisant les mésappariements aux extrémités [130].

L'édition par poly-uridylation

L'édition par poly-uridylation rajoute des uridines qui ne sont pas spécifiés par le génome aux extrémités des modules. Afin de pouvoir détecter les lectures qui possèdent ces uridines supplémentaires et une éventuelle variation du nombre d'uridines ajoutées, il est nécessaire que les transcrits mitochondriaux de référence contiennent les séquences poly-uridylées avec un nombre de U correspondant au nombre moyen observé.

L'édition par substitution groupée

Nous avons découverts que certains sites édités par substitution n'étaient pas éparpillés sur le transcrit, mais regroupés dans des régions intensivement éditées. Ces sites, en générant des mésappariements, vont diminuer le score de qualité de l'alignement. En fonction des paramètres choisis pour l'alignement, le logiciel peut éventuellement écarter les lectures qui contiennent trop de mésappariements. Afin de conserver le plus de transcrits matures, immatures ou intermédiaires d'édition, j'ai utilisé l'astuce suivante: les lectures ARN sont alignées indépendemment sur les deux sequences de reference, les transcrits pre-edites et les transcrits édités. Les deux fichiers d'alignement sont ensuite fusionnés en écartant les doublons éventuels. Cette stratégie nous a permis d'obtenir une bonne couverture des régions éditées et d'identifier les intermédiaires d'édition.

Le haut niveau d'édition.

Contrairement aux niveaux d'édition qui peuvent être faibles chez les animaux, les sites sont intensivement édités chez *Diplonema*, ce qui facilite leur détection. Nous avons donc pu appliquer des filtres de qualité très stringents pour l'appel de variant d'édition. Nous avons utilisé les logiciels FreeBayes et GATK qui présentaient des performances similaires.

3. Recherche des candidats de la machinerie d'édition et de trans-splicing dans le génome nucléaire

Assembler

L'assemblage consiste à reconstruire la séquence des chromosomes ou des transcrits à partir des lectures produites par les séquenceurs. Par chevauchement entre les extrémités des lectures, la séquence est reconstituée et l'on obtient des **contigs**. Avec des lectures en paires (*mate-paired reads* ou *paired-end reads*), l'ordre et la distance approximative des contigs sur le chromosome peuvent être déterminés bien qu'il n'y ait pas de zone de chevauchement. On les joint alors par l'intermédiaire du nucléotide indéfini 'N' pour constituer un *scaffold* .

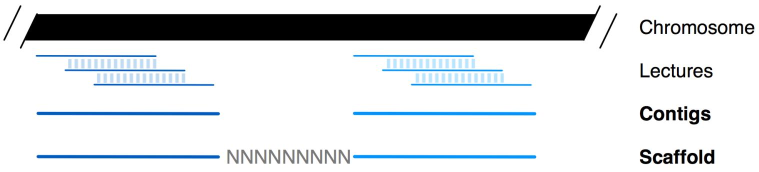


Figure 10. Lectures, contigs et scaffolds

Les défis: erreurs de séquençage, polymorphismes et répétitions

L'assemblage de chaînes de caractère chevauchantes n'est pas un problème insurmontable en soi. Une première difficulté réside dans le taux d'erreur des séquenceurs. La biologie ajoute du piquant au problème avec les deux épineuses caractéristiques que sont les répétitions et le haut degré de polymorphisme de certains génomes.

La difficulté principale des assembleurs concerne les répétitions. Les seules solutions sont l'utilisation de lectures par paires qui vont faire le pont entre les portions uniques, ou le séquençage de plus longues lectures. Cependant, les régions intrinsèquement complexes comme les très longues répétitions, les régions télomèriques et centromériques, ne peuvent être résolues. Ainsi, les génomes publiés sont très fragmentés et doivent être vus comme des "hypothèses de travail" [131], [132].

Quels critères pour déterminer la qualité d'un assemblage ?

Une fois que l'on a obtenu un ou plusieurs assemblages, il n'est pas toujours trivial de déterminer quel est le meilleur, en particulier dans le cas d'organismes non modèle. La mesure la plus utilisée est le **N50**, c'est à dire la taille du contig qui partage en deux les contigs ordonnés par taille. Plus le N50 est grand, plus les grands contigs sont nombreux. Si l'objectif est d'annoter le génome, le N50 devra être au moins de la taille médiane des gènes afin de s'assurer que 50% des gènes sont complètements contenus sur un contig. On cherchera à s'assurer également que l'assemblage obtenu contient bien **l'ensemble du génome**. On peut le vérifier en comparant la taille théorique du génome (ou la taille réelle déterminée expérimentalement) à la longueur totale des contigs ou en quantifiant la quantité de lectures ADN qui peuvent être alignées sur le génome final. Une mesure plus difficile à évaluer est l'**uniformité de la couverture** du génome par les lectures de séquençage. Une brusque chute dans la couverture peut indiquer la présence de lectures chimériques qui réalisent une jonction artificielle entre deux contigs. A l'inverse, une augmentation significative de couverture peut indiquer une compression de répétitions en tandem. Quant aux *scaffolds*, on cherchera à avoir les **trous** les moins nombreux et les plus petits possibles.

Alors que les précédentes mesures sont plutôt quantitatives, le logiciel **CEGMA** (*Core Eukaryotic Genes Mapping Approach*) propose une mesure qualitative de la complétion du génome [133]. Cet outil recherche 248 gènes supposés communs à tous les eucaryotes. Chaque gène est représenté par un modèle de Markov basé sur les orthologues présents chez six espèces modèles. La mesure absolue du nombre de gènes trouvés peut nous donner un indice sur la qualité de l'assemblage. Pour les espèces très divergentes, il est cependant rare de retrouver tous les gènes parce que certains sont soit absents, soit trop dérivés en séquence. Il est plus intéressant d'utiliser cette mesure de manière comparative, entre différents assemblages, car dans ce cas, un nombre plus élevé est forcément le signe d'un meilleur assemblage. Finalement, une autre mesure qualitative est l'analyse du **nombre de transcrits** qui s'alignent sur le génome. Cette mesure nécessite de reconstituer préalablement le transcriptome de l'espèce étudiée *de novo*. On cherchera à minimiser le nombre de transcrits qui ne sont pas alignés et les transcrits morcelés entre plusieurs contigs.

Annoter

L'annotation d'un génome consiste à identifier les éléments génétiques signifiants, principalement les gènes, les promoteurs, les autres séquences régulatrices, et les répétitions. À la suite de cette étape d'annotation syntaxique, l'annotation fonctionnelle permet d’attribuer une fonction moléculaire et/ou un rôle biologique à ces éléments.

Prédicteur de gènes

Les prédicteurs de gène sont des outils permettant de localiser les gènes. Deux grandes familles existent: les **prédicteurs *ab initio*** comme GeneMark [134] qui n'utilisent que les propriétés des séquences (la composition en bases par exemple) et les prédicteurs qui utilisent un **modèle de gène** comme Augustus [135] ou Snap [136]. Le modèle de gène contient les caractéristiques des éléments génétiques comme la taille moyenne des gènes, des exons, des introns, le nombre d'intron par gène, ou les sites d'épissage. Dans le cas d'espèces éloignées des espèces utilisées pour le modèle, il faudra définir des paramètres spécifiques en utilisant un jeu de données d'apprentissage constitué de gènes annotés par des experts.

Indice (*Evidence*)

En plus du modèle de gènes, les outils d'annotation utilisent souvent des indices ou indicateurs expérimentaux permettant d'appuyer la prédiction.

Séquençage ARN (*RNA-seq*)

Un des indicateurs expérimentaux les plus précieux est le séquençage ARN qui permet d’identifier les régions transcrites du génome (voir section III.2). Lorsque celui-ci est connu, les lectures ARN sont **alignées sur le génome** (*mapping*). En l’absence de génome de référence ou lorsque le génome est mal assemblé, l'assemblage *de novo* des lectures ARN, avec l'outil Trinity [137] par exemple, permet de reconstruire les transcrits.

Pour les espèces ayant des *Spliced leaders* (SL, voir partie II.2), la prédiction des transcrits est compliquée par la séquence SL en 5' qui n'est pas présente en amont des gènes sur le génome. Une solution consiste à exciser cette pré-séquence avant l'assemblage des transcrits avec un outil comme Cutadapt [138] normalement utilisé pour supprimer les adapteurs de séquençage. L'alignement des lectures ARN avec SL sur le génome peut être réalisé avec des outils comme STAR autorisant l'alignement local de séquences et la détection d'intron.

Similarités et homologies protéiques inférées

La similarité entre des régions du génome et des gènes connus est utilisée comme proxy pour identifier les gènes homologues. Les similarités sont plutôt recherchées au niveau protéique car la séquence est mieux conservée sur de grandes distances évolutives. La banque SwissProt [139] est généralement utilisée comme cible car elle est non redondante et très bien annotée. Comme elle est biaisée vers les organismes modèles, on enrichit cette banque avec les séquences d'organismes les plus proche phylogénétiquement de l’organisme d'intérêt. Les similarités peuvent être recherchées avec l'outil BLASTX [140] qui traduit le génome dans les six phases de lectures. Après filtrage des correspondance non significatives, les protéines similaires peuvent être ré-alignées sur le génome avec un outil qui détecte plus finement les jonctions exon-intron comme Exonerate [141]. Alternativement, il est possible de rechercher uniquement les homologues des protéines issues de la traduction des transcrits. L'outil Transdecoder (https://transdecoder.github.io/) permet de prédire la protéine la plus probable à partir d'une séquence ARN. Cependant, les gènes non exprimés dans les conditions expérimentales ne peuvent être identifiés.

Domaine protéiques

Les domaines protéiques sont des régions fonctionnellement et/ou structurellement très conservées. Ce sont des indices très précieux pour l'annotation fonctionnelle de gènes dans un contexte d'homologies lointaines car ils sont représentés sous forme d'un modèle de Markov ou d'un motif construit à partir d'alignements multiples de séquences, ce qui augmente la sensibilité de la recherche. L'outil InterProScan [142] permet de rechercher 12 banques de données de motifs, en particulier PFAM [143], ProDom [144], Prosite [145], et PANTHER [146].

ARNs non codants

Les ARNs non codants ne sont généralement pas recherchés par les prédicteurs de gènes mais plutôt par des stratégies dédiées. Les ARNs ribosomiques peuvent être identifiés avec Infernal [147] et les ARN de transfert par tRNAScan-SE [148], tous deux basés sur des modèles de covariance. Les petits ARNs non codants sont difficiles à trouver car leur caractérisation est récente et les outils pour les mettre en évidence sont émergents. De plus, ce sont de courtes séquences dont la structure secondaire est conservée mais pas la séquence primaire.

Annotation syntaxique

Les outils d'**annotation automatique** réalisent la synthèse des gènes trouvés par les prédicteurs et des indices (Séquençage ARN, similarités protéiques, domaines protéiques) à leur disposition pour produire une annotation du génome. Elle contiendra les portions codantes, l'annotation des régions non traduites (UTR), et les variants d'épissage des gènes. Il est conseillé d'utiliser les résultats de plusieurs prédicteurs différents car chacun ayant des sensibilité différentes, l'annotation finale est de meilleure qualité. Un des outils d'annotation automatique fréquemment utilisé est Maker [149]. Idéalement, une étape d'annotation manuelle permettra de corriger et parfaire cette annotation

Annotation fonctionnelle

L'annotation fonctionnelle est l'attribution d'une (ou plusieurs) fonction(s) au gène identifié. En l'absence d'analyse expérimentale, la fonction est attribuée par transfert des annotations des homologues. Cette stratégie est courante, mais hasardeuse: (i) des protéines peuvent être homologues (i.e. partager le même ancêtre) mais ne pas partager la même fonction, (ii) la fonction de la protéine homologue peut être erronée, (iii) les protéines peuvent partager un seul de leurs domaines protéiques. Des outils comme Ahrd (Automated Assignment of Human Readable Descriptions, https://github.com/groupschoof/AHRD) sélectionnent automatiquement parmi un ensemble d'homologues, l'annotation la plus pertinente. Il est inutile par exemple de propager l'annotation "Unknown Protein" si pour l'un des homologues une annotation fonctionnelle existe.

La fonction des protéines peut également être inférée à partir de la présence de domaines protéiques. Ici aussi, la prudence s'impose car certains domaines sont proches et seules des analyses *in silico* poussées (modélisation 3D, analyse phylogénétique) ou expérimentales permettent de les discriminer [150].

Les outils de classification donnent une vue globale du contenu génétique d'un génome. La Gene Ontology (GO) est un vocabulaire contrôlé classant les gènes selon le compartiment cellulaire, la fonction moléculaire et le processus biologique [151]. KEGG (Kyoto Encyclopedia of Genes and Genomes) [152] ou BioCyc classent les gènes selon leur fonction métabolique [153]. Un gène est classé par transfert des classes attribuées aux homologues. Pour un organisme non-modèle et divergent dont il est difficile d'identifier les homologues, la classification est limitée.

IV.*Diplonema papillatum*

"Constructive neutral evolution [...] suggests that neutral evolution may follow a stepwise path to extravagance." - Arlin Stoltzfus [154]

Diplonema papillatum est le diplonémide du groupe des Euglenozoa que j'ai étudié durant ma thèse. Cette section présente une vue d'ensemble des connaissances physiologiques et écologiques sur cette espèce.

1. Les Euglenozoa

L'embranchement des Euglenozoa appartient au super-groupe des **Excavata** [155], [156]. Ce groupe monophylétique comprend le groupe basal des euglenides et les deux groupes frères que sont les kinetoplastides et les diplonémides [157] (figure 11).

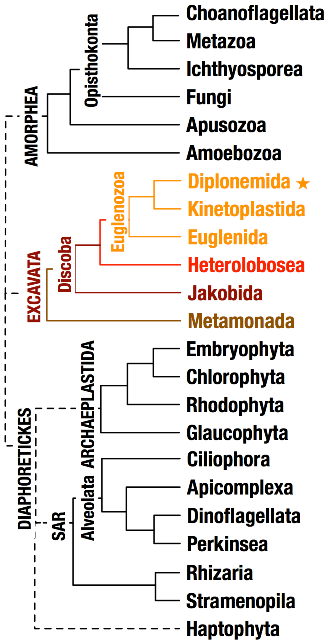


Figure 11. Phylogénie des eucaryotes mettant en évidence les Euglenozoa.

Les Euglenozoa sont des eucaryotes unicellulaires avec un ou deux flagelles en position subapicale à l'entrée du cytostome (bouche) qui se continue par un tube digestif dans lequel transite les aliments. Ils ont le plus souvent une seule mitochondrie qui présente des crêtes mitochondriales discoïdes.

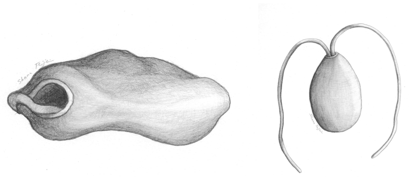


Figure12. Dessin artistique du diplonémide *Rhyncopus euleeides* par Shona Teijero (Université de Montréal).

Parmi les autres éléments distinctifs partagés par les membres du groupe des Euglenozoa, on peut citer l'utilisation d'une base rare dans l'ADN, la base J (beta-D-glucosyl-hydroxymethyluracil) qui serait un signal de terminaison de transcription pour l'ARN polymérase II [158], et l'ajout aux transcrits nucléaires d'une courte séquence épissée en trans, la séquence *Spliced Leader* [159].

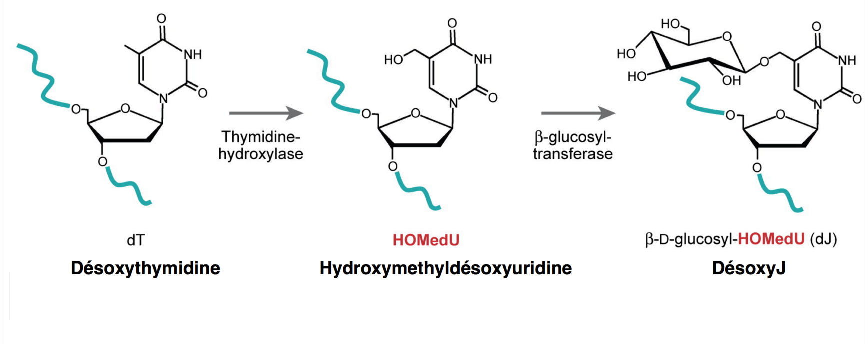


Figure13. Biosynthèse de la base J. La base J est formée par l'hydroxylation puis la glycosylation d'une thymidine [160].

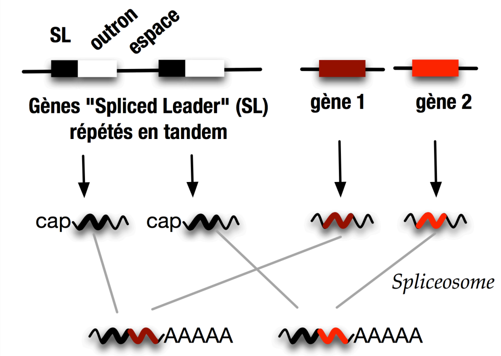


Figure 14. *Spliced leader*. La séquence transcrite à partir des gènes du *spliced-leader* (vague noire) est ajoutee par épissage au bout 5' des pré-mRNA (en rouge).

Les euglénides sont le seul groupe des Euglenozoa avec des membres photosynthétiques. Leur chloroplaste est issu de l'endosymbiose secondaire d'une algue verte (chlorophyte) [161]. Les kinetoplastides comprennent les trypanosomatides, parasites obligatoires et les bodonides dont certains sont des parasites et d'autres vivent librement. Parmi les trypanosomatides, plusieurs sont connus pour causer des maladies humaines transmises par l'intermédiaire d'insectes hématophages dont les principales sont la maladie du sommeil (parasite, *Trypanosoma brucei* ; vecteur mouche, tse tse ; principalement en Afrique), la maladie de Chagas (parasite, *Trypanosoma cruzi* ; vecteur, insecte Triatominae ; en Amérique Centrale et du Sud), et les leishmanioses (parasites, genre *Leishmania* ; vecteur, insecte Phlébotomes ; zones tropicales ou sub-tropicales). Leur pathogénicité ainsi que les caractéristiques exceptionnelles de leur mitochondrie (voir partie I.2 sur la structure du génome et partie II.3 sur l'édition) en font le groupe d'Euglenozoa le plus étudié. Enfin, les diplonémides sont des organismes libres que nous décrirons plus en détail dans la partie suivante.

2. Ecologie des diplonémides

Distribution et diversité

Les diplonémides dont le genre s'appelait autrefois *Isonema* ont été identifiés pour la première fois en 1914 et sont principalement des espèces marines [162], [163]. Jusqu'à récemment, deux genres étaient identifiés : *Rhynchopus* et *Diplonema*.

Une étude de la distribution des diplonemides avec des amorces taxon-spécifiques de l'ARNr 18S (ARN de la petite sous-unité ribosomique cytosolique) a identifié deux autres clades, DSPD-I et DSPD-II (*Deep-sea Pelagic Diplonemids*) contenant exclusivement des espèces non cultivées [164]. De plus l'étude a montré une grande diversité des diplonémides, particulièrement en eau profonde (>500 m) et une nette stratification des phylotypes (>99% d'identité de séquence) dans une colonne d'eau. Les phylotypes sont souvent cosmopolites et se retrouvent dans la plupart des échantillons analysés (figure 15). Un échantillon benthique (du fond des océans) montre un enrichissement significatif en phylotypes de la même clade (groupe monophylétique) que *Diplonema* et *Rhynchopus* ce qui laisse supposer que ceux-ci proviendraient de fonds marins.

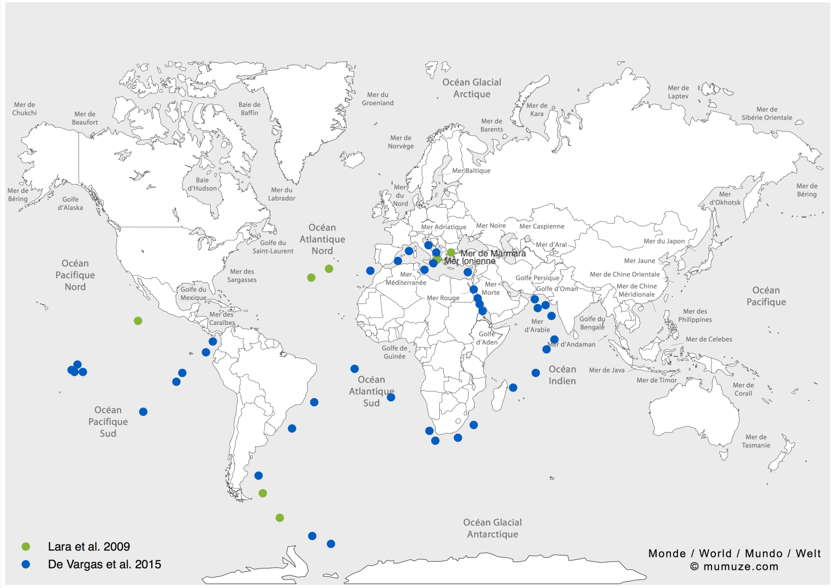


Figure 15. Distribution des diplonemides dans les océans. D'après des échantillons épipélagiques (zone photique < 200 m) et mésopelagiques (zone entre 200 m et 1000 m de profondeur) [165] et des échantillons de surface et de mer profonde (entre 5 m et 3500 m de profondeur) [164].

L'étude récente de la composition eucaryote du plancton par séquençage de la portion variable V9 de l'ARNr 18S en surface et à la profondeur maximale de chlorophylle (*Deep Chlorophyll Maximum*, DCM) a montré également l'ubiquité et la diversité génétique remarquable des diplonémides [165]. Chez les eucaryotes, les diplonémides sont le troisième taxon le plus divers génétiquement et le sixième en terme d'abondance dans les océans (Figure 16).

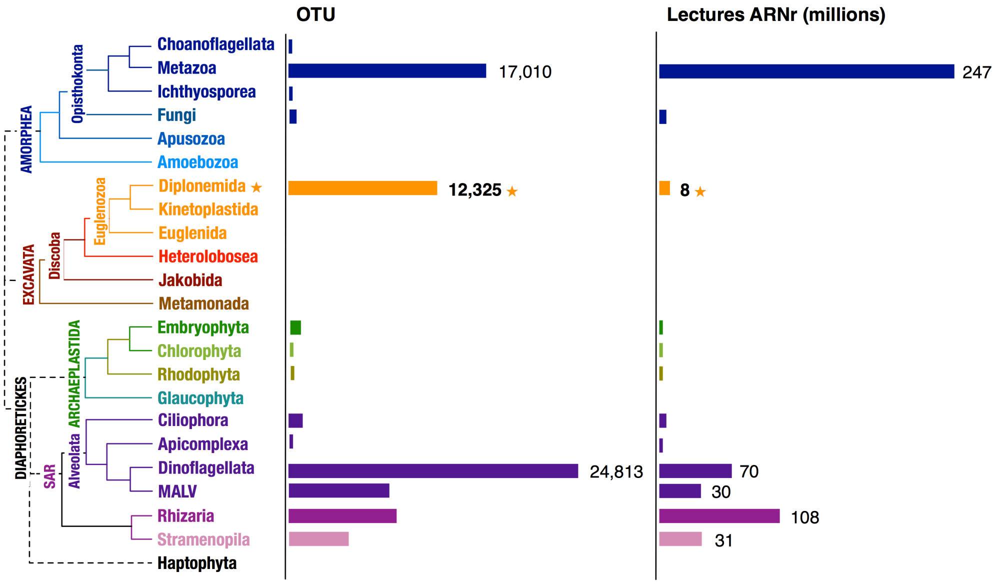


Figure 16. Diversité et abondance des diplonemides. D'après les données de l'expédition Tara [165]. OTU, *Operational taxonomic unit*, unité de base pour l'analyse phylogénétique des écosystèmes. Le nombre d'OTU et de millions de séquences ARN est indiqué pour les groupes plus divers ou plus abondants que les diplonémides.

Mode trophique

Les diplonémides sont des organismes libres, probablement osmo- ou phagotrophes dont le mode exacte de nutrition n'est pas identifié. *Rhyncopus* a été décrit comme parasite de diatomes [166] et de crustacés [167] et *Diplonema* infecterait les palourdes [162] et des plantes aquatiques d'eau douce (Cryptocorynae) [163]. Cependant, les diplonémides pourraient être également des organismes saprophytes.

3. Expression des gènes mitochondriaux

Cette section fait la synthèse des connaissances antérieures au début de ma thèse.

Structure du génome mitochondrial

Les premières informations sur le génome mitochondrial ont été publiées à l'occasion d'une étude phylogénétiques utilisant l'ARN ribosomique 18S et la sous-unité 1 de la Cytochrome Oxidase (*cox1*) pour résoudre la place des diplonemides dans l'embranchement des Euglenozoa [157]. Les auteurs ont alors constaté que le génome mitochondrial était particulièrement riche en GC, ce qui est rare pour un génome mitochondrial, et qu'il était probablement composé de molécules circulaires. Une sonde réalisée à partir de l'ADNc de *cox1* s'hybride avec différents fragments du génome et les auteurs ne parviennent pas à reconstituer le gène et concluent: "*Cloning and sequence analysis of the mitochondrial genomic COl sequence are in progress.*"

En 2005, notre laboratoire résout la structure du génome mitochondrial [43] et en 2007 sa structure génique et certains intermédiaires ARN [44]. Comme les autres Euglenozoa, le génome est composé de plusieurs chromosomes mais il ressemble d'avantage à celui des kinetoplastides car les molécules sont circulaires et monomériques. Cependant, les chromosomes ne sont pas concaténés et entrelacées comme certains kinetoplastes. Il existe deux classes de chromosomes nommés A (~6 kb) et B (~7 kb) composés d'une région commune à A et B, la **région constante partagée**, représentant 35 à 45% du chromoosome, flanquée de part et d'autre d'une **région constante classe-spécifique** droite et gauche (partagée par tous les membres d'une classe), encadrant une **cassette** différente pour tous les chromosomes. Les régions constantes sont riches en répétitions et constituent plus de 95% de la taille des chromosomes. La cassette contient une portion codante appelée **module** encadré par deux régions flanquantes uniques (figure 17). Les gènes mitochondriaux sont fragmentés en modules de 73 pb à 534 pb. Par exemple le gène *cox1* est fragmenté en neuf modules.

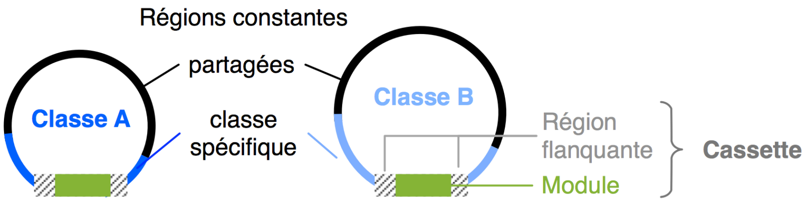


Figure 17. Structure des chromosomes mitochondriaux de *D. papillatum*

Expression des gènes mitochondriaux

Les modules mitochondriaux sont transcrits indépendamment les uns des autres. Le promoteur est probablement dans la région constante partagée car les transcripts primaires sont flanqués de séquences des régions constantes. Les cassettes peuvent être orientées dans les deux sens par rapport à la région constante et l'on détecte des transcrits dans les deux orientations pour un module donné [45]. Chaque transcrit primaire subit une batterie de traitements post-transcriptionnels qui comprend l'excision des extrémités non-codantes, l'épissage en trans des modules et la polyadénylation du dernier module C-terminal. De plus, l'édition par ajout d'uridines est détectée à la jonction de quelques modules. Par exemple, pour le premier gène analysé, *cox1*, six uridines sont ajoutées à la jonction entre les modules 4 et 5. Au total, neuf gènes scindés en 3 à 11 modules ont été identifiés: *atp6, cob, cox1, cox2, cox3, nad4, nad5, nad7, rnl* (voir figure 4 p.10 pour les fonctions des gènes). Le gène *rnl* de la grande sous-unité ribosomique, spécifié par un seul module semblait étrangement petit et il était surprenant de ne pas trouver le gène codant pour la petite sous-unité ribosomique, qui est ubiquitaire dans les mitochondries.

Les machineries réalisant ces processus de maturation post-transcriptionnels sont inconnues. Les séquences flanquant les modules ne contiennent pas les signaux typiques d'épissage que l'on trouve dans les introns connus et caractérisés (*c.f.* partie I.2) [44], [168]

L'investigation de la structure du gène *cox1* chez trois autres diplonémides (*Diplonema ambulator*, *Diplonema* sp. 2 et *Rhynchopus euleeides*) montre que le nombre de modules est systématiquement conservé bien que la séquence varie substantiellement. Etonnamment, le nombre de U ajoutés à la jonction des modules 4 et 5 de *cox1* est strictement conservé [168].

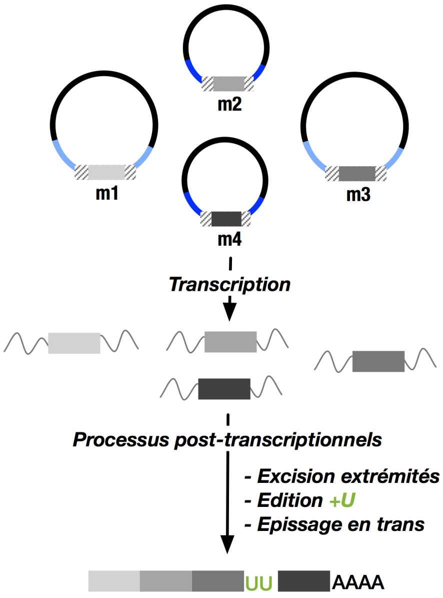


Figure 18. Expression des gènes mitochondriaux (avant 2011).

Chapitre 2: Trans-splicing and RNA editing of LSU rRNA in *Diplonema* mitochondria

Matus Valach, Sandrine Moreira, Georgette N. Kiethega and Gertraud Burger

Référence: Valach, M., S. Moreira, G. N. Kiethega, and G. Burger. 2014. "Trans-Splicing and RNA Editing of LSU rRNA in *Diplonema* Mitochondria.." Nucleic Acids Research 42(4):2660–72.

Contribution des auteurs:

GB a conçu et supervisé l'étude. GNK. a réalisé les RT-PCR. GNK. et MV. ont préparé les échantillons d'ARN pour le séquençage ARN. MV. a isolé les mitochondries et les mitoribosomes, éliminé les ARNr et analysé les structures secondaires. Les analyses préliminaires de Séquençage ARN ont été réalisées par GB. et MV. SM. a réalisé les analyses de données détaillées incluant l'alignement des lectures et les statistiques. Tous les auteurs ont contribué à la version finale du manuscrit.

I. Abstract

Mitochondrial ribosomal RNAs (rRNAs) often display reduced size and deviant secondary structure, and sometimes are fragmented, as are their corresponding genes. Here we report a mitochondrial large subunit rRNA (mt-LSU rRNA) with unprecedented features. In the protist *Diplonema*, the *rnl* gene is split into two pieces (modules 1 and 2, 534- and 352-nt long) that are encoded by distinct mitochondrial chromosomes, yet the rRNA is continuous. To reconstruct the post-transcriptional maturation pathway of this rRNA, we have catalogued transcript intermediates by deep RNA sequencing and RT-PCR. Gene modules are transcribed separately. Subsequently, transcripts are end-processed, the module-1 transcript is polyuridylated and the module-2 transcript is polyadenylated. The two modules are joined via trans-splicing that retains at the junction ~26 uridines, resulting in an extent of insertion RNA editing not observed before in any system. The A-tail of trans-spliced molecules is shorter than that of mono-module 2, and completely absent from mitoribosome-associated mt-LSU rRNA. We also characterize putative antisense transcripts. Antisense-mono-modules corroborate bi-directional transcription of chromosomes. Antisense-mt-LSU rRNA, if functional, has the potential of guiding concomitantly trans-splicing and editing of this rRNA. Together, these findings open a window on the investigation of complex regulatory networks that orchestrate multiple and biochemically diverse post-transcriptional events.

II. Introduction

Mitochondria are semi-autonomous organelles of the eukaryotic cell that contain not only a distinct genome--typically a multicopy, single type of circular-mapping chromosome--but also their own translation machinery. Although protein components of the mitoribosome are partly or completely encoded by the nuclear genome, synthesized in the cytosol and imported into mitochondria, the genes specifying the large subunit (LSU) and small subunit (SSU) ribosomal RNAs always reside on mitochondrial DNA (mtDNA) [169]. Mitochondrial rRNAs (mt-rRNAs) are sometimes fragmented, extreme cases being dinoflagellates and apicomplexans [170]-[172]. In *Plasmodium* the ~20 gene pieces are spread across the genome on both DNA strands, are separately transcribed and then assembled into the ribosome, without covalently joining of the rRNA pieces [170]. Further peculiarities observed in certain mt-rRNAs are homo-nucleotide appendages at their 3' end, e.g. oligo(A) tails in *Plasmodium* [173] and short poly(U) tails in kinetoplastids [115].

Identifying mt-rRNA genes and accurate termini mapping in mitochondrial genome sequences can be challenging, particularly in taxa that are not closely related to model organisms and whose mtDNA has diverged far away from its bacterial ancestor. This applies *in extremis* to the unicellular protozoan (protist) group diplonemids, the sistergroup of kinetoplastids. Mitochondrial genes of *Diplonema papillatum* and its relatives are not only highly divergent but also systematically fragmented in a unique way. Genes consist of up to 11 pieces (modules) that are ~80–530-nt-long, and each is encoded on a distinct circular chromosome of 6 kb (class A) or 7 kb (class B). Modules are transcribed separately and subsequently joined into continuous RNAs. With each chromosome containing only 1–6% coding sequence, the estimated genome size of *Diplonema* mtDNA is unusually large [~600 kb; [45]].

In contrast to the eccentric genome structure, the gene complement of *Diplonema* mtDNA is rather conventional. Mitochondrial genes encode components of the respiratory chain, oxidative phosphorylation and mitoribosome, notably NADH dehydrogenase subunits 1, 4, 5, 7 and 8; apocytochrome b, cytochrome oxidase subunits 1–3, ATP synthase subunit 6 and LSU rRNA. The gene for mitochondrial SSU rRNA has not yet been identified [174]. For *rnl* (encoding LSU rRNA), we only found a 352-nt long 30-terminal portion that is otherwise well conserved. Incidentally, this RNA piece is the most highly expressed transcript in poly(A) libraries. However, the complete sequence and overall organization of *rnl* has remained unrecognized for many years, partly due to technical challenges in culturing sufficient cell material and isolating mitochondria from *Diplonema*, but also, as we know now, because of the intricate structure and biosynthesis of mt-LSU rRNA. We succeeded to resolve the puzzle by high-throughput RNA sequencing (RNA-Seq) and show here that maturation of *Diplonema* mt-LSU rRNA proceeds by multiple steps including extensive RNA editing. We also identify antisense RNA molecules that have the potential for guiding both trans-splicing and RNA editing of mt-LSU rRNA, but their function has yet to be demonstrated.

III.Materials and methods

1. Sequences deposited in public-domain databases

We have deposited in GenBank the genomic sequence of *rnl*-module 1 plus adjacent chromosome regions (accession no. KF633465) and the cDNA sequences of cytosolic 5.8S, 18S and 28S rRNA of *D. papillatum* (accession nos. KF633466-KF633468). The sequence of *rnl*-module 2 was deposited previously under the accession number JQ302963. A partial sequence of *D. papillatum* cytosolic 18S rRNA had been deposited before by others (GenBank accession no. AF119811).

2. Strain, culture and extraction of mtRNA

D. papillatum (ATCC 50162) was obtained from the American Type Culture Collection. The organism was cultivated axenically at 16–20ºC in artificial seawater enriched with 1% fetal horse serum (Wisent) and 0.1% bacto tryptone. For extended large-scale cultivations, chloramphenicol (40 mg/L) was added to prevent bacterial contamination. To isolate mitochondria, cells were collected by centrifugation at 3000g for 10min, washed once with ice-cold ST buffer [0.65M sorbitol, 20mM Tris (pH 7.5), 5mM EDTA] and disrupted by nitrogen decompression at 600psi (Parr Instrument Company) in the same buffer. Mitochondrial RNA and DNA were extracted from an organelle-enriched fraction isolated by differential and sucrose gradient centrifugation essentially as devised earlier [175]. More specifically, intact cells and nuclei were removed by centrifugation at 3000g. The mitochondria-enriched fraction was obtained after centrifugation at 30 000g (20 min) followed by two consecutive separations on a discontinuous sucrose gradient [15, 25, 35, 45 and 60% sucrose supplemented with 20mM Tris (pH 7.5) and 5 mM EDTA] at 130 000g (1 h). Mitochondria accumulated at the interface between the sucrose layers of 35 and 45% (and/or 25 and 35%). Mitoribosomes were enriched via separating a cell lysate by two consecutive kinetic centrifugations, the first on a step gradient (10–35% glycerol, in steps of 5%) at 250 000xg for 2 h and the second on a continuous gradient (10–40% glycerol) at 250 000g for 4 h. Fractions enriched in mt-LSU rRNA (as determined by agarose gel electrophoresis) were pooled. RNA was extracted by a home-made Trizol substitute [175]. Residual DNA was removed from RNA preparations by either RNeasy (Qiagen) column purification or digestion with RNase-free DNase I (Fermentas), or TURBO DNase (Invitrogen) followed by phenol-chloroform extraction. Poly(A) RNA was enriched by a passage through oligo(dT)-cellulose (Amersham), after denaturation of the aqueous solution at 72ºC for 2min and subsequent chilling on ice.

3. Northern hybridization

DNase-treated RNA was separated electrophoretically in a MOPS/formaldehyde denaturing gel (1.2% agarose, 3% formaldehyde), side by side with the Riboruler High and

Low Range RNA ladders (0.2 – 6.0kb and 0.1 – 1.0kb, Fermentas). As a size marker for smaller molecules, we used single-stranded DNA, which was obtained from denatured RT-PCR products of 130–440-nt-long *rnl* segments. This marker was visualized by hybridization to a radioactively labeled oligonucleotide (see later in text). Primers used for RT-PCR (and product sizes) are dp210+dp211 (130 nt), dp72+dp211 (240 nt), dp168+dp169 (355 nt), dp210+dp208 (440 nt) and dp72+dp208 (560nt). As size markers and positive controls for mono-modules, we used RNAs synthesized by in vitro transcription of PCR products amplified with primer pairs dp230+dp216 (module 1) and dp232+dp168 (module 2). Oligonucleotides used as primers and hybridization probes are listed in Supplementary Table S1. The electrophoretically separated nucleic acids were blotted on a nylon membrane (Zeta-Probe, BioRad) and fixed by baking the membrane at 80°C for 60 min. As hybridization probes, we used oligo-deoxynucleotides radio-labeled by T4 polynucleotide kinase in the presence of [γ-32P]ATP. For the detection of antisense transcripts, we used an oligoribonucleotide probe that was *in vitro* transcribed from PCR amplicons that in turn were produced with primer pairs dp225+dp210 (antisense targeting) and dp226+dp211 (sense-targeting control); for each primer pairs, one contained the T7 promoter in addition to gene-specific sequence. *In vitro* transcription with T7 RNA polymerase [New England BioLabs (NEB)] was performed in the presence of [α-32P]UTP, for internal labeling. Membranes were hybridized overnight at 55°C in either 5x saline sodium citrate (SSC) supplemented with 5x Denhardt's solution (0.1% polyvinylpyrrolidone, 0.1% BSA, 0.1% Ficoll 400) and 0.5% sodium dodecyl sulfate (SDS) when oligonucleotide probes were used or the ULTRAhyb buffer (Ambion) when RNA probes were used. Subsequently, membranes were washed twice at 50°C in 2x SSC plus 0.1% SDS (oligonucleotide probes), or twice at 68°C in 0.1x SSC plus 0.1% SDS (RNA probes) and visualized using a phosphor-imaging screen scanned by a Personal Molecular Imager (Bio-Rad). Quantitative measurements of relative band intensities were conducted with the Image Lab 4.1 software (Bio-Rad).

4. CircRT-PCR and RT-PCR

DNase-treated RNA was incubated with tobacco acid phosphatase (TAP; Epicenter) and T4 polynucleotide kinase (PNK; NEB). For circRT-PCR experiment, we used an unmodified kinase that possesses 3'-phosphatase activity. RNA was diluted to 20ng/mL and circularized using T4 RNA ligase (Roche). 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, but for certain RT-PCR experiments, 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 [176]. Primer sequences are given in the Supplementary Table S1. For all RT-PCR experiments, a negative control was performed where no template RNA was added.

5. Cloning and sequencing of amplicons

Amplicon termini were rendered blunt with T7 DNA polymerase and the Klenow fragment of DNA polymerase I (NEB), agarose gel-purified, phosphorylated with T4 PNK (NEB) and ligated into the vector pBFL6cat, which is an in-house constructed, small pBlueScript derivative. Libraries of cDNA were cloned into pDNR- LIB (Clonetech). After transformation into Escherichia coli DH5a, plasmid DNA was extracted using the Qiagen 96-well mini-prep kit. Sequencing reactions were performed with the BigDye Terminator version 3.1 Cycle Sequencing Kit from Applied Biosystems and sequenced on an ABI 370 Analyzer.

6. High-throughput RNA sequencing

Total RNA and mitochondrial RNA-enriched samples from *D. papillatum* were depleted of cytosolic 5, 5.8, 18 and 28S rRNA using a series of 50 end biotinylated oligonucleotides (IDT) complementary to these rRNAs. For oligonucleotide design, we used the 5S rRNA sequence published earlier by others (GenBank accession no. AY007785) and the 5.8, 18 and 28S rRNA sequences reported here. The amount of the overabundant mt-LSU rRNA in mitochondrial RNA preparations was reduced by an oligonucleotide (dp72-5biosg) targeting the *rnl* module 2 (for oligonucleotides, see Supplementary Table S1). After annealing, oligonucleotide:rRNA hybrids were removed by streptavidin-coated magnetic beads (MyOne C1 and/or M- 270 Dynabeads; Invitrogen). The library PA was made from cytosolic rRNA-depleted total RNA enriched for poly(A) RNA (see earlier in text), and the libraries F1 and F2 from mitochondrial RNA, following the supplier-recommended protocol devised for strand-specific RNA-Seq libraries and using the ScriptSeqTM RNA-Seq Library Preparation Kit (Epicentre). The difference between the F1 and F2 libraries is that for F2 the RNA fragmentation step was omitted to minimize further fragmentation of short RNA molecules. The F1, F2 and PA libraries were constructed and paired-end-sequenced (2 x 101 nt; Illumina HiSeq 2000) at the commercial technology platform Macrogen (Korea). According to the service provider, spurious antisense reads are below 2% and typically at 1% with the methodology used. For the GG library, we used RNA extracted from a subcellular fraction enriched in mitoribosomes. The library was constructed using the TruSeq Stranded Total RNA Sample Prep kit (Illumina) following the suppliers instructions and paired-end sequenced (2 x 250 nt; Illumina MiSeq) at the Genome Quebec Innovation Center in Montreal.

7. RNA-Seq data analysis

From the libraries F1, F2 and PA, we obtained between 50 and 70 Mio raw fastq reads of 101-nt length, and from the library GG ~3 Mio raw reads of 250-nt length (Supplementary Table S2). Reads corresponding to cytosolic rRNAs were filtered out using Geneious 5.6 (Biomatters, New Zealand) leaving 40% (F1), 33% (F2), 95% (PA) and 15% (GG) reads. Adapters were removed from the 5' and 3' termini of reads with cutadapt version 1.2.1 (http://journal.embnet.org/index.php/embnetjournal/ article/view/200). As parameters, we used a subsequence of 12nt at the 3' end or 5' end of the 5' and 3' adapters, respectively, to allow for partial adapter sequence in the reads. The error rate was set to 0.1. Cutadapt was also used for quality clipping with a quality threshold of 20. Reads <20nt were discarded. Statistics for the cleaning steps of reads are compiled in Supplementary Table S2. The data set used for further analysis was built from paired reads; reads that lost their mate during filtering were discarded using an in-house script. As a reference on which to map the read pairs to, we constructed a set of theoretically possible reference transcript sequences, including the expected intermediary molecules from RNA processing, trans-splicing and RNA editing. Paired reads were mapped onto each of these reference transcripts using bowtie2 (http://bowtie-bio.sourceforge.net/index. shtml). Bowtie was executed independently on each reference transcript and for each sense (forward and reverse) using the corresponding –norc/–nofw option. Read pairs where only one mate maps to the reference transcript or which are discordant (i.e. not mapping to the same strand or where the forward mate maps downstream of the reverse mate) were discarded from the alignment. Finally, using in- house scripts, pairs were removed that do not overlap with any of the reference transcripts, have a mapping quality <30, or a number of deletions ≥3.

From libraries F1 and F2, we removed read pairs representing insert sizes ≥165 nt that originate from spurious dp72-amplification products primed by residual, contaminating dp72, an oligonucleotide that was used during sample preparation for the removal of cytosolic rRNA. Output files in sam format were subsequently transformed into '.bam' files with SAMtools version 1.4 (http://samtools.sourceforge.net/). Alignments were visualized with tablet version 1.13.05.17 available at URL http://bioinf.scri.ac.uk/tablet/ [177]. The statistics for the length distribution of the poly(A) tail and the poly(U) tract were calculated using an in-house script, which filters fastq files or the 'sam' alignment file, respectively, for reads that overlap the upstream and/or downstream modules by a minimum number of nucleotides (typically 10–12 nt) and which contain a minimum number of homopolymeric nucleotides (typically 4 nt). The exact parameters are given in the figure legends.

8. RNA secondary structure modeling

We searched for conserved primary sequence and secondary structure motifs of mitochondrial LSU rRNAs by using the phylogeny-based consensus model available at the Comparative RNA Web (http://www.rna.ccbb.utexas. edu) [178]. Thermodynamic folding was predicted by RNAfold 2.0 [179]. Identified conserved motifs served as anchors for manual folding of the entire sequence to fit the model. Conventional nomenclature for sequential numbering of secondary structure elements has been used [e.g. [180]]. The secondary structure was drawn with XRNA 1.1.12 (http://rna.ucsc.edu/rnacenter/xrna/xrna. html) and finalized using CorelDRAW X4.

IV.Results

1. Identification of mt-LSU rRNA and its gene in *Diplonema*

The 352-nt-long 3'-terminal portion of mt-LSU rRNA from *Diplonema* was early on recognized as a top candidate for an unidentified rRNA, due to its extremely high abundance (representing 1% of all ESTs) in cDNA libraries constructed from total poly(A) RNA [[45]; GenBank record JQ302963]. This RNA species carries an A tail of >25 nt and as we show here, is a precursor transcript of mt-LSU rRNA (see section later in text). For identification of mt-LSU rRNA from *Diplonema*, neither BLAST nor Rfam searches, nor comparison with mitochondrial rRNA sequences from other taxa was successful. Counterparts from euglenozoan species (i.e. the euglenid *Euglena gracilis* and kinetoplastids) not only are as highly divergent as mt-LSU from *Diplonema* but also display an extremely dissimilar nucleotide composition (15–20% G+C-content in kinetoplastids and *Euglena* versus ~50% in *Diplonema*).

Mature *Diplonema* mt-LSU rRNA was first detected by northern hybridization, using an oligonucleotide as a probe that is specific for the 3'-terminal *rnl* portion. In total RNA, this probe lights up a major band of ~0.9 kb, together with a weaker band at 0.4 kb (Figure 1A, right panel). The same band pattern is seen when using the entire 3'-terminal piece as a probe (Supplementary Figure S1). The 0.9-kb band is most likely the mature mt-LSU rRNA, whereas the smaller one, present in >20-times lower steady-state concentration, corresponds to the polyadenylated 3'-terminal portion. A size of 0.9kb may appear small for mt-LSU rRNA, but the kinetoplastid counterpart is not much longer [1.1 kb; GenBank acc. no. TRBKPGEN; [181]]. In poly(A) RNA, the RNA species of 0.4kb is highly enriched, whereas that of 0.9kb is nearly undetected [Figure 1A, lane 'poly(A)'; Supplementary Figure S1], which is in accordance with evidence from cDNA sequencing. Apparently, mature mt-LSU rRNA has a shorter A tail than the 352-nt RNA species, so that only a small fraction of is pulled down during the poly(A) enrichment procedure.

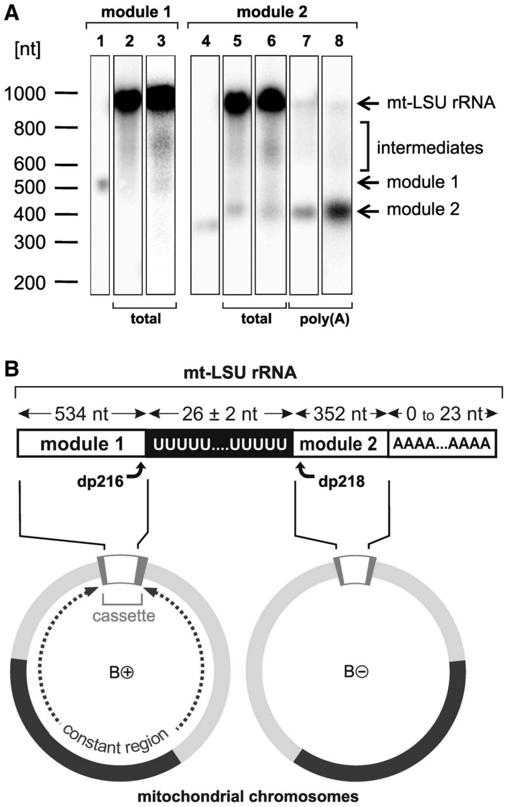


Figure 1. Mitochondrial LSU rRNA of *Diplonema*. (A) Northern blot hybridization. Lane 1, *in vitro* transcription product of *rnl* module 1 (540 nt); lane 4, *in vitro* transcription product of *rnl* module 2 (359 nt; synthetic RNAs are 6 and 7 nt longer than the corresponding modules); lanes 2, 3, 5 and 6, total RNA (~5 μg); lanes 7 and 8, poly(A) RNA (~0.5 μg) extracted from whole cells. RNA in lanes 2 and 5 is from one preparation; that in lanes 3 and 6 is from an independent preparation. Blotted RNA was probed with radioactively labeled oligonucleotides dp216 (lanes 1–3) and dp218 (lanes 4–8) that target module 1 and module 2 of *rnl*, respectively. Bands represent the mature mt-LSU rRNA (~900 nt), mono-module 1 transcripts (~550 nt; the weak band in lane 3 is clearly visible on the original image), mono-module 2 transcripts (~450 nt) and presumptive end-processing intermediates of single-module transcripts. The size markers are indicated on the left. The signal ratio of mt-LSU rRNA versus mono-module 1 transcripts varies noticeably from one preparation to another; it is 100:1 in lane 2 and 60:1 in lane 3. The signal ratio of mt-LSU rRNA versus mono-module 2 transcripts (lanes 5 and 6; total RNA) is ~20:1. This ratio is ~1:5 to ~1:17 in poly(A)-enriched RNA (lanes 7 and 8), a variation depending on the particular oligo(dT) pull-down experiment. Notably, the steady-state of mono-module 1 transcript is lower than that of mono-module 2. The same is seen in RNA-Seq experiments (see Figure 4). (B) Upper part, schematic sequence of mtLSU rRNA. The U-tract between modules 1 and 2 (black box) is not encoded by mtDNA, but added post-transcriptionally. Regions with which northern hybridization probes dp216 and dp218 anneal are indicated. Lower part, coding regions of mt-LSU rRNA on mitochondrial chromosomes. Modules 1 and 2 are contained in cassettes of B-class chromosomes, but oriented in opposite direction relative to the chromosome's constant region [indicated as B(+) and B(–), see text]. Non-coding regions within the cassettes ('unique flanking regions') are shown in dark gray. The constant region of chromosomes (light gray) is ~95% identical across all B-class chromosomes [45]. The black part of the constant region is also present in A-class chromosomes ('shared constant region').

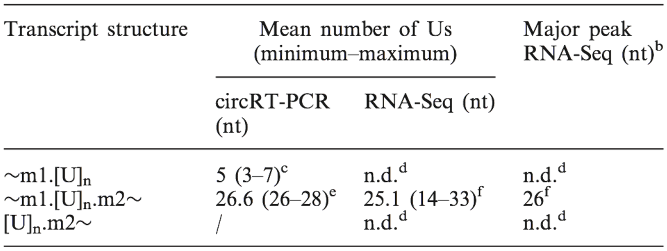
The 5'-terminal region of mt-LSU rRNA was identified by RT-PCR applied to circularized RNA (circRT-PCR) using a pair of 'divergent' primers annealing with the molecule's 3' end region (see 'Methods'). Subsequent cloning and sequencing revealed a 534-nt-long stretch upstream of the 3' end portion of *rnl*. As only two such clones were obtained (in multiple experiments), we confirmed their authenticity by northern hybridization. An oligonucleotide specific to the presumed 5'-terminal portion lights up the 0.9-kb product and in addition a faint 0.5-kb band that corresponds to the 5'-terminal portion alone (Figure 1A, left panel). RNA-Seq data (later in text) provided the ultimate confirmation for the 534-nt-long sequence being the 5' moiety of mt-LSU rRNA in *Diplonema*.

The most remarkable sequence feature of *Diplonema* mt-LSU rRNA is a run of ~26 uridines (Us) immediately upstream of its 3' moiety (Figure 1B, upper part; Table 1 and Supplementary Figure S2). This homopolymer tract was confirmed independently by RT-PCR using a primer pair that anneals upstream and downstream of this tract (Supplementary Table

S5 and Supplementary Figure S3). The observed U-tract length varies by about ±3, which is apparently due to experimental rather than biological variation (Supplementary Table S6); errors probably occur during PCR amplification or the sequencing reaction itself, as commercial RT-enzymes have high synthesis fidelity. We posit that this long U-tract is the reason why RT-PCR-based experiments yielded extremely low numbers of reads. This bias is observed also in RNA-Seq (see later in text).

The gene specifying *Diplonema* mt-LSU rRNA was pinpointed by mapping the rRNA sequence on the available mtDNA sequence, revealing two previously unannotated coding regions embedded in cassettes of separate B-class chromosomes (for a definition of 'cassette', see legend of Figure 1B). These coding regions are referred to as *rnl* modules 1 and 2 (Figure 1B, lower panel). With 534bp length, *rnl* module 1 is the longest among all known gene modules in *Diplonema* mtDNA, whereas *rnl* module 2 (352 bp) is of average size. Gene module 2 of *rnl* lacks a 3' terminal A-homopolymer stretch, which is obviously added by post-transcriptional polyadenylation. Also absent from both the module 1 and module 2-coding regions is a terminal T tract, otherwise present in the center of the mt-LSU rRNA cDNA sequence. The sequence of gene module 1 ends precisely upstream, whereas that of gene module 2 starts exactly downstream of the U tract in mt-LSU rRNA. Therefore, these non-encoded nucleotides must be added post-transcriptionally, resulting in U-insertion RNA editing. This is by far the longest stretch of non-encoded Us seen in *Diplonema* mitochondria and also the largest number of nucleotides added at a single editing site ever observed.

Table 1. Non-encoded U-tract length of mt-LSU rRNA and its precursor transcripts[a]



Transcript structure	Mean number of Us (minimum–maximum)		Major peak RNA-Seq (nt)[b]
	circRT-PCR (nt)	RNA-Seq (nt)	
~m1.$[U]_n$	5 (3–7)[c]	n.d.[d]	n.d.[d]
~m1.$[U]_n$.m2~	26.6 (26–28)[e]	25.1 (14–33)[f]	26[f]
$[U]_n$.m2~	/	n.d.[d]	n.d.[d]

[a] m1, m2, rnl modules 1 and 2; [U]n, uridine-homopolymer of length n; m1.[U]n, module 1 with 3'-terminal U tract; m1.[U]n.m2, LSU-rRNA; [U]n.m2, module 2 with 5'-terminal U tract; and ~, exact module terminus not determined. n.d., not identified; /, not observed.

[b] Peak positions of tract length distribution is taken from Supplementary Figure S2. Libraries F2, PA and GG display similar U-tract length as F1 shown here.

[c] Four clones (dp11056, dp11060, dp11084, dp11088).

[d] This type of transcript could not be identified unambiguously.

[e] Seven clones (dp9540, dp10594, dp11008, dp11009, dp11012, dp11017, dp11064).
[f] Not-quality clipped individual reads from the F1 library.

2. 2D structure modeling of mt-LSU rRNA

The secondary (2°) structure of the 3' moiety from *Diplonema* mt-LSU rRNA was modeled based on comparison with the mitochondrial consensus structure--the homologs from kinetoplastids and *E. gracilis* are too divergent for a meaningful comparison of covariant residues (Figure 2A). Only domains IV to VI [as defined for *E. coli* (Figure 2B)] are conventional, albeit reduced. Domain V encompasses the peptidyl-transferase center (PTC) and is the most conserved region of LSU rRNAs. As in many other reduced mt-LSU rRNAs, the *Diplonema* molecule lacks the helices H76-H79 that in *E. coli* bind the ribosomal protein L1 and H83-H86 that associate with 5S rRNA. Domain VI lacks major parts, and the sequence that connects H73 and H95 in most other mt-LSU rRNAs [178], [182] is unusually short. Just a few of the universally conserved sequence motifs are readily recognizable in the *Diplonema* molecule, namely, those corresponding to the basis of helix H90 and its single-stranded junctions to H89 and H93, as well as the terminal loops of helices H80, H92 and H95 (the latter is also known as the a-sarcin/ricin loop). Nonetheless, domain V of *Diplonema* mt-LSU rRNA resembles bacterial 23S rRNA somewhat more closely than that of kinetoplastids, the latter lacking for example H97 [183], [184].

Domain IV is most likely constituted by the 3' third of the module 1 sequence. We recognize the conserved helices H69 and H71 with their surrounding single-stranded regions that are involved in the majority of inter-subunit contacts with ribosomal SSU and functionally important interactions with ribosome-bound tRNAs [185]. Two other consensus helices of this domain lack a substantial peripheral portion in *Diplonema* as well as in kinetoplastids and several other taxa. The structure model places the poly(U) tract at the 3' end of domain IV. Two 4-nt-long purine stretches upstream in module 1 might base-pair with poly(U) to form a helix analogous to H61. However, this region could also remain single-stranded as in the 2° model of kinetoplastid and nematode mt-LSU rRNAs [182], [186].

Although we were able to reconstruct a reasonable 2° structure model of the 3' half of mt-LSU rRNA from *Diplonema*, folding the 5' half of this molecule (domains I-III) is challenging due to several reasons (but see Supplementary Figure S7). First, this part of the molecule is in general moderately conserved. In addition, comparative modeling was not feasible due to low sequence similarity between *Diplonema* mt-LSU rRNA and homologs with available 2° models. Finally, modeling based on thermodynamic folding leads to an excessive number of alternatives because the G+C rich (51%) sequence allows profuse base-pairing possibilities. As to length and structure, the 5' half of mt-LSU rRNA from *Diplonema* is more reduced and shorter than that from kinetoplastids, yet comparably deviant as that from certain animals as detailed in the Discussion section.

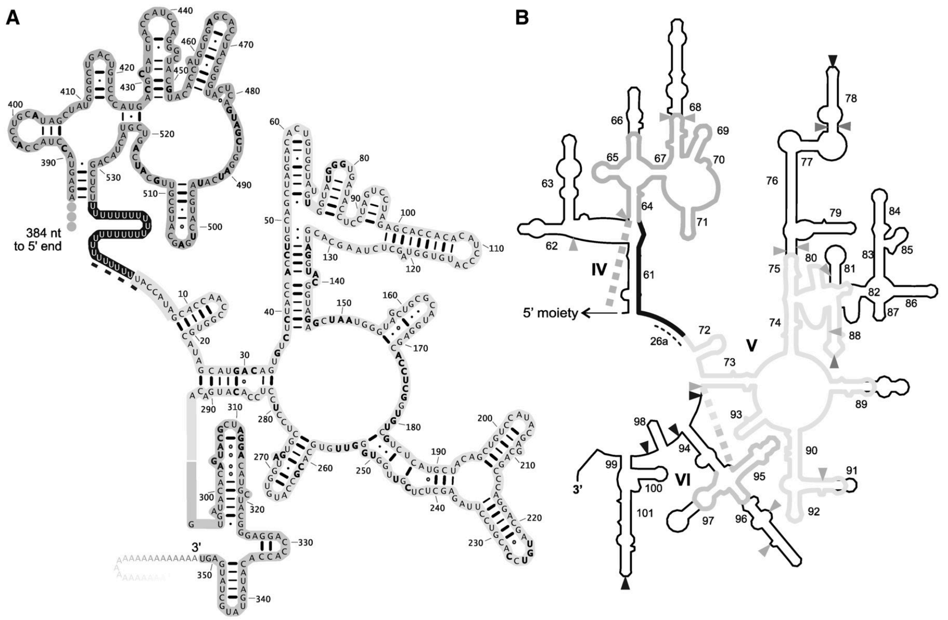


Figure 2. Putative secondary structure of the mt-LSU rRNA (3′ moiety) from *Diplonema*. (A) The structure was modeled according to the mitochondrial reference sequence and structure (http://www.rna.icmb.utexas.edu). Residues identical to the universal consensus sequence [178], [182] are shown in bold. Domain IV is composed of the 3′ portion of module 1 (dark gray shading) and the post-transcriptionally added U-tract (black shading). Domains V and VI are encoded by module 2 (light gray shading). The thin dashed line marks helix 26a (see 'Discussion'). Base pairing is indicated as thin lines, thick lines, dots and open circles corresponding to A:U, G:C, G:U and other base pairs, respectively. Residues are numbered according to nucleotide positions in *rnl* modules 1 (upstream of U-tract) and 2 (downstream of U-tract). The nucleotide pair U305:A314 in the module 2 corresponds to a conserved trans Watson-Crick/Hoogsteen pair in the *E. coli* structure. (B) The 2° structure of the 3′ moiety from *Diplonema* mt-LSU rRNA mapped onto the structure from *E. coli* LSU rRNA. Helices are numbered according to [180]. H95, α-sarcin/ricin loop. Thick gray and black lines indicate the structure elements present in the *Diplonema* model [same shading as in (A)]. Triangles indicate breakpoints in the 3′ half of fragmented LSU rRNAs from apicomplexan [170], [187], [188] and dinoflagellate [171], [172] mitochondria (light gray triangles), several green algal mitochondria (gray triangles; [57], [189]-[191], and the kinetoplastid [192] and euglenid [193] cytosol (black triangles). It is noteworthy that among all known cases of discontinuous

domain-IV LSU rRNA (apicomplexans and dinoflagellates), none is split in the 3′ half of H61.

3. Deep sequencing and RT-PCR analysis of the rnl transcript population

To capture *rnl* transcripts of *Diplonema* in a comprehensive way, we performed massively parallel sequencing (RNA-Seq) of three RNA samples (F1, F2, PA). Samples F1 and F2 were extracted from a subcellular fraction enriched in mitochondria; sample PA was enriched for poly(A) RNA. The applied RNA-Seq approach involved paired-end library construction by RNA fragmentation for F1 and PA (but not F2), random hexamer priming and strand-specific sequencing. Average fragment (insert) length is 300 nt, read length is 101nt and read depth is ~60 Mio reads per sample. Primer and quality trimming resulted in ~100 Mio paired reads of ≥20-nt length for all three libraries together (60%). Of these, 1.066 Mio paired reads (1%) contain rnl sequences. A fourth small library (GG) was constructed with an RNA sample that was extracted from a mitoribosome-enriched subcellular fraction of *Diplonema*. Reads of this library were used to characterize the mitoribosome-associated LSU rRNA. Information on RNA-Seq data are compiled in Supplementary Tables S2 and S3.

First, we mapped read pairs to the sequence of mt-LSU rRNA. Read coverage of the mitochondrial libraries is depicted in Supplementary Figure S4A. Detailed inspection of coverage showed that only 14 (quality-clipped) reads span completely the internal U-tract and include ≥10nt of both adjacent modules, although ~150000 reads map to the module-1/module-2 junction region; the majority of U-tract containing reads maps to either the3'end of module 1 or the 5' end of module2 (Supplementary Table S4). This bias is due to low sequence quality in homopolymer tracts. More than 99.9% reads have quality values <20 from the 13th U-tract position on, so that all sequence beyond this position is removed by quality clipping during the read preprocessing step (Supplementary Figure S5). Therefore, we used the inferred 'inserts' (i.e. the interval inferred from paired-end reads) instead of reads for mapping onto mt-LSU rRNA (Figure 3; for logarithmic scale, see Supplementary Figure S4A) and most of the other analyses described later in text.

For targeted detection of long transcripts and accurate mapping of their termini, we conducted in addition RT- PCR using specific primers that anneal within module 1 or module 2 of *rnl*. In experiments with circularized RNA, primers point in divergent direction, otherwise they are oriented in convergent fashion.

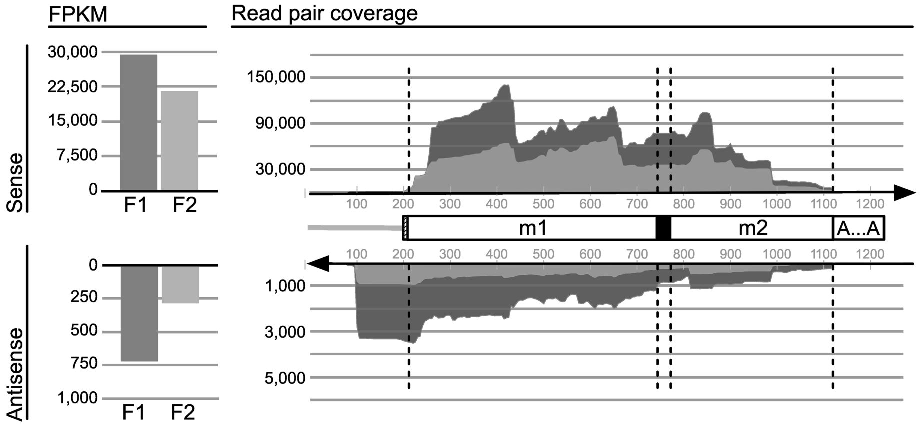


Figure 3. Coverage of *Diplonema* mt-LSU rRNA by RNA-Seq data. Mapping of inferred inserts from two mitochondrial libraries, F1 (dark gray) and F2 (light gray). Vertical scales, counts of inserts. Cartoon in the center, schematic representation of the virtual reference transcript to which inserts were mapped. Unfilled boxes labeled m1 and m2, *rnl* modules 1 and 2, respectively. Black box, poly(U) of ~26 length added by RNA editing; dashed box upstream module 1, unique flanking region; gray line, transcribed constant region of B-class chromosomes (see Figure 1B). 'A...A', A-tail. It should be noted that inserts (and reads) cannot be mapped unambiguously beyond ~80 nt upstream and downstream of modules because these regions are nearly identical in sequence with those from other modules residing on B-class chromosomes. Stacked-area chart on the right side, coverage by sense (upper area) and antisense (lower area) inserts, respectively. The bar charts to the left represent the total number of reads covering the corresponding area in the stacked-area chart. The scales for sense and antisense transcripts differ by a factor of 3'. Sharp drop-off in antisense read coverage ~100 nt upstream of *rnl* module 1 (a zone corresponding to the constant region of B-class chromosomes) reflecting a discrete 3' end of antisense RNAs. Uneven read coverage along the sequence is probably due to sequence bias.

4. Maturation intermediates of rnl transcripts

To characterize intermediates of mt-LSU rRNA, we mapped RNA-Seq inserts from the mitochondrial libraries F1 and F2 to three virtual reference transcript sequences, which represent the primary transcript of each individual module and a trans-spliced, edited and polyadenylated transcript. LSU rRNA precursors were also characterized by RT-PCR and circRT-PCR experiments.

End-processing intermediates are of two types, transcripts including an *rnl* module plus either both adjacent non-coding regions or a single adjacent region retained on either end (Figure 4). Fully processed module transcripts are seen as well (Table 2). Notably, not only fully processed modules but also end-processing intermediates engage in trans-splicing. For example, we detected a transcript with joined modules 1 and 2, whose 30 end still has non-coding sequence attached. Mapping of RNA-Seq data to unprocessed reference sequences is shown in Supplementary Figures S4B and C.

RNA editing almost certainly takes place before trans- splicing, because neither RNA-Seq nor RT-PCR detected reads where the 30 end of *rnl* module 1 is immediately upstream-adjacent to the 5' end of module 2. Uridine residues are most likely added 30 to module 1 and not 5' to module 2 according to circRT-PCR experiments (Table 1). For *Diplonema* cox1, U-appendage editing of the module upstream of the editing site has been validated more rigorously. Only after 3' dephosphorylation of RNAs did we observe upstream modules with Us appended at the 3' end, but under no condition was the downstream module found with Us attached to its 5' end [174].

RNA editing intermediates of *rnl* that have excess or deficit Us cannot be determined reliably, because sequences containing homopolymers are of low quality, especially those where the U-tract is at the 5' end of the read (Supplementary Figure S5). At present, the two following editing scenarios remain indistinguishable: (i) uncontrolled addition of numerous Us and subsequent precise trimming as is the case for U-insertion editing of trypanosome mitochondria [194] and (ii) controlled addition of the exact number of nucleotides.

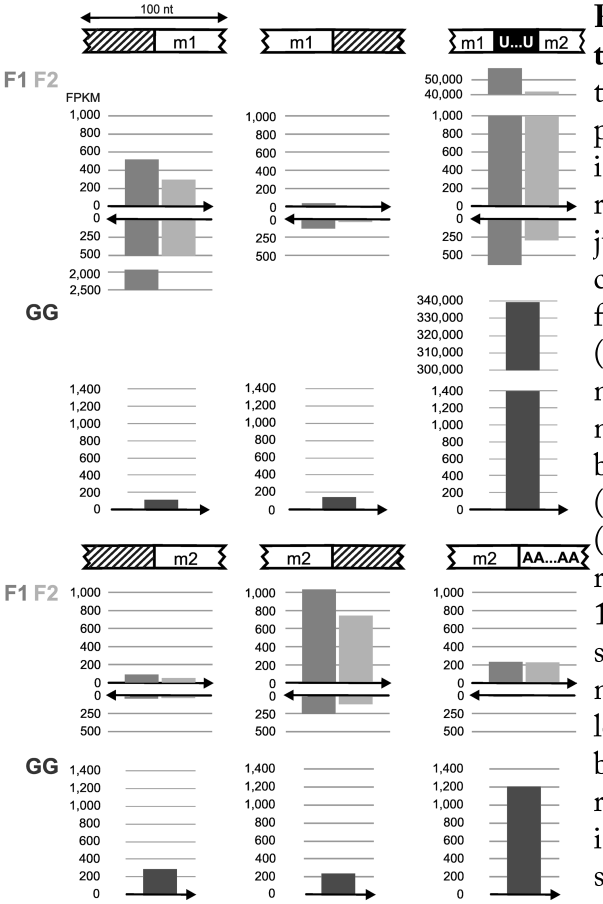
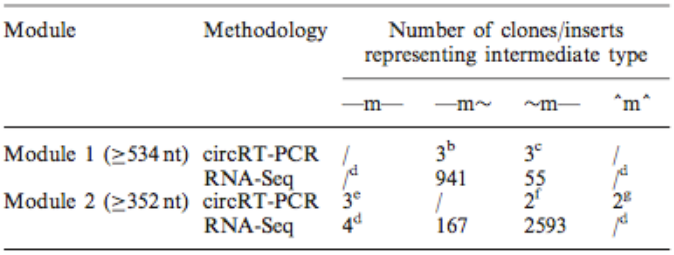


Figure 4. Maturation intermediates of *rnl* transcripts. Cartoons depict schematically the regions where maturation processes take place. White, hatched and black boxes indicate modules and the A tail, non-coding regions and the U-tract at the module junction, respectively. Bar charts beneath cartoons show the number of paired reads from the mitochondrial libraries F1 (medium gray) and F2 (light gray), and the mitoribosome library GG (dark gray) that map to the designated regions. The arrow below the bars specifies reads in sense (pointing to the right) and antisense (pointing to the left) direction. Counted reads suffice the following criteria: within a 100-nt-long region around the maturation site, reads (forward or reverse read of mapped read pairs) are required to cover at least 55 nt of this window, i.e. overlap boundaries (between modules and other regions) by at least 5nt. The proportion of immature *rnl* transcripts in the library GG serves as a measure for mitoribosome enrichment.

Table 2. End processing intermediates of rnl modules transcripts[a]



Module	Methodology	Number of clones/inserts representing intermediate type			
		—m—	—m~	~m—	ˆmˆ
Module 1 (≥534 nt)	circRT-PCR	/	3[b]	3[c]	/
	RNA-Seq	/[d]	941	55	/[d]
Module 2 (≥352 nt)	circRT-PCR	3[e]	/	2[f]	2[g]
	RNA-Seq	4[d]	167	2593	/[d]

[a] Number of observed clones in RT-PCR experiments or inserts in RNA-Seq libraries F1 and F2 (latter data taken from Figure 4). Symbols and abbreviations used: --, non-coding adjacent region; m, rnl-module 1 or 2; ^m^, module end-processed at both termini; ~m, m~, nature of module's 5' end or 3' end, respectively, is unknown (may be unprocessed or processed); /, not observed.

[b] Three clones (dp11008, dp11034, dp11059); length of non-coding regions is 324, 20 and 69 nt, respectively.

[c] Three clones (dp9411, dp9613, dp11051l); length of non-coding regions is 163, 22 and 3nt.

[d] Low probability of observation, because the libraries have an insert size average of 300nt.

[e] Three clones (dp9408, dp10574b, dp10586).

[f] Two clones (dp9411, dp9613).
[g] Two clones (dp10439rb, dp10526a).

The A-tail of module 2-containing *rnl* transcripts displays substantial differences in length (Table 3). Mono-module 2 is polyadenylated by addition of up to 90 As, whereas trans-spliced transcripts have predominantly ~20-nt-long A-tails, and mt-LSU rRNA incorporated in the mitoribosome has virtually no A-tail. These differences are seen consistently in all three experimental approaches used in this study. In northern hybridization, we observe different signal ratios of mono-module 2 versus mature rRNA. The ratio in total RNA is ~1:20, but nearly inverse in the poly(A) RNA-enriched fraction (see Figure 1 and Supplementary Figure S1). In circRT-PCR experiments, A-tails of rnl mono-module 2 transcripts are up to ~50-nt-long, whereas those of the trans-spliced transcript are not longer than 26nt. Finally, A-tail size distributions in RNA-Seq data from total-cell poly(A) RNA exhibit a broad crest up to 80 nt, those from total mitochondrial RNA peak at ~20nt and the ones from mitoribosomal RNA have a dominant maximum at 0 nt (Table 3 and Supplementary Figure S6). The possible biological significance of this variation in A-tail length will be examined in the 'Discussion' section.

5. Antisense RNA covering module junction and editing site of mt-LSU rRNA

We posited earlier that trans-splicing and RNA editing of the mitochondrial protein-coding gene *cox1* in *Diplonema* might be instructed by antisense RNAs. Preliminary evidence for antisense transcripts of a protein-coding gene came from targeted RT-PCR experiments [174]. However, the yield of products was low and the informative sequence obtained (after subtraction of primer sequences) was only a few nucleotides long. Here we re-examine whether guiding antisense RNAs exist in *Diplonema* mitochondria by focusing on one of the most highly expressed mitochondrial genes, *rnl*, and by exploiting strand-specific RNA-Seq data.

Strikingly, putative antisense transcripts of mt-LSU RNA are detected at ~2.5%, which is significantly above background (see 'Materials and Methods' section and Figure 3A, lower panel). The existence of such transcripts is also seen in RT-PCR experiments (Supplementary Figure S3 and Supplementary Table S5). Remarkably, antisense read coverage drops off sharply

~100nt upstream of module 1, a zone corresponding to the constant region of B-class chromosomes (Figure 3A). This drop-off reflects a discrete 3' end of *rnl* antisense RNAs. The same phenomenon is seen in read mapping of antisense transcripts from cox1-module 1 (not shown), which is likewise a first module encoded on a B-class (+) chromosome (see Figure 1B). Whether the *rnl*-antisense 3' terminus is generated by transcription termination or processing remains to be investigated.

In contrast to their 3' end, the 5' terminus of *rnl* antisense RNAs appears variable in the read coverage profile. We attempted to determine the length of these transcripts by northern experiments using either single-stranded oligo-deoxynucleotides or in vitro transcribed RNAs as a probe, but the signals were extremely weak (not shown). Antisense transcripts might be a heterodisperse assemblage of different length that do not form a homogenous band in gel electrophoresis; an already weak signal spread out instead of concentrated in a band would be difficult to detect by northern hybridization. Neither could we find the potential gene encoding the anti-mt-LSU RNA in the available ~250kb mtDNA nor in the currently draft assembly of nuclear DNA. It is possible that the gene was not found because it is encoded in a yet unsequenced genomic region, or alternatively, because there is no such gene as elaborated in the 'Discussion'.

Putative antisense transcripts of unprocessed modules are also seen in RNA-Seq data. These RNAs apparently originate from bi-directional transcription of *rnl*-module- carrying chromosomes (Supplementary Figure S4B and C). Transcription in *Diplonema* mitochondria starts in the shared, constant region of chromosomes located opposite to modules [174]. As modules are oriented in either sense relative to the shared region (as for example *rnl* modules 1 and 2 Figure 1B), the promoter(s) must be able to drive transcription of both strands.

V.Discussion

1. Regulation of rnl gene expression in *Diplonema* mitochondria

Based on the observed types of *rnl*-transcript intermediates, two diametrically opposite maturation pathways of mt-LSU rRNA can be postulated. One interpretation of the results is that polyadenylation is a dead-end reaction, tagging molecules that failed to be trans-spliced or incorporated into the ribosome (Figure 5A). However, this view does not explain why only module 2 but not module 1 is polyadenylated. The other hypothesis, which we favour, considers that polyadenylation is crucial for mt-LSU maturation. We posit that module 2 is first polyadenylated and then deadenylated in two subsequent steps, with the particular A-tail length being the check-points for trans-splicing of modules 1 and 2, and then for assembly of the trans-splicing product into the ribosome (Figure 5B). This view would explain the difference in predominant A-tail length of mt-LSU rRNA from total mitochondrial RNA extractions (~20 nt) versus mitoribosome-extracted RNA (~0nt) as follows. The former RNA preparation may contain mainly rRNA that is not incorporated into the ribosome. Still, we cannot fully exclude technical variation because different protocols were used for constructing the two libraries.

The various biochemical reactions involved in the expression of *Diplonema* mt-LSU rRNA, module-end processing, adenylation, uridylation, trans-splicing and potentially A-tail trimming of the molecule's 3' end, must be catalyzed by an assortment of activities (ribonuclease, polymerase and ligase), as well as trans-factors that guide trans-splicing and editing. Traditionally, multi-step biochemical pathways are pictured as a cascade of catalytic steps, where the product of a given reaction is the sub- strate for the subsequent step. However, in *Diplonema* mitochondria, most transcript maturation-steps proceed independently from one another in the sense that the reaction at one extremity of the transcript is not influenced by the nature of the other extremity. This excludes a strictly linear, assembly line-like maturation pathway in this system. Parallelization, thought to accelerate this multi-step process, might be achieved by a molecular machine that combines all activities in one ('processo-edito-spliceosome'), i.e. one that would properly position guiding factors relative to its catalytic domains and allow that the two extremities of a given transcript are sculpted in an independent fashion and in no particular order. The only steps where the nature of the 'other' end seems to matter in mt-LSU rRNA maturation of *Diplonema* is polyadenylation or deadenylation, which, according to the two above pathway hypotheses, appear to be the 'rubbish' or 'quality' stamps of molecules.

In contrast to the here proposed integrated multi- functional complex in *Diplonema* mitochondria, the current view of kinetoplastids mitochondrial (m)RNA maturation postulates the sequential action of two major complexes, each having dedicated functions. The RNA editing core complex conducts cleavage of pre-mRNA at the editing site, removal or addition of Us and resealing of the transcript, whereas the mitochondrial RNA-binding complex 1 recruits guide RNAs and interfaces with gRNA processing and mRNA tailing [reviewed in [195]].

2. Antisense transcripts

We detected two types of antisense RNAs, anti-*rnl*-mono-modules and anti-mt-LSU rRNA transcripts. Anti-mono-module transcripts most likely arise by bidirectional transcription of chromosomes, as the promoter(s) in the shared region must accommodate modules encoded on the plus and the minus strand [see Figure 1B and [174]]. The observed higher steady-state concentration of the *rnl* sense transcript could be achieved by either an elevated transcription rate in sense direction or faster degradation of antisense transcripts. Either scenario calls for controlled strand-dependent transcript regulation, whose nature is yet to be unraveled.

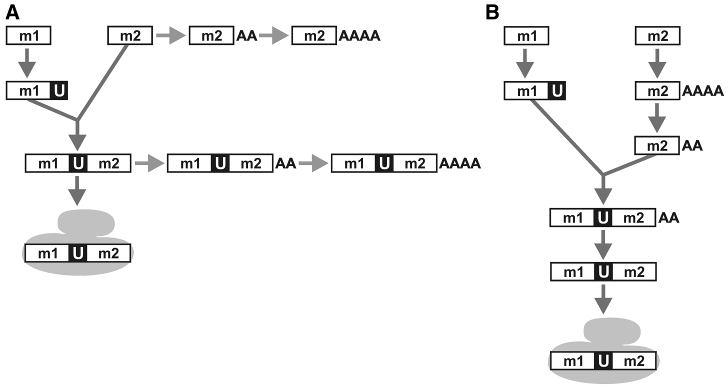


Figure 5. Maturation process of mt-LSU rRNA in *Diplonema* mitochondria. For clarity, the cartoon disregards end-processing of module 1 and module 2 precursor transcripts. m1, m2, rnl module 1 and 2, respectively. U, post-transcriptionally-added U tract. AA, AAAA, poly(A) tails of ~20nt or ~40–90nt length, respectively. The gray-filled shape symbolizes the mitoribosome. (A) Hypothetical pathway where polyadenylated rnl- transcripts represent dead ends instead of maturation intermediates. (B) Alternative pathway (preferred hypothesis) where the polyadenylation status plays a key role in mt-LSU rRNA maturation: a poly(A)tail length of ~20 As signals a check point for trans-splicing, and absence of an A-tail from the trans-spliced product is a requirement for incorporation of the transcript into the mitoribosome.

The origin of anti-mt-LSU rRNA is less obvious, as a corresponding gene has not been detected. Either the gene is encoded in yet unsequenced portions of the mitochondrial or nuclear genomes or alternatively, no such gene exists in *Diplonema*. The antisense RNA might be transcribed from mature mt-LSU rRNA and inherited epigenetically from generation to generation. Antisense transcription templated by mt-LSU rRNA would require an RNA-dependent RNA polymerase (RdRp). As this activity has broad taxonomic distribution [196]-[200], the *Diplonema* nuclear genome might well encode a mitochondrion-targeted enzyme. Epigenetic inheritance of RNAs has precedents as well, for example, in ciliates [201] and *C. elegans* [202], where RNAs transmitted to daughter cells are involved in genome rearrangement and antiviral response, respectively.

3. *Diplonema* mt-LSU rRNA is extraordinarily short and derived

With only ~910 nt, mt-LSU rRNA of *Diplonema* is among the smallest known, but still longer than that of certain nematodes, bryophytes and rotifers [529–729 nt; (38–41)]. It is the module-1 portion (534nt) that is substantially shorter in *Diplonema* (and even more in the aforementioned animals) compared with counterparts from other euglenozoans and heteroloboseans [~730 nt in kinetoplastids (e.g. GenBank accession no. NC_000894), >800 nt in *Euglena* [40], and 1485 nt in *Naegleria* (GenBank accession no. AF288092)].

As stated in the 'Results' section, folding the *Diplonema rnl* sequence into the consensus 5'-half 2° structure of mt-LSU rRNA is challenging. The problems include low conservation, absence of comparative data from close relatives and the possibility to build numerous alternative structures with this G+C-rich sequence, making selection of the single most likely model difficult. For illustration, one of the multiple equally probable structure models is shown in Supplementary Figure S7. With the availability of mt-LSU rRNA sequences from other diplonemids, it should become feasible to model confidently this portion of the molecule. Finally, it is conceivable albeit not likely, that a separate 50 mt-LSU rRNA piece exists. Whereas the mitoribosome-enriched fraction analysed here contains a highly abundant 350-nt molecule (not shown), this RNA species lacks 2° structure motifs typical for mt-LSU rRNA, but instead displays remote similarity to phylogenetically conserved mt-SSU rRNA signatures [helices h18, h44 and h45; numbering as in [185]]. Whether this molecule represents indeed mt-SSU RNA is currently uncertain, because its 5' tier consists virtually exclusively of Gs and Ts impeding meaningful secondary structure modeling, and its length is much shorter than ever reported for this rRNA. These issues could be re-examined rigorously once a protocol is available for isolating pure mitoribosomes from *Diplonema* and by sequencing a mitoribosomal library prepared specifically for small RNAs.

The 3' half is the most conserved portion of all mt-LSU rRNAs. The corresponding 2° structure of *Diplonema* mt-LSU rRNA was modeled based on comparison with the mitochondrial consensus structure--the homologs from kinetoplastids and *E. gracilis* are too divergent for a meaningful comparison of covariant residues. Overall, the fold of domains V and IV is less deviant in *Diplonema* than in kinetoplastids, where the PTC-abutting helices H89 and H91 are considerably truncated. The absent masses of these two helices appear to be the cause of the positionally shifted a-sarcin/ricin loop (H95) toward the PTC [186], seen in the cryo-electron microscopy map of the mitoribosome from *Leishmania tarentolae*. We posit that the extremely short single-stranded segment between helices H73 and H95 in *Diplonema* mt-LSU rRNA induces an even more pronounced overall shift of H95 together with H89 and H91 and stronger domain V/IV compaction in the mitoribosome.

4. Role of extensive U-'insertion' editing in mt-LSU rRNA from *Diplonema*

To our knowledge, LSU rRNA of *Diplonema* mitochondria is the only example of a massively edited rRNA and represents the most extensive editing ever observed at a single site. Other cases of rRNA editing include sparse nucleotide insertion or substitutions that mostly restore secondary structure elements and conserved sequence motifs [203], [204]. In kinetoplastids, mt-rRNAs are virtually never edited. Eukaryotic cytosolic rRNAs are chemically modified [guided by small nucleolar RNAs; ref. [205]], but *sensu stricto* RNA editing has not been described for these molecules.

The region occupied by the U-tract in our model of *Diplonema* mt-LSU rRNA corresponds to the 3' half of H61 in the *E. coli* structure, a helix that plays an important role in the ribosome. The part of this helix abutting H64 ensures correct positioning of the SSU/ LSU-connecting domain IV [180], [182], [185], whereas the part adjacent to H72 is deeply embedded in the ribosome (as are H72 andH73).

According to a most recent 2° structure model of LSU rRNA [206], the segment corresponding to the six 3' terminal nucleotides in the U-tract together with the three first nucleotides in module 2 constitute the 3' half of the newly proposed helix H26a. The corresponding 5' half of this helix is a stretch traditionally modeled as single-strand connecting H26 and H47 in domain II. Helix 26a is thought to be a pivotal structural element of the proposed core domain 0, to which the traditional domains I-VI would be rooted. With the U-tract not only substituting the 3' half of H61 but also being part of H26a, RNA editing of *Diplonema* mt-LSU rRNA would be function-critical.

Accession numbers

GenBank KF633465, KF633466, KF633467, KF633468.

Acknowledgements

The authors acknowledge M. Aoulad-Aissa for excellent technical assistance and thank M.W. Gray (Dalhousie University, Halifax, Canada) for help in modeling the rRNA secondary structure and B.F. Lang (Université de Montreal) for reading the article. G.B. designed and supervised the study. G.N.K. conducted RT-PCR. Both G.N.K. and M.V prepared RNA samples for RNA-Seq. Isolation of mitochondria and mitoribosomes, secondary structure analyses and cytosolic rRNA depletion was performed by M.V. Preliminary RNA-Seq data analysis was conducted by G.B. and M.V. S.M. performed the detailed data analyses including read mapping and statistics. All authors contributed to the final manuscript version.

Funding

Canadian Institute for Health Research [CIHR, grant MOP-79309; to G.B.]; Ph. D. scholarship from the Programme Canadien de Bourses de la Francophonie (PCBF; to G.N.K.). Funding for open access charge: Canadian Institute for Health Research.

Conflict of interest statement.

None declared.

VI.Supplementary data

Table S1. Oligonucleotides used in this study

Oligonucleotide	Sequence (5' to 3')	Targeted region
RT-PCR primers and hybridization probes		
dp71	CTCCCGTAGCATGTCCTAGCG	*rnl* module 2
dp72	CCATTAGCTCTACCGTACCTA	*rnl* module 2
dp168	TTTTGCTCATAGCATACTATGTGG	*rnl* module 2
dp169	TACCATAGCGCCAACCGGTG	*rnl* module 2
dp179	CCACCACATAGTATGCTATGAGC	*rnl* module 2
dp180	CACCGGTTGGCGCTATGGTA	*rnl* module 2
dp208	AGCGTACTCACATGGGTAGCT	*rnl* module 1
dp209	GTAGGTGCTCCACATGGTGTAC	*rnl* module 1
dp210	AGACACTGTCATGCTATGCACC	*rnl* module 2 (junction)
dp211	CTACGGGTACTCAGTAGCTGGA	*rnl* module 1 (junction)
dp213	CATGCGTTGCATCAGTCGTACTAC	*rnl* module 1
dp214	CACCTGTACCTACAGCATCCTGTGTAGTGCA	*rnl* module 1
dp215	CATGCGTTGCATCAGTCGTACTACAGCTC	*rnl* module 1
dp216	GAGCTGTAGTACGACTGATGCAACGCATG	*rnl* module 1
dp218	TGGTAGAGACACTGTCATGCTATGCACCG	*rnl* module 2
dp219	CAGGATGCTGTAGGTACAGGTG	*rnl* module 1
dp220	ACAGTGATACACAGTACGCTAGGA	*rnl* module 2
dp225	TAATACGACTCACTATAGGGCTGACTGTCCCA	*rnl* module 1 (internal) with T7 promoter sequence (RNA probe)
dp226	TAATACGACTCACTATAGGGATGTGTGGTGCTCTA	*rnl* module 2 (internal) with T7 promoter sequence (RNA probe)
dp230	TAATACGACTCACTATAGGGATGTACTGTAGAGGATG	*rnl* module 1 (5′ end) with T7 promoter sequence for a full-length transcript
dp232	TAATACGACTCACTATAGGGTACCATAGCACCAACC	*rnl* module 2 (5′ end) with T7 promoter sequence for a full-length transcript
Biotinylated oligonucleotides used for rRNA depletion		
dp72-5biosg	CCATTAGCTCTACCGTACCTA	mt-rRNA 3' moiety
dp183-5biosg	GATACAACACCTGGGGTTCC	cytosolic 5S rRNA
dp185-5biosg	CGTTCGACGACCTGATGAGT	cytosolic 5.8 rRNA
dp187-5biosg	CCCAAACTGAGCGACTCGT	cytosolic LSU rRNA
dp188-5biosg	ATTGGCACATCTCCCTTTCA	cytosolic LSU rRNA
dp189-5biosg	TCGGCAGGTGAGTTGTTACA	cytosolic LSU rRNA
dp190-5biosg	CCACAAGACACCCTACTACACG	cytosolic LSU rRNA
dp191-5biosg	CCATTCATGCGCGTCTTTAA	cytosolic LSU rRNA
dp192-5biosg	AGACCAAAGGATCGTTAGGC	cytosolic LSU rRNA

dp193-5biosg	CCATTCGGACTCGTATAGACC	cytosolic LSU rRNA
dp195-5biosg	GCTATTGGGCAATTTGCGTAC	cytosolic SSU rRNA
dp196-5biosg	TCGTTTTGATTATTCCTTCTCG	cytosolic SSU rRNA
dp197-5biosg	GACAAATCACTCCACCAACCA	cytosolic SSU rRNA
dp198-5biosg	CACCTACAGCAACCTTGTTACG	cytosolic SSU rRNA

Table S2. Pre-processing of RNA-Seq reads[a]

Libraries	Number of reads			
	Raw	cytosolic rRNAs removed	Adapter trimmed	Quality clipped
F1	71,647,714	28,527,038	28,503,370	28,112,156
F2	53,564,184	18,152,466	18,147,895	17,791,943
PA	61,236,282	57,913,591	57,906,024	57,530,626
Total (F1, F2, PA)	186,448,180	104,593,095	104,557,289	**103,434,725**
GG	6,079,760	2,447,047	6,078,383	6,064,962

[a] Minimum length of retained reads is 20 nt. Libraries F1 and F2 were prepared from mitochondrial RNA, PA from poly(A) enriched total cellular RNA, and GG from a mitoribosome-enriched subcellular fraction. See Methods section of main text for the procedures used. Order of processing steps for libraries F1, F2, PA is (1.) removal of cytosolic rRNAs, (2.) quality clipping and (3.) adapter trimming. For library GG, the order was (1.) adapter clipping, (2.) quality clipping, and (3.) removal of cytosolic rRNAs.

Table S3. Mapping of RNA-Seq reads to mt-LSU rRNA[a]

Library	Number of total reads	Number of paired reads (pairs)[b]	Number of reads (pairs) mapped to m1.Us.m2[b]	FPKM for m1.Us.m2	Number of reads (pairs) mapped to anti-m1.Us.m2[b]	FPKM for anti-m1.Us.m2
F1	28,112,156	26,660,076 (13,330,038)	714,682 (357,341)	29393.9	17,344 (8,672)	713.3
F2	17,791,943	16,993,786 (8,496,893)	332,936 (166,468)	21482.1	4,492 (2,246)	289.8
PA	57,530,626	28,535,314 (14,267,657)	14,444 (7,222)	555.0	346 (173)	13.3
Total (F1, F2, PA)	103,434,725	72,189,176	1,062,062 (531,031)	NA	22,182 (11,091)	NA
GG	2,247,047	2,421.982 (1,210,991)	1,489,136 (744,568)	678,064.0	1,924 (963)	880.1

[a] Numbers are the counts of reads that mapped (using Bowtie) to the reference transcript sequence indicated as m1.Us.m2, which corresponds to the mature mt-LSU rRNA sequence. Anti-m1.Us.m2 is the reverse-complementary sequence. FPKM, Fragment Per Kilobase of transcript per Million of mapped reads. N.a., not applicable.

[b] Concordant mapping of mate-1 and mate-2 reads using a mapping quality (MAPQ; SAMtools) >30 and allowing <3 single-nucleotide deletions per read. For details, see Methods of main text.

Table S4. Mapping of RNA-Seq reads to the module-1/module-2 junction of mt-LSU rRNA[a]

Reference transcript	Library	Number of **reads** including poly(U)			Number of **pairs** with reads including poly(U)	
		Total	Terminal	Internal	Mate1 or mate2	Mate1 and mate2

			poly(U)	poly(U)	(cases 1-5)	(case 6)
Sense	F1	714,682	108,737	13	71,826	36,911
	F2	332,936	35,419	1	27,917	7,502
	PA	14,444	5,461	0	3,630	1,831
	Total	**1,062,062**	**149,617**	**14**	**103,373**	**46,244**
Antisense	F1	17,344	1,190	0	870	320
	F2	4,492	272	1	223	49
	PA	346	10	0	8	2
	Total	**22,182**	**1,472**	**1**	**1,101**	**371**

```
                        744           769
                          |             |
        __________________|             |_______________________
        ________m1________|UUUUU...UUUUU|____________m2_________

Case 1       --->......<--|--           |                          Case 1
Case 2                ----|--->.........|...<------                Case 2
Case 3                    |          ---|---->.........<------     Case 3
Case 4                  --|-------------|---->.........<------     Case 4
Case 5       --->......<--|-------------|--                        Case 5
Case 6                ----|---------><--|---                       Case 6
```

[a]Reads were mapped to the sense and antisense reference sequence including *rnl* module 1, the poly(U) tract and module 2 as depicted at the bottom.

Table S5. RT-PCR across the U-tract of mt-LSU rRNA and its antisense transcripts[a]

Targeted orientation	Primer combination		Amplicon size (expected size)	Confirmed by sequencing	U-tract length[b]
	RT	PCR			
Sense	dp72	dp72 + dp211	~250 (246) nt	+	24-30 nt
	dp210	dp210 + dp208	~450 (446) nt	+	24-29 nt
	dp72	dp72 + dp208	~580 (562) nt	/	/
	dp220	dp220 + dp219	~730 (729) nt	/	/
Antisense	dp219	dp211 + dp72	~250 (246) nt	/	/
	dp219	dp208 + dp72	~580 (562) nt	/	/
	dp219	dp219 + dp210	~ 580 (577) nt	/	/
	dp219	dp219 + dp220	/ (729) nt	/	/
	dp208	dp208 + dp72	~ 580 (562) nt	+	24-27 nt
	dp211	dp211 + dp72	~250 (246) nt	/	/
	dp211	dp211 + dp71	~420 (419) nt	/	/

[a]The template was total RNA. PCR products separated on agarose gel are shown in **Figure S3**. PCR products were sequenced directly, without cloning.

[b]The range in U-tract length was determined by visual inspection of chromatograms, with the number of consecutive 'clean' T-peaks given as the lower boundary and clean peaks plus the subsequent mixed T-peaks (i.e., T-peaks superimposed by signals from other nucleotides) given as the upper boundary. This range reflects the heterogeneity between individual molecules within the amplicon population.

Table S6. Artifactual T-tract length variation[a]

Clone	T-tract length[a]
dp9540 (input)	31
dp11501-a	31
dp11501-b	29
dp11502	31
dp11503	31
dp11504-a	30
dp11504-b	30
dp11505	30
dp11506	29
dp11507-a	29
dp11507-b	28
dp11508	30
dp11509	30
dp11510	30
dp11511-a	29
dp11511-b	28
dp11512-a	29
dp11512-b	28

[a] Clones dp11501-dp11512 were derived from dp9540 by PCR using primer pairs dp211+dp220 and dp211+dp72 (see **Table S1)** and subsequent cloning of the amplicon population. -a, -b, inserts of clones with two inserts. The T-tract length variation between clones is similar to that seen in RT-PCR experiment, pointing to a PCR artifact (**Table 1**).

SUPPLEMENTARY FIGURES

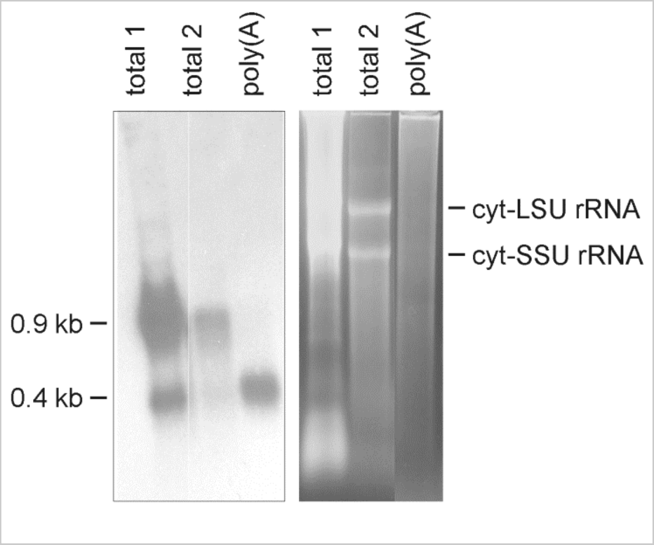


Figure S1. **Northern hybridization of mt-LSU rRNA from *Diplonema*.** Left panel, autoradiogram. Blotted total RNA ('total 1', ~3 µg; 'total 2', ~0.3 µg) and poly(A) RNA extracted from whole cells ('poly(A)', ~0.5 µg) were electrophoretically separated on a denaturating agarose gel, transferred to a membrane, and hybridized using the radiolabeled *rnl* module 2 as a probe. Right panel, ethidium-bromide stained RNAs prior to the transfer. cyt-LSU rRNA, cyt-SSU rRNA, cytosolic RNA species. As in most eukaryotes, *Diplonema* cytoplasmic LSU rRNA is bi-partite (28S + 5.8S), whereas the counterparts in trypanosomes and *Euglena* consist of multiple rRNA pieces.

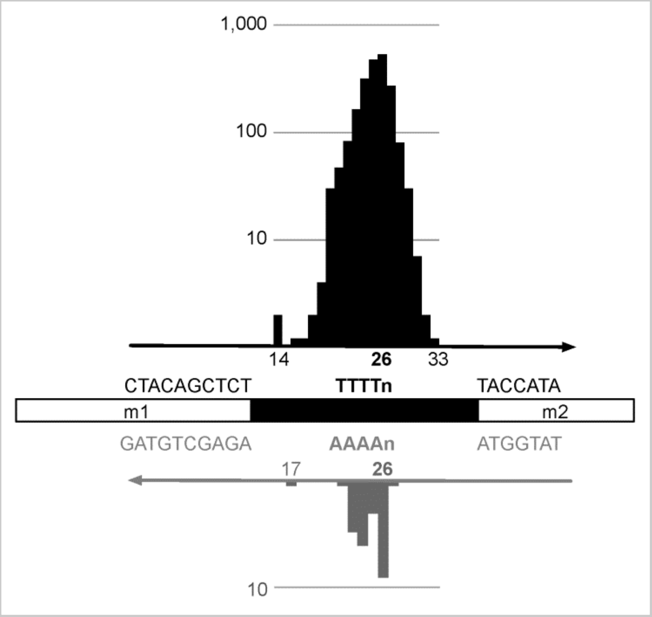


Figure S2. **U-tract length distribution of non-quality-trimmed RNA-Seq reads**. Individual reads from library F1 that bridge the *rnl* module-1/module-2 junction were analyzed here; the other libraries show similar distributions. The majority classes of sense and antisense reads contain 530 and 8 reads, respectively.

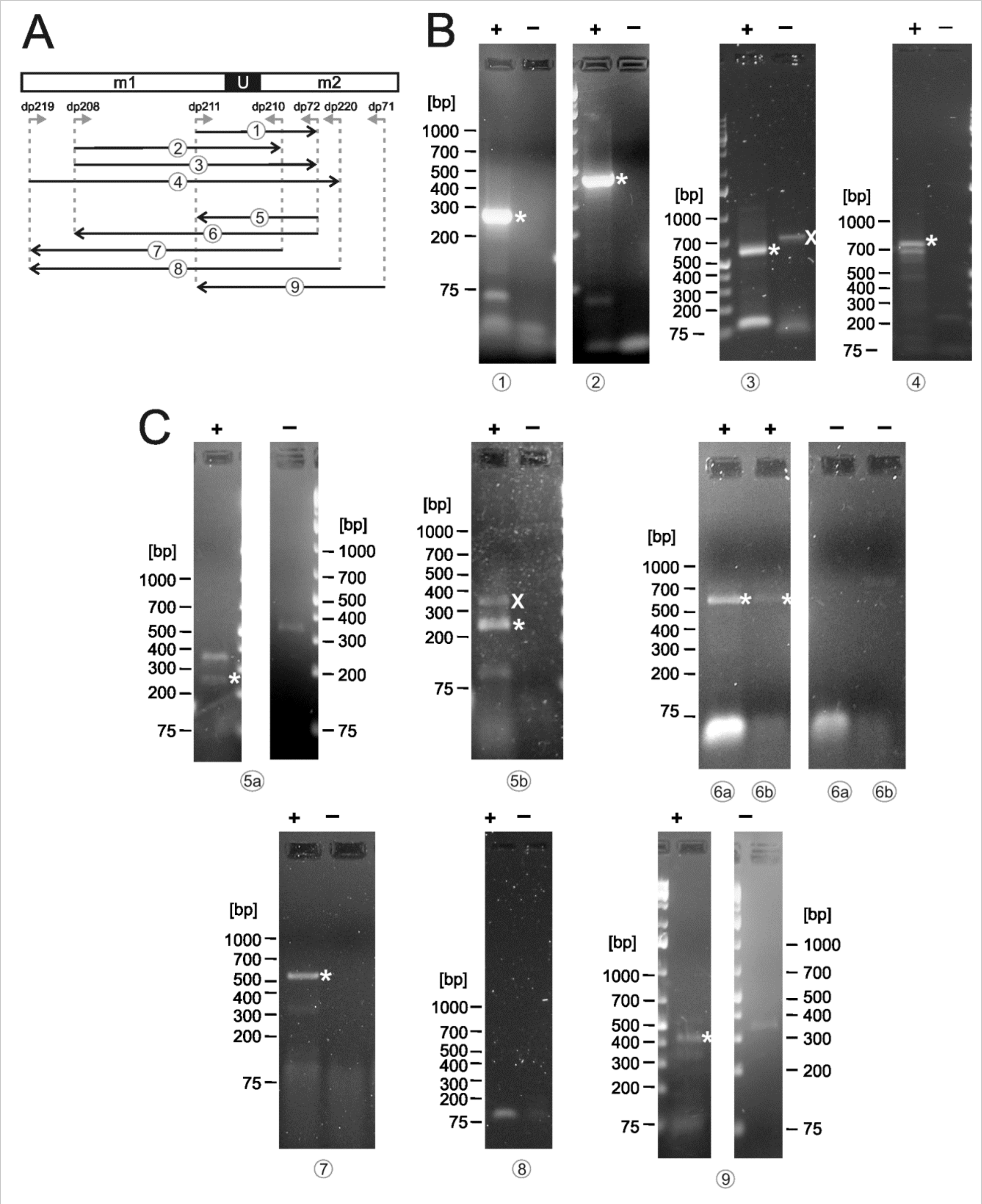
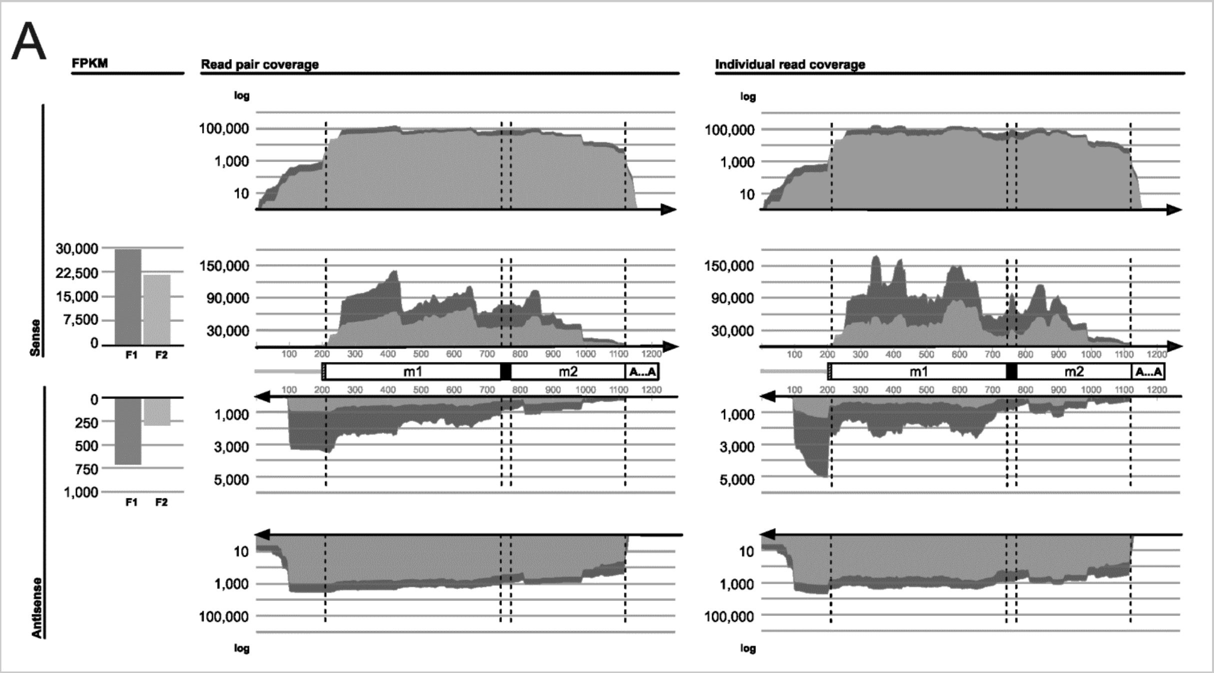


Figure S3. **RT-PCR across the U-tract of mt-LSU rRNA and its antisense transcripts. (A)** Schema of mt-LSU rRNA, positions of annealing primers (dp71 to dp220; for primers, see Table S1) and expected RT-PCR products labeled (1) – (9). **(B)** and (C) RT-PCR for detection of sense and antisense transcripts, respectively, using total RNA as a template. Size markers are shown in the margins. +, –, reverse transcriptase added or omitted, respectively. Primer combinations are (RT/PCR): (1), dp72/dp72+dp211; (2), dp210/dp210+dp208; (3), dp72/dp72+dp208; (4), dp220/dp220+dp219; (5a), dp219/dp211+dp72; (5b), dp211/dp211+dp72; (6a), dp208/

dp208+dp72; (6b), dp219/dp208+dp72; (7), dp219/dp219+dp210; (8), dp219/dp219+dp220; and (9), dp211/dp211+dp71. Asterisks indicate amplicons of expected size. Amplicons labeled by 'X' were sequenced and found to be artifacts due to unspecific annealing of primer dp72 to a partially complementary sequence in the constant region downstream of the module 1 gene. See also the Results section in the main text for further evidence of bi-directional transcription of mitochondrial chromosomes in *Diplonema*. The results of this figure are summarized in Table S5.



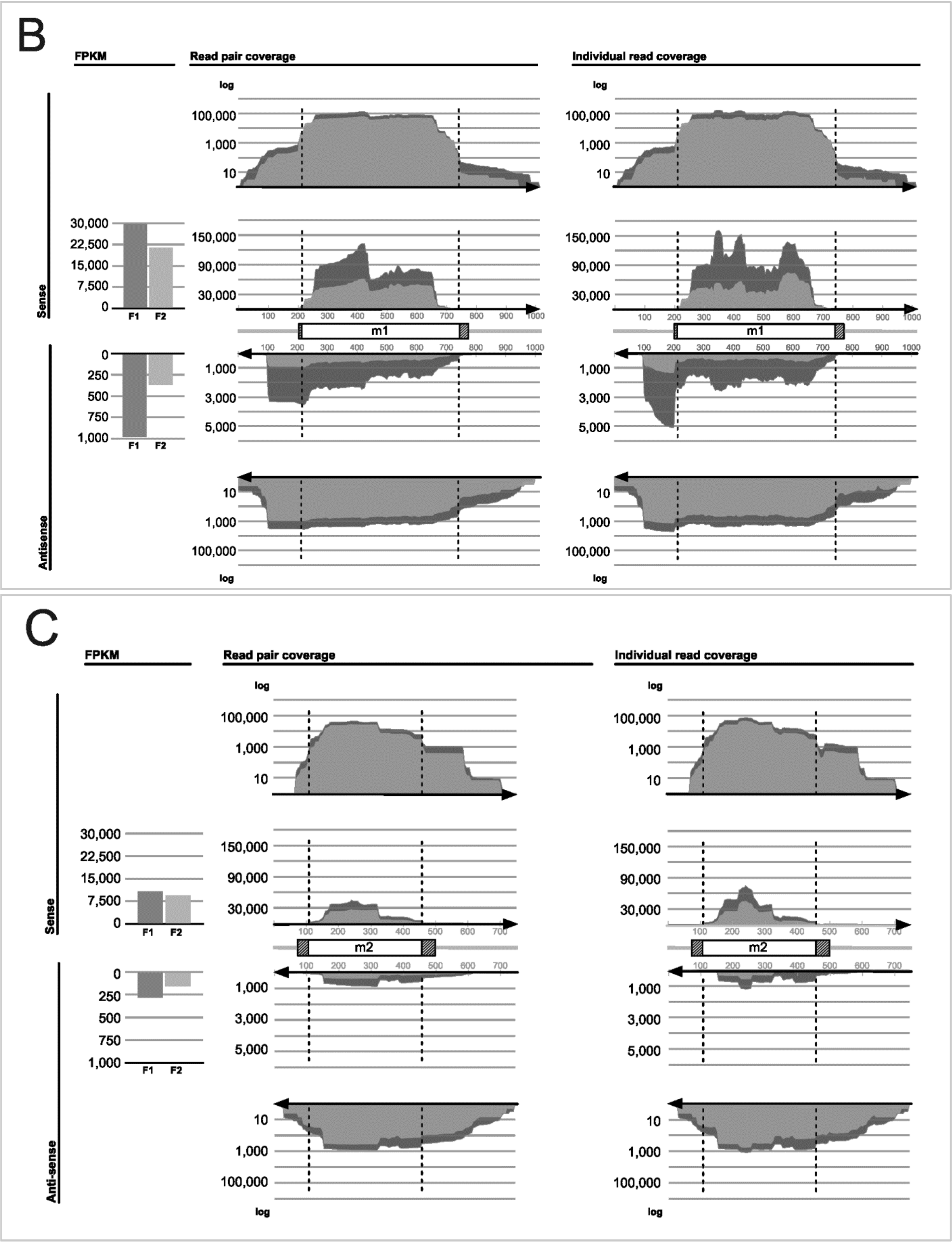
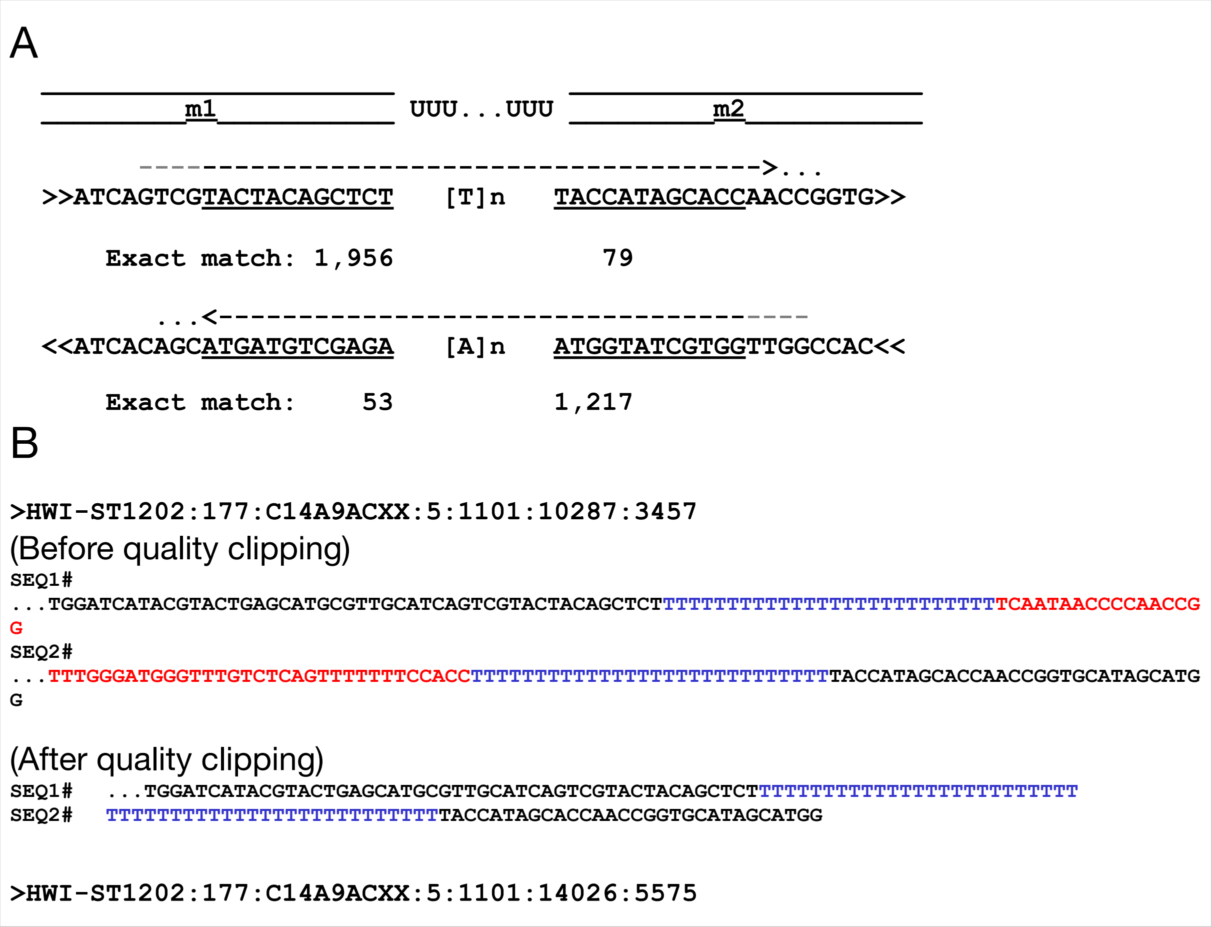


Figure S4. Coverage of RNA-Seq reads along reference transcript sequences of *rnl*. (A) Mitochondrial LSU rRNA. **(B)** *rnl* module 1 with non-coding, adjacent regions. **(C)** Module 2 with non-coding, adjacent regions. The

reference sequence is represented by the central cartoons, where the open and hashed boxes indicate coding and non-coding regions, respectively, and the grey line symbolizes constant regions of B-chromosomes (see Fig. 1B). Mapping to sense and antisense reference is indicated in the left margin. The stacked-area charts depict the coverage along the sequence (window size 5) and have linear and logarithmic vertical scales. The bar charts represent the total number of reads covering the region of the corresponding stacked-area chart. Note the 30-fold difference between the scales for sense and antisense transcripts.



```
A

 ______________________         _______________________
 __________m1__________ UUU...UUU __________m2___________

       ------------------------------------->...
 >>ATCAGTCGTACTACAGCTCT   [T]n   TACCATAGCACCAACCGGTG>>

     Exact match: 1,956              79

          ...<-------------------------------------
 <<ATCACAGCATGATGTCGAGA   [A]n   ATGGTATCGTGGTTGGCCAC<<

     Exact match:    53          1,217
B

>HWI-ST1202:177:C14A9ACXX:5:1101:10287:3457
(Before quality clipping)
SEQ1#
...TGGATCATACGTACTGAGCATGCGTTGCATCAGTCGTACTACAGCTCTTTTTTTTTTTTTTTTTTTTTTTTTTTCAATAACCCCAACCG
G
SEQ2#
...TTTGGGATGGGTTTGTCTCAGTTTTTTCCACCTTTTTTTTTTTTTTTTTTTTTTTTTTTTTACCATAGCACCAACCGGTGCATAGCATG
G

(After quality clipping)
SEQ1#   ...TGGATCATACGTACTGAGCATGCGTTGCATCAGTCGTACTACAGCTCTTTTTTTTTTTTTTTTTTTTTTTTTT
SEQ2#   TTTTTTTTTTTTTTTTTTTTTTTTTTTACCATAGCACCAACCGGTGCATAGCATGG


>HWI-ST1202:177:C14A9ACXX:5:1101:14026:5575
```

Figure S5. Effect of quality clipping on reads that bridge the *rnl* module-1/module-2 junction. (A) Analysis of not-quality-clipped reads. Top, cartoon depicting schematically the examined transcript region. Right- and left-pointing dashed arrows symbolize mate-1 and mate-2 reads, respectively. Underlined residues are those used for counting reads with exact sequence matches. >>, <<, 'regular' and reverse-complementary sequence, respectively. **(B)** Two typical read pairs mapping to the module-1/module-2 junction. Sequence before and after quality clipping. SEQ1#, SEQ2#, forward read and reverse read of the mate pair, respectively.

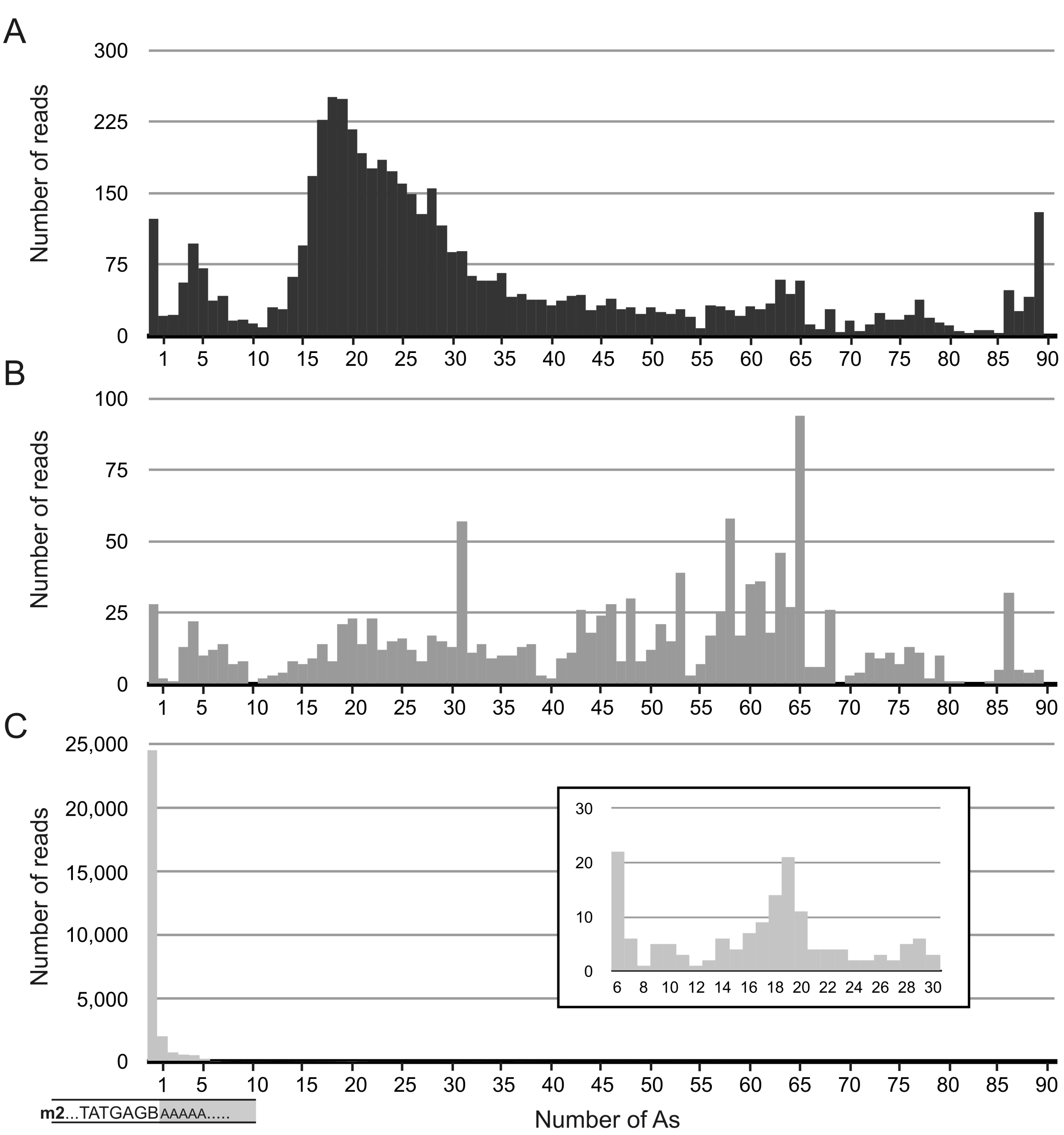


Figure S6. Length distribution of poly-(A) tails downstream of *rnl* module 2 in RNA-Seq reads. Reads contain at least 10 nt of the adjacent module. See also Table 1 of main text. **(A)** Reads from library F1; 96% of *rnl*-module 2 containing transcripts represent (mature) mt-LSU rRNA. The predominant length peaks at ~20 nt. **(B)** Reads from library PA; 94% of *rnl*-module 2 containing transcripts correspond to mono-module 2 (see text). A-tails >30 nt are ~3-times more abundant than tails <30 nt. The PA library contains less than 10% of mature mt-LSU rRNA than the F1 library, as inferred from the number of reads spanning the module junction. The cartoon in the left bottom corner depicts the pattern used for identifying the

reads, which includes the 10 terminal nucleotides of *rnl* module 2 followed by As. The last nucleotide in module 2 is permitted to be a B (C, G, or T in the IUB code), since this position is polymorphic, with a preponderance of T. **(C)** Reads from library GG. A total of 99.9% of *rnl*-module 2 containing transcripts represent mt-LSU rRNA. For estimates of the *rnl*-mono-module 2 to mt-LSU rRNA ratio, see Figures 1A and 4.

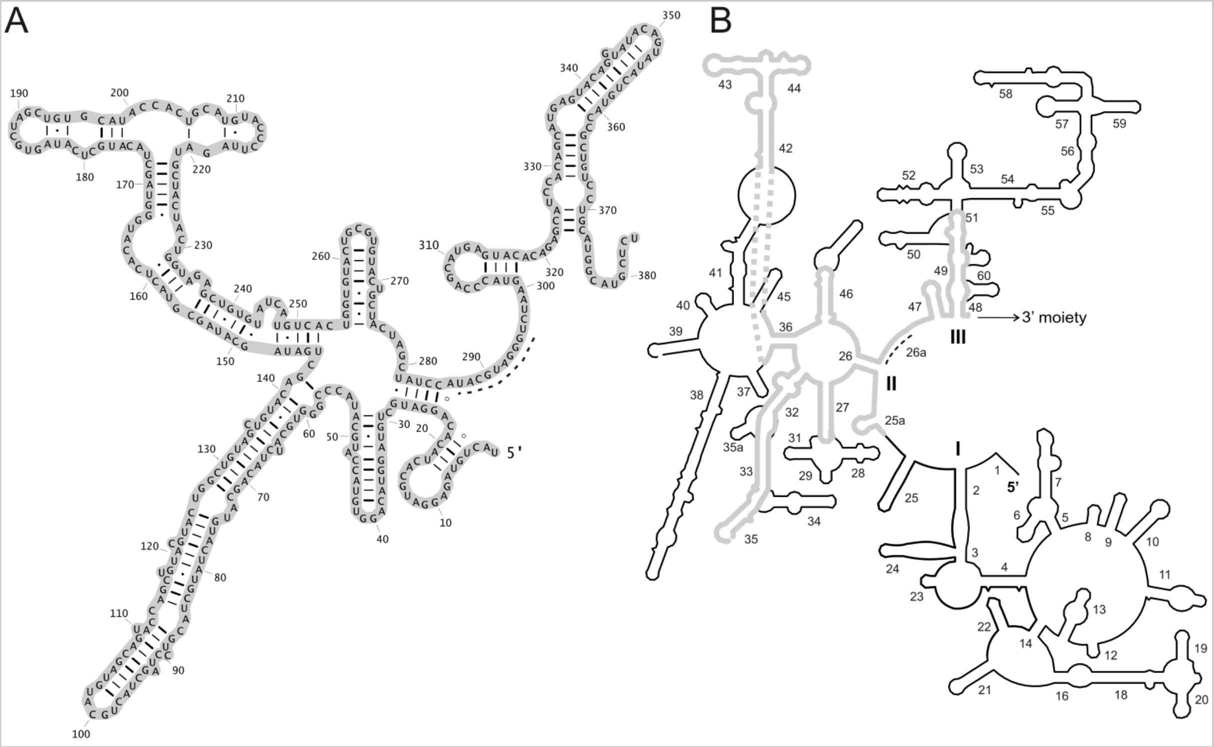


Figure S7. Potential 2° structure of the 5′ half of mt-LSU rRNA from *Diplonema*. (A) One of several possible structures is shown. In domains I-III of the mitochondrial consensus sequence and structure (http://www.rna.icmb.utexas.edu), sequence motifs are much shorter and less conserved than in domains IV-VI (3′ half), and 2°-structure motifs are considerably more variable. Therefore, the 5′ half of this molecule is hard to model in a system as derived as *Diplonema* mitochondria. Nucleotides identical to the reference mt-LSU sequence are shown in bold. For base pairing and other details, see Fig. 2. **(B)** The 2° structure from (A) mapped onto that of the 3' half from E. coli LSU rRNA. Helices are numbered as in Fig. 2. Thick grey lines indicate the structure elements present in the *Diplonema* model [same shading as in (A)]. Thin black dashed line indicates the position of H26a.

Chapitre 3: Novel modes of RNA editing in mitochondria

Sandrine Moreira*, Matus Valàch*, Mohamed Aoulad-Aissa, Christian Otto and Gertraud Burger

** These authors contributed equally to the paper as first authors.*

Reference: S. Moreira, M. Valach, M. Aoulad-Aissa, C. Otto, and G. Burger. 2016. "Novel Modes of RNA Editing in Mitochondria." *Nucleic Acids Research* gkw188.

Contribution des auteurs: GB a conçu et orchestré l'étude. SM et MV ont conçu et conduit les analyses bio-informatiques et l'approche expérimentale respectivement. MAA a aidé dans la préparation de l'ADN et de l'ARN, les expériences de PCR et la mise au point de la détection des inosines. GB, SM, MV et CO ont réalisé les analyse de données. Le manuscrit a été écrit par GB avec MV et SM

I.Abstract

Gene structure and expression in diplonemid mitochondria are unparalleled. Genes are fragmented in pieces (modules) that are separately transcribed, followed by the joining of module transcripts to contiguous RNAs. Some instances of unique uridine insertion RNA editing at module boundaries were noted, but the extent and potential occurrence of other editing types remained unknown. Comparative analysis of deep transcriptome and genome data from *Diplonema papillatum* mitochondria reveals ~220 post-transcriptional insertions of uridines, but no insertions of other nucleotides nor deletions. In addition, we detect in total 114 substitutions of cytosine by uridine and adenosine by inosine, amassed into unusually compact clusters. Inosines in transcripts were confirmed experimentally. This is the first report of adenosine-to-inosine editing of mRNAs and ribosomal RNAs in mitochondria. In mRNAs, editing causes mostly amino-acid additions and non-synonymous substitutions; in ribosomal RNAs, it permits formation of canonical secondary structures. Two extensively edited transcripts were compared across four diplonemids. The pattern of uridine-insertion editing is strictly conserved, whereas substitution editing has diverged dramatically, but still rendering diplonemid proteins more similar to other eukaryotic orthologs. We posit that RNA editing not only compensates but also sustains, or even accelerates, ultra-rapid evolution of genome structure and sequence in diplonemid mitochondria.

II.Introduction

DNA sequence alone does not always indicate what a genome encodes. One reason is RNA editing, the programmed alteration of a transcript, with the result that the RNA sequence differs from that of its genomic template. All kinds of transcripts can be affected by editing: mRNAs, intron RNAs, structural RNAs and regulatory RNAs. RNA editing plays an important role across the Tree of Life, and unsurprisingly, alterations in RNA editing can lead to human disease [207]. In the following, we will use the term 'RNA editing' for processes that change the sequence of a transcript, not including chemical modifications such as pseudouridylation, 2'-O methylation, etc. [208].

RNA editing is a post-transcriptional process that changes the sequence of the precursor transcript. RNA editing can act either on full-length transcripts or on nascent RNAs prior to 3′ end formation. This latter case has been referred to as 'cotranscriptional RNA editing' [209], although nucleotides are changed post-transcriptionally. Traditionally, cotranscriptional RNA editing describes a scenario discovered in myxomycete (slime mold) mitochondria where changes are intimately linked to RNA synthesis, and pre-edited (nascent) transcripts seem not to exist [100], [210]. Therefore, in a strict sense, the term 'RNA editing' does not apply to myxomycetes, because not the RNA sequence is changed but rather the DNA template is 'incorrectly' transcribed. In fact, the term RNA editing is often employed to generically describe differences in gene versus transcript sequences, although in many cases the origin of these changes remain unknown as in dinoflagellates [171], [211].

Post-transcriptional RNA editing is classified in three distinct types. The first type results in insertions or deletions (indels), by addition of new, or removal of existing, nucleotides in transcripts. The second type involves nucleotide substitutions, which are generated in situ by either deamination or (trans) amination, most commonly pyrimidine exchange (i.e. cytidine (C) to uridine (U; C-to-U) and U- to-C) and adenosine-to-inosine (A-to-I) deamination. Reverse transcriptases read Is in RNA as Gs, and similarly, the translation machinery is thought to interpret Is in mRNA as Gs [212]. In structural RNAs, A-to-I replacement has consequences as well. It influences RNA folding stability, and tRNAs extend codon recognition when the altered nucleotide is part of the anticodon. Note that for tRNAs, A-to-I deamination has been traditionally classified as nucleotide modification, but is now widely considered as RNA editing [213]. The third type of RNA editing causes nucleotide substitution as well, but acts exclusively on the 5′ and 3′ ends of the acceptor stem of mitochondrial tRNAs. In this case, mis-paired nucleotides are removed from one side of the helix and replaced by ones matching the complementary portion of the helix (reviewed in [214]). New types of post- transcriptional indel and substitution RNA editing are the topic of this work.

RNA editing has been discovered first in mitochondria [84]. It is quite common and highly diverse in this organelle. Post-transcriptional substitution of Cs by Us is most frequent, with land plant mitochondria featuring up to 2000 distinct events of this kind [215], [216]. Mitochondrial C-to- U editing is sporadically observed in other taxa, such as heteroloboseans [111], [217] and metazoans [218], [219]. Also, certain plastids perform C-to-U RNA editing [211]. Elsewhere, only a few such instances have been reported: notably one in an archaean tRNA [220] and a few dozen in metazoan nuclear mRNAs, nearly all within 3′ untranslated regions [221]. The prototype of mammalian C-to-U editing acts on apolipoprotein B (*apoB*) mRNA, and remains the only case of this type that impacts a coding region [102]. The inverse reaction, U-to-C substitution, occurs much more rarely than C-to-U, with the majority of sites in plant mitochondria [222]. Unheard of is A-to-I editing of organellar mRNAs or rRNAs, whereas this kind of substitution is pervasive in the metazoan nucleus [223], see also [224].

Mitochondria also perform post-transcriptional insertion and deletion RNA editing, which is extremely rare in other systems. The flagship organisms are kinetoplastids (Euglenozoa), where solely Us are inserted in, or deleted from, mitochondrial pre-mRNAs, up to nearly 600 in a single gene. Kinetoplastid indel editing involves site-specific cleavage of pre-mRNAs, U-insertion or deletion and religation. All steps are directed by small guide RNAs [225].

The sister clade of kinetoplastids is a group of ocean-thriving unicellular flagellates, the diplonemids. With only two genera recognized, *Diplonema* and *Rhynchopus*, diplonemids are seemingly an insignificant protist taxon. However, recent environmental explorations revealed that these organisms are among the most abundant and genetically most diverse eukaryotes in the oceans [164], [165], [226]. Diplonemids are notorious for their eccentric genome architecture and gene structure in mitochondria [44]. Specifically, mitochondrial genes of the type species *Diplonema papillatum* are systematically split in up to 11 pieces (modules) that are ~40–550 nt long. Each such piece is encoded on one of the ~80 distinct circular chromosomes of 6 kbp (class A) or 7 kbp (class B) length.

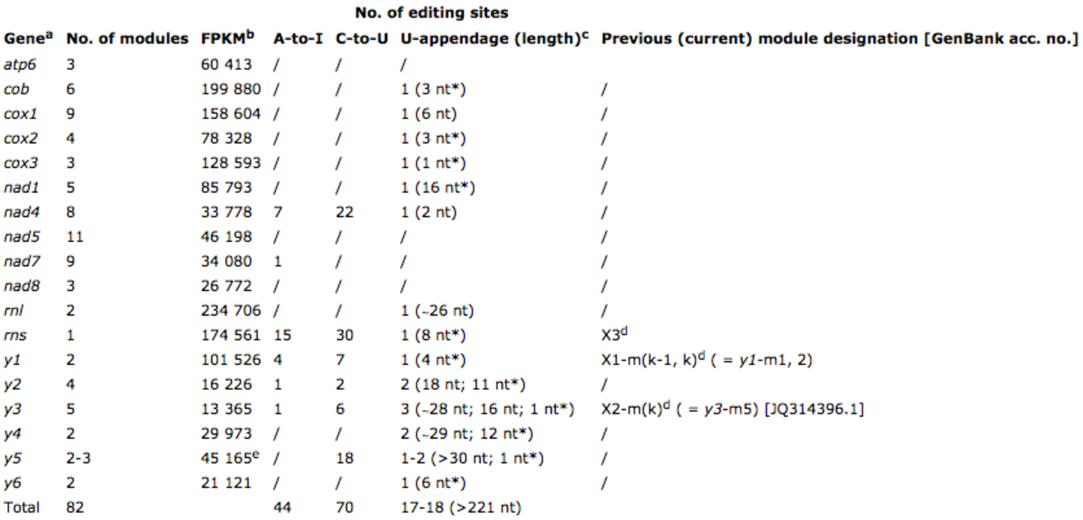
Chromosomes have a surprisingly regular structure (**Supplementary Figure S1A**; [45]). Coding regions are flanked by on average 50-nt unique sequence and together, they make up a distinctive cassette that is unique to a given chromosome. The rest of the circle (~90%) mostly consists of repeats. Specifically, adjacent to each cassette are two 'class-specific constant regions' of 1–3 kbp whose sequence is conserved across all chromosomes of a given class. In addition, opposite to the cassette resides a ~2.5-kbp 'shared constant region', which is common to A- and B-class chromosomes [45].

Gene modules in *Diplonema* mitochondria are transcribed separately as RNA precursors, then end-processed, and subsequently joined into contiguous RNAs [174]. The molecular mechanism of this unique trans-splicing process is yet to be unraveled. Collectively, modules specify a relatively 'standard' set of 12 recognized genes, including two ribosomal RNAs (mt-rRNAs) as well as protein components of the respiratory chain, oxidative phosphorylation and the mito-ribosome (**Table 1, column 1**); as in kinetoplastids, tRNAs appear to be imported from the cytosol.

In diplonemids, we previously noted a mode of mitochondrial RNA editing that somewhat resembles U-insertion editing in kinetoplastid mitochondria, as it involves the addition of multiple Us at 3′ ends of modules (therefore termed 'U-appendage' editing). For example, the module 4- transcript of the gene encoding cytochrome c oxidase subunit I (*cox1*) is extended by six Us that are retained in the trans-spliced mRNA, consisting of a total of nine modules [44], [168]. An even more spectacular U-appendage occurs during maturation of the mitochondrial large-subunit ribosomal RNA (mt-LSU rRNA). The corresponding gene (*rnl*) is split into two modules. At the 3′ end of the *rnl* module 1-transcript, ~26 Us are added prior to trans-splicing. We showed that the U-tract-containing mt-LSU rRNA is indeed incorporated into the mito-ribosome of *D. papillatum* [227].

Here we examine comprehensively RNA editing in *D. papillatum* based on deep transcriptome sequencing data, uncovering a second type of post-transcriptional RNA editing in diplonemid mitochondria: nucleotide substitution. Remarkably, replacements include A-to-I substitutions in mRNAs and rRNAs, which has never been seen in organelles before. These nucleotide changes will be investigated experimentally. A second focus of this study is on the conservation and diversification of RNA editing pattern during the evolution of diplonemids, and possible evolutionary relationships between RNA editing in diplonemid mitochondria and those in other systems.

Table 1. Genes and RNA editing sites in *D. papillatum* mitochondria



			No. of editing sites			
Gene[a]	No. of modules	FPKM[b]	A-to-I	C-to-U	U-appendage (length)[c]	Previous (current) module designation [GenBank acc. no.]
atp6	3	60 413	/	/	/	
cob	6	199 880	/	/	1 (3 nt*)	/
cox1	9	158 604	/	/	1 (6 nt)	/
cox2	4	78 328	/	/	1 (3 nt*)	/
cox3	3	128 593	/	/	1 (1 nt*)	/
nad1	5	85 793	/	/	1 (16 nt*)	/
nad4	8	33 778	7	22	1 (2 nt)	/
nad5	11	46 198	/	/	/	/
nad7	9	34 080	1	/	/	/
nad8	3	26 772	/	/	/	/
rnl	2	234 706	/	/	1 (~26 nt)	/
rns	1	174 561	15	30	1 (8 nt*)	X3[d]
y1	2	101 526	4	7	1 (4 nt*)	X1-m(k-1, k)[d] (= *y1*-m1, 2)
y2	4	16 226	1	2	2 (18 nt; 11 nt*)	/
y3	5	13 365	1	6	3 (~28 nt; 16 nt; 1 nt*)	X2-m(k)[d] (= *y3*-m5) [JQ314396.1]
y4	2	29 973	/	/	2 (~29 nt; 12 nt*)	/
y5	2-3	45 165[e]	/	18	1-2 (>30 nt; 1 nt*)	/
y6	2	21 121	/	/	1 (6 nt*)	/
Total	82		44	70	17-18 (>221 nt)	

[a]Gene products are: *atp6*, subunit 6 of ATP synthase; *cob*, apocytochrome
mt-SSU rRNA. Gene products of *y1-y6* are unknown; *y1-y4* code for proteins. GenBank accession numbers of transcripts determined here are listed in Supplementary Table S6.

[b]Fragments Per Kilobase of transcript per Million of mapped reads in library DPA2 made from poly(A) RNA, determined by Star/Cufflinks (see 'Materials and Methods' section).

[c]Asterisks indicate terminal modules.

[d] (30).

[e]FPKM average over three transcript variants, differing in their 3′-terminal region; see Figure 1.

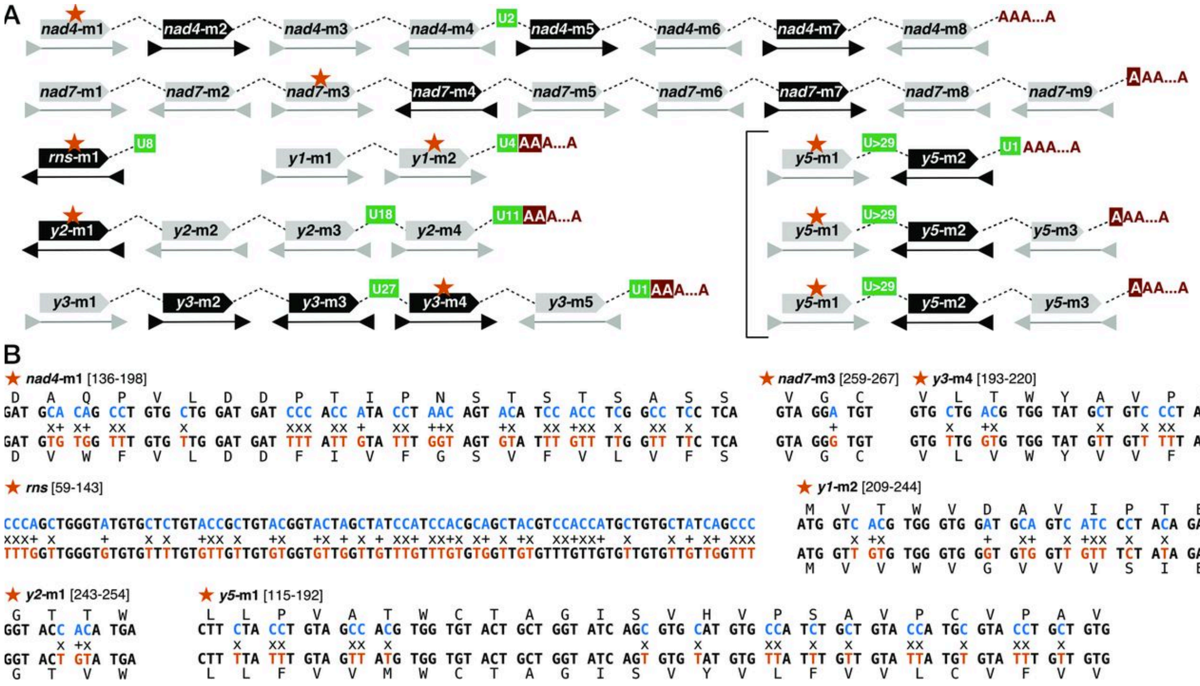

A
nad4-m1
nad4-m2
nad4-m3
nad4-m4
U2
nad4-m5
nad4-m6
nad4-m7
nad4-m8
AAA...A
nad7-m1
nad7-m2
nad7-m3
nad7-m4
nad7-m5
nad7-m6
nad7-m7
nad7-m8
nad7-m9
AAA...A
rns-m1
U8
y1-m1
y1-m2
U4 AAA...A
y5-m1
U>29
y5-m2
U1 AAA...A
y2-m1
y2-m2
y2-m3
U18
y2-m4
U11 AAA...A
y5-m1
U>29
y5-m2
y5-m3
AAA...A
y3-m1
y3-m2
y3-m3
U27
y3-m4
y3-m5
U1 AAA...A
y5-m1
U>29
y5-m2
y5-m3
AAA...A
B
nad4-m1 [136-198]
D A Q P V L D D P T I P N S T S T S A S S
GAT GCA CAG CCT GTG CTG GAT GAT CCC ACC ATA CCT AAC AGT ACA TCC ACC TCG GCC TCC TCA
GAT GTG TGG TTT GTG TTG GAT GAT TTT ATT GTA TTT GGT AGT GTA TTT GTT TTG GTT TTC TCA
D V W F V L D D F I V F G S V F V L V F S
nad7-m3 [259-267]
V G C
GTA GGA TGT
GTA GGG TGT
V G C
y3-m4 [193-220]
V L T W Y A V P T
GTG CTG ACG TGG TAT GCT GTC CCT ACC
GTG TTG GTG TGG TAT GTT GTT TTT ACC
V L V W Y V V F T
rns [59-143]
CCCAGCTGGGTATGTGCTCTGTACCGCTGTACGGTACTAGCTATCCATCCACGCAGCTACGTCCACCATGCTGTGCTATCAGCCC
TTTGGTTGGGTGTGTGTTTTGTGTTGTTGTGTGGTGTTGGTTGTTTGTTTGTGTGGTTGTGTTTGTTGTGTTGTGTTGTTGGTTT
y1-m2 [209-244]
M V T W V D A V I P T E
ATG GTC ACG TGG GTG GAT GCA GTC ATC CCT ACA GAG
ATG GTT GTG TGG GTG GGT GTG GTT GTT TCT ATA GAG
M V V W V G V V V S I E
y2-m1 [243-254]
G T T W
GGT ACC ACA TGA
GGT ACT GTA TGA
G T V W
y5-m1 [115-192]
L L P V A T W C T A G I S V H V P S A V P C V P A V
CTT CTA CCT GTA GCC ACG TGG TGT ACT GCT GGT ATC AGC GTG CAT GTG CCA TCT GCT GTA CCA TGC GTA CCT GCT GTG
CTT TTA TTT GTA GTT ATG TGG TGT ACT GCT GGT ATC AGT GTG TAT GTG TTA TTT GTT GTA TTA TGT GTA TTT GTT GTG
L L F V V M W C T A G I S V Y V L F V V L C V F V V

Figure 1. Substitution RNA editing of mitochondrial genes from *Diplonema papillatum*. (**A**) Module composition of edited genes. Gray and black pentagons represent modules encoded on A- and B-class chromosomes, respectively. Arrows under pentagons indicate the orientation of the module within the chromosome (see Supplementary Figure S1A). Orange stars point to clusters of substitution RNA editing. Green boxes depict post-transcriptional U-insertions with the indicated number of appended Us. 'AAA...A', poly(A) tail. As being part of a stop codon are shown on red background. The square bracket groups transcript isoforms of *y5*. (**B**) Sequences of substitution editing clusters. Numbers in square brackets specify the position of the depicted sequence in the corresponding cassette. Upper rows, genomic and lower rows cDNA-derived nucleotide sequence and conceptual translation (one-letter code). +, x: A-to-G and C-to-T sites, respectively. Blue and orange nucleotides, pre-edited and edited states of substitution sites, respectively.

III.Materials and methods

Detailed descriptions of applied methods are available in **Supplementary Materials and Methods**.

1. Strains, culture, and DNA and RNA extraction

Diplonema papillatum (ATCC 50162), *Diplonema ambulator* (ATCC 50223), *Diplonema* sp. 2 (ATCC 50224) and *Rhynchopus euleeides* (ATCC 50226) were obtained from the American Type Culture Collection. Organisms were cultivated axenically as described earlier [168], [227]. To isolate mtDNA, mitochondria were enriched by differential and sucrose gradient centrifugation. Mitoribosomes were separated from whole cell lysates by kinetic glycerol-gradient ultracentrifugations [227]. After extraction of RNA [228], residual DNA was removed by column purification or digestion with RNase-free DNase followed by phenol-chloroform extraction. Poly(A) RNA was enriched by a passage through oligo(dT)-cellulose.

2. *In vitro* transcription and RNase cleavage of glyoxalated RNA

The DNA templates for in vitro transcription of synthetic pre-edited or edited mt-SSU rRNA were generated either by PCR on mtDNA or by RT-PCR on purified mt-SSU rRNA. To detect inosines in RNA, we followed a protocol devised by others [229] that exploits the fact that guanosines, but not inosines, can be modified by glyoxal/borate treat- ment, which protects against RNase T1 cleavage [230]. After glyoxalation and RNase T1-treatment, the RNA sample was deglyoxalated and then used in RT-PCR, northern blot hybridization, or primer extension assays. Oligonucleotides used as primers and hybridization probes are listed in Supplementary Table S1. For details, including deviations in electrophoretic migration behavior of certain RT-PCR products, see **Supplementary Methods** and https://www.protocols.io/u/matus-valach.

3. DNA library construction, sequencing, read processing and assembly

A genomic paired-end library was constructed from total DNA and sequenced with Illumina MiSeq. Details on libraries are compiled in **Supplementary Table S2**. Cutadapt version 1.2.1 (http://journal.embnet.org/index.php/embnetjournal/article/view/200) was employed for adapter clipping, quality trimming and elimination of reads shorter than 20 nt. Reads were assembled with the Celera software runCA version 8.3rc2 (http://sourceforge.net/projects/wgs-assembler/ [231]) using default parameters. Contigs originating from the mitochondrial genome were identified by sequence identity with previously determined mitochondrial chromosomes and modules (GenBank acc. nos. EU123536- 8 and HQ28819-33) using BLAST at a hit-reporting threshold of 99% and then clustered with CD-HIT version 4.6 [232] employing the option -c 0.9.

4. RNA-Seq library construction, sequencing and read processing

We depleted *D. papillatum* RNA from cytosolic rRNAs and mt-LSU rRNA (32) using biotinylated oligonucleotides complementary to these rRNA species. Libraries were prepared from total cellular RNA enriched for poly(A) RNA (PA, DPA2), mitochondrial RNA (F1 from a fragmented, F2 from an un-fragmented sample) and a mito-ribosome- enriched RNA fraction (GG). Libraries from the three other diplonemids were made from total RNA depleted from cytosolic rRNAs. For details on libraries, see **Supplementary Table S2**.

5. Mapping of Illumina reads to reference sequences and calculation of FPKM

Illumina reads were mapped to reference sequences with Bowtie 2 [130]. If not specified otherwise, we employed the options –local –no-unal (removing unaligned reads), and default values for the alignment and scoring parameters. Output files in sam format were subsequently transformed into '.bam' files with SAMtools v1.4 (http://samtools.sourceforge.net/). Alignments were visualized with the Integrative Genomes Viewer (IGV; https://www. broadinstitute.org/igv/) [233]. FPKM (Fragments Per Kilobase of transcript per Million of mapped reads) values were obtained after mapping RNA-Seq reads of the library DPA2 against the mito-transcriptome reference using TopHat v2.0.14 (https://ccb.jhu.edu/software/tophat/index. shtml) [234] or STAR version 2.4.2a (https://github.com/alexdobin/STAR) [235] with default parameters, followed by assembling mapped reads into contigs with Cufflinks (https: //github.com/cole-trapnell-lab/cufflinks).

6. Mitochondrial reference genome sequence

From a Celera assembly of mitochondrial-genomic MiSeq Illumina reads (see above), we extracted contigs holding cassettes, i.e. those containing the distinctive left-hand and right-hand class-specific constant regions of chromosomes, but not the shared constant region (see **Supplementary Figure S1A**). The contig sequences were validated and polymorphisms determined simultaneously by mapping back the MiSeq reads to the contigs with Bowtie 2.

7. Transcript de novo assembly and function annotation

RNA-Seq reads from the PA and DPA2 libraries (**Supplementary Table S2**) were assembled using SOAPdenovoTrans [236] using various kmers, and the resulting contigs were assembled again using SOAPdenovo [237] with a kmer size 127. We also used Trinity with default parameters (http://trinityrnaseq.github.io/, [137]). Mitochondrial rRNA sequences were assembled from reads of library GG. Function assignment of newly detected mitochondrial transcripts was attempted with various approaches (see **Supplementary Methods**), but failed.

8. Mitochondrial transcriptome reconstruction and assignment of orphan modules

Transcripts were also reconstructed via a split-read approach developed in-house. First, RNA-Seq reads of the poly(A) library were mapped to the genome reference with Bowtie 2 (paired-end and local mode). The custom python script, findTransSplicedRNA.py, then identified read pairs whose two partners do not map to same genomic module and analyzed their sequence portions that were soft-clipped during mapping in local mode. If the soft-clipped sequence overlaps with another genomic module, then the two modules must belong to the same gene and be adjacent in the trans-spliced transcript.

9. *In silico* identification of polymorphic genomic sites

Variant sites in mtDNA were determined with the UnifiedGenotyper module of the Genome Analysis Toolkit (GATK) v3.3.1 (https://www.broadinstitute.org/gatk/ [127], [238]) and FreeBayes (Garrison E and Marth G. (2012) In arXiv (ed.), Vol. 1207.3907v2 [q-bio.GN]). For both tools we set the ploidy to 100, the minimum number of observed variants to 2, the minimum base quality and mapping quality to 30 and the minimum allele frequency to 0.01. The output files of DNA–DNA comparison is referred to as DDd.vcf files. Only sites with at least 10% allele frequency were considered. We validated the obtained genome variants by visual inspection with IGV v2.3.40 (https://www.broadinstitute.org/igv/) [233], as well as with reads generated by Sanger and 454-FLX (Roche). For linked sites, we consolidated the allele frequencies reported by the variant caller software by calculating the mean across all linked sites.

10.*In silico* identification of RNA editing sites

We mapped RNA-Seq reads from all libraries against the genome reference, the pre-edited (virtual) and the fully edited transcriptome, using Bowtie 2 (see above) in strand-specific mode. The returned bam files were merged by the custom script mergeSAM.py. To identify DNA-RNA differences (DRds), the merged bam file was used as input for the GATK UnifiedGenotyper [127] and FreeBayes, using the same parameters as for calling genomic variants described above. To differentiate between genomic polymorphisms and RNA editing sites, DRd.vcf and DDd.vcf files were compared with the tool vcf-isec of the VCFtool kit (https://vcftools.github.io/perl_module.html#vcf-isec). The called sites were inspected and validated visually in the bam files.

11.Analysis of RNA processing intermediates and partially edited transcripts

Using the in-house script editpop.py (see **Supplementary Methods**), we analyzed the correlation between the status of RNA editing sites, and between RNA editing, module processing and trans-splicing in read pairs. U-appendage sites were analyzed by searching motifs in individual reads using GNU grep. For short internal U-insertions (1–2 nt) we requested a full match of ≥6 adjacent nucleotides in the two neighbor modules. For terminal Us, the search requested eight adjacent nucleotides in the module's 3′ end followed by ≥2 Us. To detect partially edited substitution sites, RNA-Seq reads from library DPA2 were mapped with Bowtie 2 against the pre-edited (virtual) and edited transcriptome sequences, the resulting sam alignment files were merged as described above and then parsed with the in-house script editedSitesStat.py (see **Supplementary Methods**) to extract the particular nucleotides present at RNA editing sites. Site positions were obtained from the vcf file Dp_mito_SNP-RNA_20160212.vcf (**Supplementary File 1**).

12.Search for *cis* elements and RNA trans-factors that guide RNA editing

We searched for recurrent cis-motifs near individual RNA editing sites using the in-house script editbysite.py (see **Supplementary Methods**). Sequence motifs in *cis* that flank clusters of substitution editing were searched using MEME and GLAM2 (MEME web server [239] (http://meme-suite.org/tools/meme) for ungapped and gapped motifs, respectively. Common 2D *cis*-motifs were searched using RNAalifold [179], [240], [241], LocARNa [242], [243], RNAstructure and Multilign [240], [241], with default parameters. *Trans*-acting guide

RNAs were searched in reads of library F2 (Supplementary Table S2) employing the GNU grep utility supplied with query motifs that are reverse-complements of the edited sequence and adjacent regions. Finally, for detection of anti-sense reads with mismatches, reads of the F2 library were mapped against the sequences of pre-edited (virtual) and full-length mature transcripts, using Bowtie2 with the options –local –nofw and default mismatch settings. Resulting bam files were inspected visually.

13.Analysis of Nad4 protein sequences

Using MUSCLE with default parameters [234], we built a multiple alignment of Nad4 protein sequences deduced from edited and pre-edited *nad4* genes from four diplonemids and 15 moderately divergent homologs from other taxa (see **Supplementary Methods**). From this alignment, we extracted a sub-alignment including columns 1–130, which corresponds to the N-terminal region up to the last 'edited' amino acid in diplonemids. Based on this alignment, protein conservation was determined with MstatsX (https://github.com/gcollet/MststX) which uses the trident statistics [235]. To determine potential trans-membrane helices, the protein sequences of 'pre-edited' and 'edited' diplonemid Nad4 variants were scanned with TMHMM2.0 [244] (http://www.cbs.dtu.dk/services/TMHMM/), Phobius [245] and TMpred from the ExPASy web suite [246].

IV.Results

1. Assignment of modules to genes

As a first step, we compiled all cassette sequences of *D. papillatum*, i.e. the unique portions of mitochondrial chromosomes (**Supplementary Figure S1A**). Cassettes were extracted from a whole genome assembly, built with Illumina reads. These consensus sequences, 23 kbp in total, include 81 different cassettes of 166–638 nt length. In an earlier assembly [247], the cassette/module count was 75. We then determined genomic variants by mapping reads to the consensus sequences and analyzing the alignment maps by variant calling software tools. We detected nearly 100 genomic variants (with ≥10% allele frequency) spread across 37 cassettes. Variants represent 80% transitions and 20% transversions, and are mostly biallelic and linked, suggesting two versions for each chromosome (**Supplementary Table S3**). Finally, the consensus sequences were corrected so that they represent the majority of reads, for use as mitochondrial reference genome in the following analyses.

Mapping of RNA-Seq reads against the *Diplonema* mito-genome reference confirms that all predicted modules are transcribed (including allelic variants). The majority of modules are pieces of 11 previously reported, assigned mitochondrial genes. For nearly 20 modules (designated orphans), the genes they belong to were unknown at the outset of this study (e.g. modules X1 to X3 [174]; **Table 1**).

To pinpoint the mature transcripts to which orphan gene modules belong, we assembled transcripts from RNA-Seq reads, yielding a 15-kb mitochondrial transcriptome. Orphan modules were assigned to transcripts based on sequence identity, revealing seven new genes: X3 consists of a single module, whereas the others (denoted *y1, y2*, etc.) are composed of two to five modules. All mitochondrial genes and their modules detected to this point are listed in **Table 1** and depicted in Supplementary **Figure S2**.

2. Identification of the tentative gene for mt-SSU rRNA

Functional annotation of newly discovered transcripts was attempted initially by BLAST similarity searches in GenBank's non-redundant database, followed by searches with profile Hidden Markov Models and Covariance Models representing all known mitochondrial protein-coding genes and mitochondrial rRNAs, respectively (see 'Materials and Methods' section). However, no significant hit was obtained.

Gene X3 stands out because its transcript is highly abundant, with a steady-state level comparable to *Diplonema*'s mt-LSU rRNA. In addition, the transcript is highly enriched in the library made from a mito-ribosome-enriched fraction (**Supplementary Table S2**), suggesting that X3 represents the elusive mitochondrial small subunit (mt-SSU) rRNA. Indeed, highly divergent structural domains (5′- and 3′-minor domains) of the SSU-RNA are recognizable (**Supplementary Figure S1B**), although the remainder of the secondary (2D) structure could not be modeled, and this is due to several reasons. One is that sequence similarity is very low between X3 from *Diplonema* and mt-SSU rRNAs from other organisms for which secondary structure models are available. Further, thermodynamics-based modeling is inconclusive, because high G + C content of the sequence generates numerous equally probable alternatives. The same issues, but to a lesser degree, were encountered when modeling mt-LSU-rRNA from *D. papillatum* [227].

The tentative (366 nt) mt-SSU rRNA of *Diplonema* is among the shortest ever reported, slightly shorter than the highly derived mitochondrial *rns* in certain animals [248].

Yet, as of now, it cannot be ruled out that X3 represents only one of several molecules of a mitochondrial rRNA in pieces, as seen in *Euglena gracilis* [40], apicomplexans [249] and dinoflagellates [171], for example.

The *y* genes still remain unidentified. For *y1* to *y4*, we have mass-spectrometry data demonstrating that these genes code for proteins (data not shown). Unravelling the biological role of *y*-genes will require detailed biochemical studies.

3. Uncovering RNA editing events in mitochondrial transcripts

RNA editing sites manifest as differences between gene and transcript sequence. DNA–RNA-differences will be referred to in the following as DRds. Since mitochondrial genes in *Diplonema* are fragmented, we used as a reference sequence the joined gene pieces (equivalent to pre-edited full-length transcript sequences), against which we mapped RNA-Seq reads from the various libraries (**Supplementary Table S2**). Based on the genome/ transcriptome alignment maps, we determined DRds using variant calling tools, and all these sites were visually inspected and validated. Note that 'pre-edited' refers to sites not yet edited, while 'unedited' characterizes sites that are never edited.

Two positions returned as editing sites coincide with genomic variants determined earlier. One is located in *y2*-module number 3 (*y2*-m3) (position 131 in the corresponding cassettes). It is an A/G dimorphism with a ratio of 4:6 in DNA versus 1:9 in RNA. The second site falls in *rns* with a C/T dimorphism of 2:8 in DNA (see **Supplementary Table S3**), but only U in RNA. In both cases, it cannot be distinguished whether the dimorphic sites in RNA arise from transcription of the two genomic variants or rather by transcription of the A- and C-alleles with subsequent partial RNA editing to G and T (U), respectively.

Table 1 summarizes the number and types of RNA editing sites that are frequently edited (>50%) and do not coincide with genomic variants. Positions and frequencies of sites displaying at least 5% editing are listed in **Supplementary File 1**.

4. Catalog of nucleotide insertions and substitutions in the D. papillatum mito-transcriptome

Inspection of the global landscape of RNA editing in *D. papillatum* mitochondria (**Table 1**) reveals two important features. First, post-transcriptional insertions involve only Us and no other nucleotide. Not a single event of deletion-RNA editing was detected in the 15-kb transcriptome examined. Inserted U-tracts are between 1 and ~30 nt long and coincide with module junctions or transcript ends. Earlier we demonstrated experimentally for *cox1* and mt-LSU rRNA that Us inserted at module junctions in the mature transcript are actually appended at the 3′ end of the upstream module prior trans-splicing [174], [227]. Our comprehensive mito-transcriptome data corroborate the 3′ appendage mechanism, since no cases of U-extensions at 5′ ends of full-length transcript were observed. Finally, 3′ U-additions are independent of the module position (5′, internal or ultimate) within the mature transcript.

The second important feature is that *Diplonema* mitochondria perform substitution RNA editing (**Table 1**), which had remained unnoticed until now. We observe more than 110 C-to-U and A-to-I substitutions. Except for one solitary substitution in *nad7*, sites are congregated in six clusters of 3–85 nt length, which are located in *nad4*, the tentative *rns*, and four y genes (*y1, 2, 3, 5*; **Figure 1A**). Most densely packed are the substitution editing clusters in *nad4* and the tentative *rns* with sites often placed immediately adjacent to one another. The cluster in *nad4* is 56 nt long and positioned in module 1, close to the 5′ terminus. In this transcript, a total of 29 sites undergo substitutions, notably all Cs and one half of As. The 85-nt long cluster in the tentative *rns* is also situated near the transcript's 5′ end and includes 45 sites; all Cs and As in this cluster are edited (**Figure 1B**).

5. Searches for potential cis-elements and trans-factors that guide RNA editing

We scrutinized the sequence context of mitochondrial RNA editing sites in *Diplonema* by searching for shared sequence motifs and 2D structure element (*cis*-motifs) in close vicinity of the locations where RNA editing occurs. These analyses were performed separately for sites of U-addition, A- to-I substitution, and C-to-U substitution, and for substitution clusters. However, we did not detect motifs specifically associated with either type of RNA editing site (**Supplementary Figure S3A–C**). The absence of discernable recurrent sequence patterns around RNA editing sites suggests that these sites are defined by specific *trans* factors.

We searched for potential *trans*-acting RNAs that guide RNA editing, postulating a population of site-specific factors with the propensity to pair with RNA editing sites. But again, convincing candidates were not detected (**Supplementary Figure S3D**). For both *cis*-elements and *trans*-factors, our search strategies, results and interpretations are detailed in '**Supplementary Results and Discussion**'.

6. Biochemically, A-to-I substitution RNA editing proceeds by deamination

C-to-U and A-to-G differences between genomic and cDNA sequences usually originate from *in situ* base deamination in transcripts, but nucleotide excision and replacement is conceivable as well. In the case of deamination, a substituted G in cDNA corresponds to an inosine (I) in RNA. Therefore, we determined the presence of Is in the transcripts of two function-assigned genes that undergo substitution RNA editing in *Diplonema*, *nad4* and the tentative *rns*. We treated RNA from *Diplonema* with glyoxal, which forms a stable adduct with G, but not I, in the presence of borate [230]. RNase T1 then cleaves RNA after (unmodified) Is, while glyoxalated Gs are protected (**Figure 2A–C**).

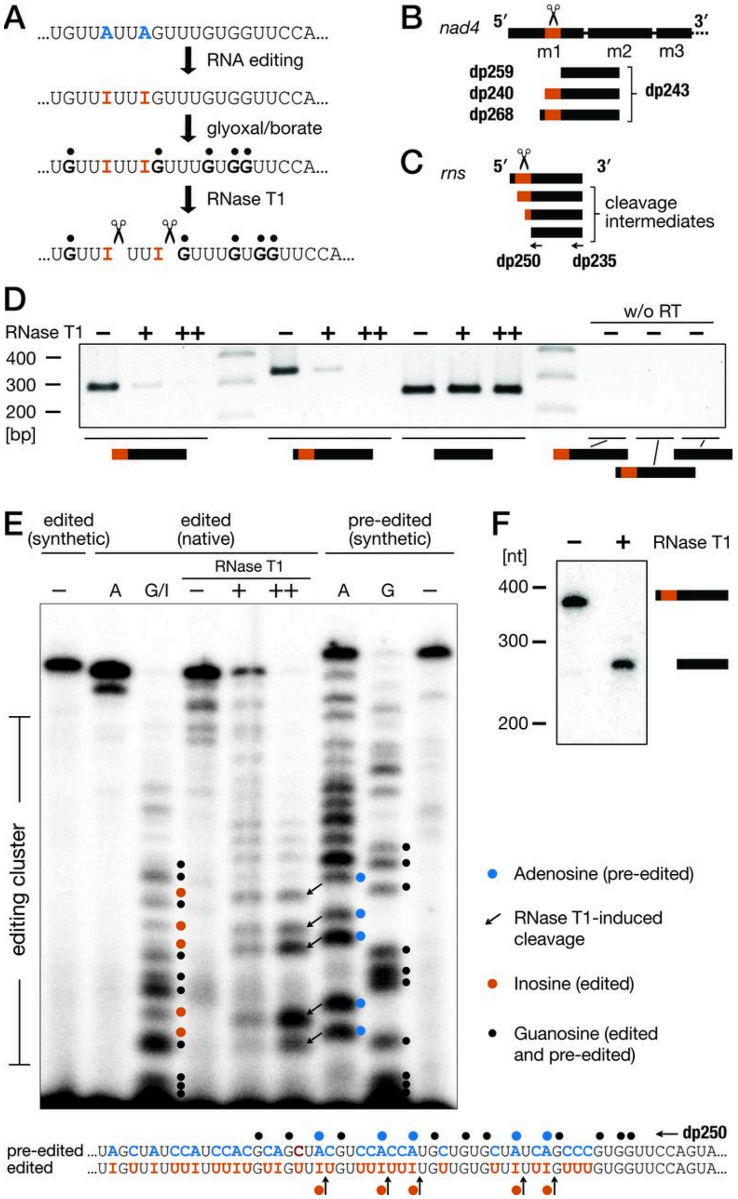


Figure 2. Demonstration of inosines in substitution RNA editing clusters of *Diplonema papillatum*. (**A**) Experimental approach involving glyoxal/borate and RNase T1 treatment of transcripts. (**B**) Design of the RT-PCR assays for the detection of RNase T1 cleavage at unprotected inosines (Is) in the editing cluster of *nad4*-m1. Forward primers bind downstream (dp259), within (dp240), or upstream (dp268) of the editing cluster; the reverse primer (dp243) anneals in the downstream module. (**C**) Oligonucleotides designed for detection of mt-SSU rRNA; dp250 was used for the primer extension assay (see panel **E**), and dp235 served as a probe in northern blot hybridization (see panel **F**). (**D**) RT-PCR products after digesting glyoxalated *nad4*-m1 with increasing RNase T1 concentrations (0, 100, 1000 U). Yield of those RT-PCR products that overlap the edited cluster is reduced progressively, but not of those that do not overlap the cluster. w/o RT, control amplification in the absence of reverse transcriptase (RT). (E) Primer extension of mt-SSU rRNA to map RNase T1 cleavage

sites at Is. As templates served the following glyoxal/borate-treated samples: *in vitro* transcribed substitution-edited mt-SSU rRNA (edited (synthetic)); total RNA (edited (native)); and *in vitro* transcribed pre-edited mt-SSU rRNA (pre-edited (synthetic)). A, sequencing lane where ddTTP was used as a chain terminator. G/I, G: ddCTP was used as dideoxy terminator. –, untreated templates. +, ++: digestion with 10 U and 50 U RNase T1, respectively, amounts which allow detection of cleavage intermediates. In the lanes labeled + and ++, bands represent reverse transcriptase-stops one nucleotide prior to I. The sequence schema at the bottom illustrates the positions of glyoxalation, predicted Is, as well as reverse-transcription stops. (For details on assay optimization and explanation of apparent size shifts, see the Supplementary Figure S4). (**F**) Northern blot hybridization of mt-SSU rRNA, demonstrating the expected size-reduction of the transcript by ~110 nt after digestion with RNase T1 (1000 U) that cleaves off the 5' portion of the rRNA.

For treated *nad4* transcripts, RNase cleavage manifests as a >10-times reduction of RT-PCR amplification across the edited region compared to amplification of the adjacent unedited region (**Figure 2D**). In the tentative mt-*rns*, Is were mapped by differential RNase digestion followed by primer extension (for assay optimization, see **Supplementary Figure S4**). This experiment demonstrates that at least five out of 15 Gs in cDNA are indeed Is in RNA (**Figure 2E**). We also mapped the editing cluster with northern hybridization (**Supplementary Figure S4B**). After extended digestion with RNase T1, the band corresponding to the full-length transcript disappears, showing that the steady-state level of pre-edited tentative *rns* is extremely low (**Figure 2F**).

Given that A-to-I substitutions in *Diplonema* mitochondria occur by deamination, we presume that the same applies to C-to-U RNA editing, since deamination is the only molecular mechanism of C-to-U substitutions encountered in systems that are biochemically characterized [220].

7. RNA editing precedes trans-splicing and progresses stochastically within editing clusters

In kinetoplastid mitochondria, pre-mRNAs are transcribed full-length and subsequently edited progressively from 3′ to 5′ [225]. With deep transcriptome data at hand, we examined whether the same temporal order and directionality applies to RNA editing in *Diplonema* mitochondria. Specifically, we examined read pairs that both span a region encompassing editing sites in a given module, as well as extend beyond the edited module. For both types, U-appendage and substitution RNA editing, we encountered two predominant transcripts forms: edited + trans-spliced and pre-edited + unprocessed, while edited +

unprocessed and pre-edited + trans-spliced intermediates are extremely rare (**Supplementary Table S4**). An independent experiment inquiring *nad4*-m1 intermediates by RT-PCR confirmed this result: a pre-edited trans-spliced module was not detected among poly-adenylated transcripts (results not shown). Thus, in contrast to kinetoplastids, RNA editing in mitochondria of *Diplonema* takes place prior to the occurrence of a full-length transcript, and essentially in parallel with module- end processing. Further, analysis of nad4 transcript intermediates shows that edited and pre-edited substitution sites are interspersed, indicating a stochastic order in the deamination of individual sites (**Figure 3**). Therefore, again unlike RNA editing in kinetoplastids, there appears to be no directionality of editing progression in *Diplonema*.

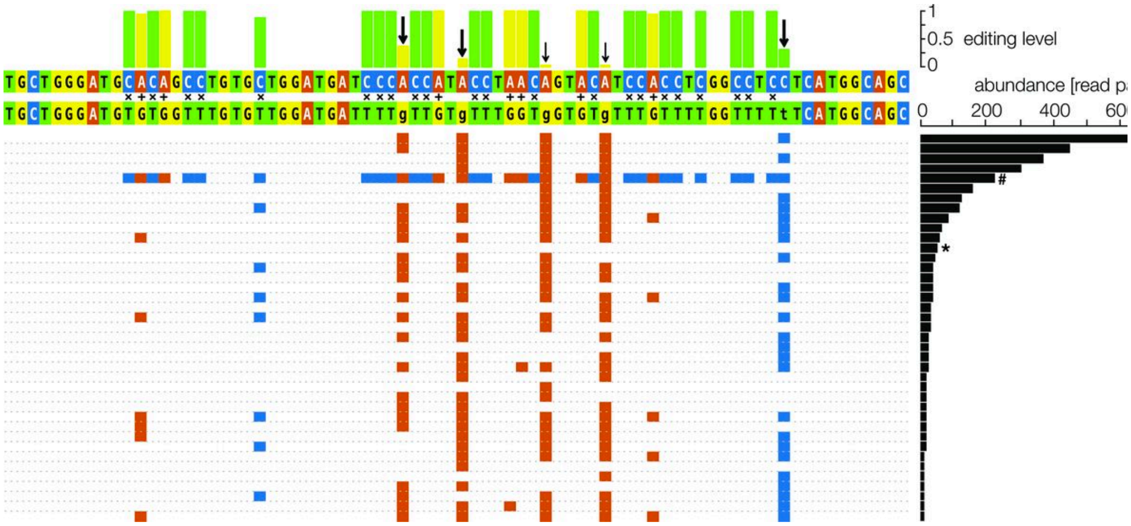


Figure 3. Partial substitution RNA editing in *nad4*-m1. RNA-Seq reads from two poly(A) libraries combined (DPA2 and PA) were analyzed for low- frequency RNA editing in the substitution editing cluster (positions 129-204 in the corresponding cassette). Editing patterns observed in RNA-Seq reads from total-RNA libraries are shown in Supplementary Figure S5. The histogram on the top indicates the editing level within the cluster. Exclusion threshold is <5% frequency, i.e. reported editing sites occur in ≥5% of the reads. Green and yellow bars represent the portion of Is and Us generated by deamination RNA editing, respectively. Arrows point to less-frequently edited sites (frequency 5–38%). Thick arrows indicate the three sites (nucleotide positions 143, 148 and 185 in cassette *nad4*-m1), where both editing states were confirmed to exist in *nad4*-mRNA (Supplementary Figure S6). Top-most nucleotide sequence, genomic sequence. Nucleotide sequence below, cDNA sequence showing the edited state of all sites observed to be edited above threshold. Lower- case letters, sites edited below 50%. +, x: predominant A-to-G and C-to-T substitution sites, edited above 50%. Chart below sequences: editing patterns observed in RNA-Seq reads. RNA and DNA sequence are

identical except for red and blue squares, which indicate pre-edited As and Cs, respectively. The patterns shown are supported by at least 10 reads. The right-hand histogram represents the number of reads per editing pattern. It is the predominant form of *nad4*-mRNA, ranking first, that is shown in all other figures. #, the entirely pre-edited version, ranking fifth (~6% of all reads). *, the maximally edited transcript (i.e. where all, high- and low-frequency sites are in the edited state), ranking 12th (~1% of all reads).

8. Exceptionally high overall RNA editing rate, but several incompletely edited sites

In mitochondria of plants, about 15% of substitution sites (C-to-U) are partially edited (at 90% or less;[215], [250], [251]. Our analyses show that the situation is quite different in *Diplonema*. Sites with partial editing are rare. For example, in 95% of RNA-Seq reads covering the *nad4*-m1 substitution-editing cluster, the totality of sites is either edited or pre-edited (**Figure 3**, **Supplementary Figure S5**). Among partially edited sites, there are three with nearly equal proportion of both editing states. Positions 148 (A-to-I) and 185 (C-to-U) are silent sites, while nucleotide 143 (A-to-I) occupies the first position in a codon and causes a non-synonymous amino acid substitution. An RT- PCR experiment confirms that both A-143 and I-143 exist in polyadenylated *nad4* transcripts (**Supplementary Figure S6**). It is conceivable that both versions of *nad4* mRNAs are translated, because the two alternative codons AUU and IUU specify functionally similar amino acids, Ile and Val, respectively. This finding contrasts with plant organelles, where nearly all partially edited substitution coincide with silent codon positions or fall in pseudogenes (for exceptions see e.g. [252]). It appears that incompletely edited transcripts in plant mitochondria are either not translated or, if translated, the resulting proteins are instable and readily degraded ([253]; reviewed in [254]).

9. Crucial consequences of RNA editing for the tentative mt-SSU rRNA and nad4 protein

We examined the effect of RNA editing on the gene products of two *Diplonema* mitochondrial genes, tentative *rns* and *nad4*. In the tentative mt-SSU rRNA, the substitution editing cluster coincides with the 5′ domain of the 2D structure model. The seven 3′-terminal editing sites in the cluster contribute to helix h18 (**Supplementary Figure S1**) that contains the so-called '530-loop', which is involved in A-site tRNA selection [185]. This region is one of the evolutionary most conserved portions in SSU rRNAs [178]. U-appendage RNA editing of *rns* involves addition of eight non-encoded Us at the transcript's 3′ end, and it is this Utailed RNA that is incorporated in mito-ribosomes. Oligo-(U) tails on mt-rRNAs are also known from kinetoplastids, but the biological role of this ornament remains unclear ([255] and references therein).

In *nad4*, ~80% of post-transcriptionally substituted nucleotides correspond to first and second codon positions resulting in 14 non-synonymous out of 15 codon changes (**Figure 4A and B**). Within the editing cluster, in the stretch corresponding to amino acids 48–64 in the *D. papillatum* protein (referred to as Nad4), RNA editing renders the protein sequence more hydrophobic (**Figure 4C and D**). This outcome of deamination RNA editing has been reported repeatedly before (e.g. [251], [256]) but has not been recognized as an inherent consequence of nucleotide deamination. It is not the particular editing pattern, but rather the mere increase of deaminated bases (i.e. Us and Gs) in codons that leads to amino acids with a higher hydrophobicity index.

The effect of RNA editing on Nad4's secondary structure is even more pronounced. Protein structure prediction indicates that only the 'edited' Nad4 has the potential to form the canonical trans-membrane helix in the N-terminal region (**Supplementary Figure S7A**). Finally, U-appendage RNA editing of *nad4* results in the addition of two Us between the modules 4 and 5 of the trans-spliced transcript. This post-transcriptional event rectifies the reading frame and prevents premature chain termination in translation (Figure 4E and F).

Thus, for both tentative mt-SSU rRNA and Nad4 of *D. papillatum*, RNA-editing appears to be crucial for mitochondrial function and survival of the cell.

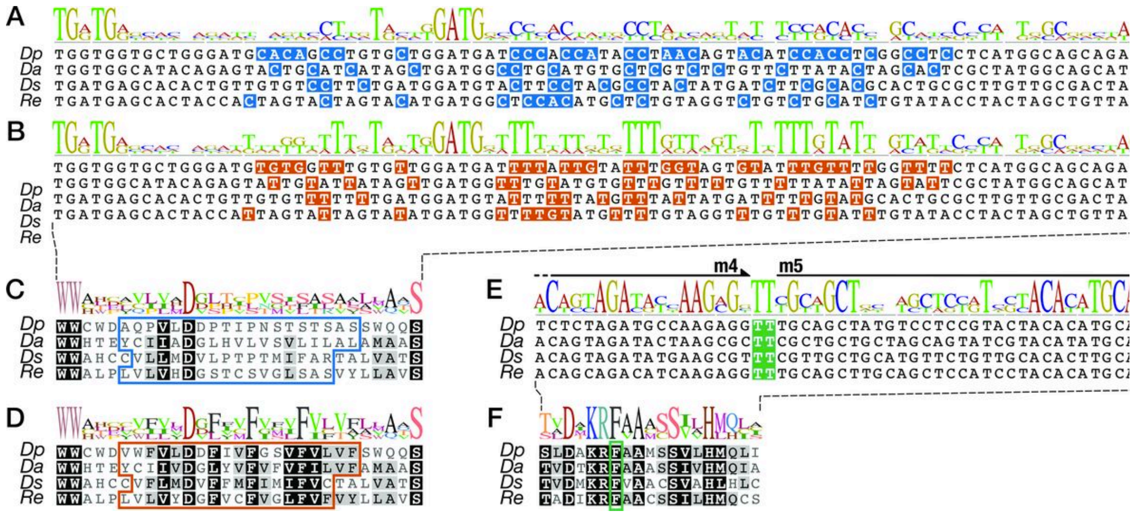


Figure 4. Multiple alignment of *nad4*-m1 genomic (A) and cDNA sequences (B) from diplonemids. Genomic sites that undergo RNA editing and the entire editing cluster are

highlighted in blue and orange, respectively. Sequence logos, with underscored codon triplets, show the impact of RNA editing on sequence identity. (**C** and **D**) Multiple alignment of proteins inferred from genomic (C) and fully edited transcript (D) sequences, with the editing cluster boxed. Background shading of residues indicates the extent of similarity. (**E**) Multiple alignment of the junction of *nad4*-m4 and m5 in cDNA sequences. Module boundaries were annotated based on the available genomic sequences. The Us appended to the module 4 are conserved across all examined diplonemids. (**F**) Multiple alignment of deduced protein sequences of the m4/m5 junction. Abbreviations: *Dp*, *Diplonema papillatum*; *Da*, *Diplonema ambulator*; *Ds*, *D.* sp.; *Re*, *Rhynchopus euleeides*.

10.Comparison of RNA editing in *nad4* across diplonemids

To investigate the conservation of RNA editing across diplonemids, we used *nad4* as a test case. **Figure 4** shows the multiple alignments of pre-edited and edited *nad4* sequences and derived proteins from four diplonemids, *D. papillatum*, *D. ambulator*, *D.* sp. 2 and *R. euleeides* [257]. In all taxa a cluster of substitution editing sites occurs in *nad4*-m1, but there are several variations to the theme. The cluster is located at different positions and it is longer in *D. papillatum* (55 nt) compared to that of the other diplonemids (44–53 nt). Furthermore, while every C in these clusters is edited, the proportion of A-to-I sites ranges from 7/11 in *D. papillatum* to 0/8 in *D. ambulator* and *D.* sp. 2; not a single substitution editing site is conserved throughout these species (**Figure 4A and B**). Together, non-synonymous changes of codons amount to more than 90% in *D. papillatum*, but to ~50% in the other diplonemids. Remarkably, this inter-taxon diversity of substitution RNA editing results in a three times higher sequence identity between the diplonemid transcripts compared to the genes (Figure 4C and D).

In contrast to the inter-species variations in substitution sites, U-appendage RNA editing of the nad4 transcript is strictly conserved throughout the four species, with exactly two Us added post-transcriptionally between modules 4 and 5 (Figure 4E). In sum, RNA editing renders the Nad4 proteins of the examined diplonemids more similar to each other, as well as more similar to orthologs from other eukaryotes (Supplementary Table S5; Supplementary Figure S8).

V.Discussion

1. RNA editing types and sites in *D. papillatum* mitochondria

Comprehensive comparative analysis of the mitochondrial genome and transcriptome from *Diplonema* uncovered two types of post-transcriptional RNA editing: (i) insertions of Us and (ii) substitutions of Cs by Us and As by Is. Together, nucleotide insertions and substitutions in *Diplonema* affect ~80% of all mitochondrial transcripts and account for ~130 RNA editing sites and ~350 nucleotides are generated or altered by editing (Table 1).

Mitochondrial U-insertion RNA editing is extremely rare with only two other, distinct instances outside diplonemids, notably in kinetoplastids [258] (also featuring U deletions) and certain sponges [259]. Substitutions of C-to-U occur more broadly in mitochondria, as well as in plastids. However, the here reported A-to-I RNA editing (of non-tRNA transcripts) is a first in organelles and the interspersed cooccurrence of A-to-I and C-to-U substitutions is unparalleled across all systems.

The distribution of RNA editing sites in *Diplonema* mitochondrial transcripts is conspicuously uneven. U-insertions occur either at module junctions or at the 3′end of transcripts, immediately upstream of the poly (A) tail. This is due to the particular mechanism of U-based RNA editing in this protist and consists in 3′-terminal nucleotide addition prior to trans-splicing or polyadenylation. A-to-I and C-to-U substitutions, on the other hand, are remarkably clustered with up to 45 sites per cluster and up to six sites directly adjacent to one another (Figure 1B). For comparison, in land plant organelles, congregated substitution sites (C-to-U and U-to-C) are rather exceptional [185]. In metazoan nuclear transcripts, clustering is a hallmark feature of substitution editing sites [223], although intervals between sites are much larger compared to *Diplonema* mitochondria. Nowhere else are substitution editing sites as tightly packed as in the system investigated here.

Ribosomal RNAs are rarely ever edited, but in *Diplonema* mitochondria these transcripts undergo massive U-appendage and both C-to-U and A-to-I substitutions. In fact, rRNAs including Is have never been observed before [260]; only one instance of an inosine derivative, O2′-methylinosine is known to occur in cytosolic rRNA of *Crithidia* [261]. Since I has a greater repertoire of potential base pairs than either of the classical nucleotides (it pairs with A, C and U) [262], Is in rRNA probably destabilize the secondary structure of the molecule due to the larger number of alternative pairing possibilities. We speculate that the effect of 15 Is in *Diplonema* mt-SSU rRNA is compensated by proteins in the mito-ribosome.

2. How does the cellular machinery recognize RNA editing sites in *Diplonema* mitochondria?

In certain organisms, targets of RNA editing are recognized by a common sequence or structure element in cis. For example, Apobec-1-dependent C-to-U RNA conversion sites in the mammalian nucleus are characterized by a particular primary sequence context, such as an A + U-rich region along with a downstream 11-nt long motif ('mooring' sequence) [263], [264]. Similarly, ADAR-dependent A-to-I editing substrates in the metazoan nucleus and in viruses share a particular RNA secondary/tertiary structure [223], [265].

In contrast, our analyses of the sequence context around RNA editing sites in *Diplonema* mitochondria did not identify common primary or secondary structure motifs in cis. Therefore, the yet elusive RNA editing machinery is probably directed by an array of distinct site-specific recognition factors acting in trans.

Our search for site-recognition factors resembling guide RNAs in trypanosome mitochondria [266] was inconclusive, suggesting that proteins might assume this task such as the pentatricopeptide repeat (PPR) trans-factors known from plant mitochondria [222]. However, irrespective of the biochemical nature of the postulated recognition factors, there is a dilemma with respect to the crammed substitution RNA editing sites in clusters. How may trans-factors recognize an RNA editing site when its neighboring nucleotides are variable, since they too are subject to RNA editing? In our view, uncovering the nature of the postulated editing guides in *Diplonema* will require an unbiased experimental approach, notably isolation and dissection of mitochondrial complexes having in vitro RNA editing activity.

3. The biological role of organellar RNA editing

Editing of pre-mRNAs in organelles, including mitochondria of diplonemids, is function-critical since the large majority of events generates start codons, abolishes in-frame stop codons or changes codons to specify conserved amino acid positions [211]. Indel editing of mitochondrial mRNAs is particularly essential, since it corrects frameshifts as observed in mitochondria of diplonemids and trypanosomes. Similarly function-critical is editing of tRNAs, not only for proper folding of the molecule, but often for end-processing of their precursor transcripts as well [213].

The situation is different for nuclear metazoan mRNAs where editing is typically partial. Both edited and 'pre'- edited transcripts are translated, and the corresponding proteins play different biological roles. For example, multiple combinatorial codon changes of an mRNA may lead to a large spectrum of protein isoforms that are all encoded by a single, genomic locus. Nuclear RNA editing also acts on intronic regions or UTRs, controlling alternative splicing, efficiency of translation, transcript stability and localization [267]. To summarize, RNA editing compensates disadvantageous mutations in organelles, while it is a means for diversification and regulation in the nucleus.

4. Emergence and diversification of the two RNA editing types in diplonemid mitochondria

The question arises why sites of post-transcriptional U- appendage are strictly conserved, whereas sites of C-to- U and A-to-I substitutions are highly variable. We postulate the following evolutionary scenario. Both RNA editing types probably have been present already in the common ancestor of diplonemids. U-additions are likely compensating detrimental consequences of mtDNA fragmentation that has led to a multi-partite mitochondrial genome. Specifically, post-transcriptional U-appendage may fill in critical nucleotides that were lost from the ancestral genome during double-strand repair at mtDNA breakpoints. Conservation of a fragmented mtDNA throughout the descendants would therefore entail faithful conservation of U- appendage RNA editing. On the other hand, substitution RNA editing likely compensates rapid sequence evolution of diplonemid mtDNAs, explaining why nucleotide substitution pattern are species-specific. In both cases we favor the model of constructive neutral evolution (for more discussion, see Supplementary Data), which posits that prior to genome fragmentation and acceleration of sequence evolution, basic tools to add or change ribonucleotides were already in place, and needed only fine-tuning by evolutionary tinkering [268], [269].

VI. Outlook

We show that diplonemids employ unique post- transcriptional RNA editing in mitochondria involving two distinct molecular processes, U-appendage and base deamination. But what are the enzymes that perform these reactions? As a working hypothesis, we postulate that the machinery that carries out U-appendage editing in diplonemids includes components known from the editosome of trypanosome mitochondria (for a review see [270]). Similarly, the enzymes for deamination RNA editing in diplonemid mitochondria might resemble either ADATs catalyzing A-to-I editing of tRNAs [213], ADARs and Apobecs responsible for A-to-I and C-to-U RNA editing of mRNAs and regulatory RNAs in the metazoan nucleus [223], or PPR-E and PPR-DYW proteins required for C/U-exchange editing in land plant organelles [222].

The latter scenario is quite plausible, because of the finding of plant-like mitochondrial RNA editing in heteroloboseans, a group that shares a common most recent ancestor with Euglenozoa (euglenids + diplonemids + kinetoplastids). Specifically, mitochondrial mRNAs of *Naegleria* and *Acrasia* undergo several C-to-U editing events and the corresponding nuclear genomes encode homologs of the PPR-DYW protein family involved in organelle RNA editing of plants [111], [217].

A glance at the first draft of the nuclear genome sequence from *D. papillatum* identified genes that specify protein domains characteristic for TUTases, PPRs and deaminases. However, it is currently unclear whether the inferred proteins are indeed involved in mitochondrial RNA editing, or rather in basic cellular processes such as RNA turnover, RNA end processing and nucleotide metabolism.

Finally, given the unique features of RNA editing in diplonemid mitochondria, the underlying molecular mechanisms might be entirely novel and may have evolved from unexpected molecular processes. It would not be the first time that the study of protists leads to the first discovery of novel molecular mechanisms that had remained unrecognized in model systems.

Accession numbers

GenBank accession numbers: KU356490-570 (mtDNA cassettes, D. papillatum), KU341361-80 (mitochondrial transcripts, D. papillatum), KU341385-86 (edited + pre- edited nad4 mRNA, D. ambulator), KU341387-88 (edited + pre-edited nad4 mRNA, D. sp. 2), KU341389-90 (edited + pre-edited nad4 mRNA, R. euleeides). See listing in Supplementary Table S6.

Acknowledgements

We acknowledge G. N. Kiethega and S. Teijeiro (both Université de Montréal, Montréal, Canada) for culturing diplonemids and extracting DNA and RNA. We thank M.W. Gray (Dalhousie University, Halifax, Canada) for advice in rRNA structure modeling, Lilyanne Peras for attempting, in the context of an undergraduate summer internship, folding rRNA consensus structures, B. F. Lang and M. Sarrasin (Université de Montréal, Montréal, Canada) for critically reading the manuscript and Peter Stadler (Universität Leipzig, Leipzig, Germany) for discussions on the bioinformatics approaches.

VII. Supplementary data

1. Supplementary File

`Dp_mito_SNP-RNA_20160212.vcf`: vcf file of RNA editing sites in D. papillatum transcripts. RNA-Seq reads of the DPA2 libraries were mapped with Bowtie2 against the unedited (virtual) transcript sequence, and variants were called with FreeBayes (see Supplementary Methods).

2. Supplementary Methods

Strains, culture, and DNA and RNA extraction

Diplonema papillatum (ATCC 50162), *D. ambulator* (ATCC 50223), *Diplonema* sp. 2 *(*ATCC 50224), and *R. euleeides* (ATCC 50226) were obtained from the American Type Culture Collection. Organisms were cultivated axenically as described earlier [168], [227]. Mitochondrial DNA was extracted from an organelle-enriched sub-cellular fraction isolated by differential and sucrose gradient centrifugation. Mitoribosome-enriched fractions were obtained from whole cell lysates by two consecutive kinetic glycerol-gradient ultracentrifugations [227]. RNA was extracted with a home-made Trizol substitute [228]. Residual DNA was removed from RNA preparations by either RNeasy (Qiagen) column purification, or digestion with RNase-free DNase I (Fermentas), or TURBO DNase (Invitrogen) followed by phenol-chloroform extraction. Poly(A) RNA was enriched by a passage through oligo(dT)-cellulose (Ambion), after denaturation of the aqueous solution at 80 °C for 2 min and subsequent chilling on ice.

RT-PCR, *in vitro* transcription, and labeling

Reverse transcription and PCR were performed with the AMV reverse transcriptase (Roche) and Q5 High-Fidelity DNA Polymerase (New England BioLabs), respectively. To produce synthetic pre-edited or edited mt-SSU rRNA, we first made the corresponding DNA templates with the primer pair dp244 and dp245 performing PCR on mtDNA or RT-PCR on purified mt-SSU rRNA. *In vitro* transcription was performed with T7 RNA polymerase (New England BioLabs). Nucleic acids were separated electrophoretically in agarose gels (0.5× TBE; for PCR products) or denaturing polyacrylamide gels (1× TBE with 7 M urea; for RNAs and primer extension products), side by side with the markers GeneRuler 1kb Plus DNA ladder (75-20,000 bp; ThermoScientific) or RiboRuler Low Range RNA ladder (100–

1000 nt; ThermoScientific), respectively. For the purification of PCR products, we used the Wizard SV Gel and PCR Clean-Up system (Promega). *In vitro*-generated transcripts, as well as RNA extracted from the mito-ribosomal fraction (see above) were separated by denaturing PAGE and the desired RNA fragments were purified from the gel by the 'crush-and-soak' method (e.g. [271]). Oligonucleotides used as primers and hybridization probes are listed in **Supplementary Table S1**. Radio-labeling was performed using T4 polynucleotide kinase (New England BioLabs) in the presence of γ^{32}[P]-ATP.

RNase cleavage of glyoxalated RNA, Northern blot hybridization, and primer extension

To detect inosines in RNA, we exploited the protection of guanosine residues against RNase T1 digestion after incubation in a glyoxal/borate solution [230] essentially as devised before [229]. Briefly, 2.5 µg of total RNA were incubated for 45 min at 37°C in phosphate-DMSO buffer (10 mM sodium phosphate pH7.0, 50% DMSO), and deionized glyoxal (2.4%). Alternatively, we used as a substrate 1 µg of poly(A) RNA or 50 ng of purified mt-SSU rRNA or a synthetic RNA, complemented to the final amount of 2.5 µg RNA with total yeast tRNAs (Roche) to avoid over-digestion. After the addition of an equal volume of sodium borate (1 M, pH7.5), the sample was precipitated with five volumes of ethanol at 25,000×g (30 min, 4°C), washed once with 70% ethanol, dried thoroughly, and solubilized in 50 µL of Tris-borate buffer (10 mM Tris-HCl pH7.8, 1 M sodium borate pH7.5). After adding RNase T1 (ThermoScientific) at indicated concentrations, the sample was incubated for 3 min at 37°C, followed by dilution with an equal volume of ice-cold water and RNA extraction using the Trizol substitute (see above). The pelleted RNA was solubilized in phosphate-DMSO buffer, de-glyoxalated by incubation for 3 h at 65°C, and Trizol-extracted. The resulting RNA sample was then used in RT-PCR, Northern blot hybridization, or primer extension assays. Detailed experimental procedures are available at https://www.protocols.io/u/matus-valach.

For Northern blotting, RNA separated by denaturing PAGE was electrophoretically transferred overnight in 1× TBE to a Zeta-Probe nylon membrane (BioRad) and fixed by baking the membrane for 2 h at 80°C. After an overnight hybridization at 42 °C in the ULTRAhyb-Oligo buffer (Ambion) with a radio-labeled oligonucleotide probe, membranes were washed twice at 42°C in 2× SSC, 0.5% SDS. Signals were visualized using a phosphor-imaging screen scanned by a Personal Molecular Imager (Bio-Rad) and analyzed by the Image Lab 5.0 software (Bio-Rad). Primer extensions and dideoxy-nucleotide sequencing reactions with a radio-labeled oligonucleotide and AMV reverse transcriptase (RT) were performed essentially as in [272]. Reaction products were separated on 12% denaturing polyacrylamide gels and visualized as above. Quantitative measurements of RT-PCR product ratios were conducted with the ImageJ software [273].

In the primer extension experiments, we noted that in the G/I sequencing lane of edited *rns*, RT extension pauses somewhat more at Gs than at Is. Inosines are probably more readily passed because they can pair with A, C, and T. We also noted that RT-extension products appear slightly larger than expected when using the pre-edited *in vitro* transcript as template (e.g. compare size of the longest extension products of edited and pre-edited templates in **Figure 2E** of main text). As demonstrated in **Supplementary Figure S4D**, this is due to the particular nucleotide bias of the edited region; the RT-synthetized strand is extremely A+C-rich (87.5%) so that the molecule is less bulky than that synthesized from the pre-edited template (54.5% A+C).

DNA library construction, sequencing, read processing and assembly

We used total DNA for the construction of a genomic library using the kit from Illumina, and sequenced the library with Illumina MiSeq. Details on technology, read counts and length are compiled in **Supplementary Table S2**. Both library construction and sequencing were outsourced to the Genome Innovation Centre, Montreal, Canada. Reads were processed by the custom script cleanNGS.py as follows. Adapters were removed from the 5' and 3' termini of reads with cutadapt version 1.2.1 (http://journal.embnet.org/index.php/embnetjournal/article/view/200). As parameters we used 12-nt long sub-sequences, i.e. the 3' end of the 5'-adapter and the 5' end of the 3'-adapter (to detect partial adapter sequence in the reads) and an error rate of ≤0.1. Cutadapt was also used for quality trimming with a phred quality threshold of 20. Reads shorter than 20 nt were discarded. The dataset used for most analyses contained paired reads only; widow reads that had lost their partner during filtering were discarded using the in-house script fastqPairedAndOrphans.py. Overlapping paired-end reads were combined with FLASH v1.2.7 (http://sourceforge.net/projects/flashpage/). For mtDNA assembly, reads originating from the mitochondrial genome were identified by sequence identity with previously determined mitochondrial chromosomes and modules (GenBank acc. nos. EU123536-8 and HQ28819-33) using BLAST at a hit-reporting threshold of 99%. Mitochondrial reads were clustered with CD-HIT version 4.6 [232] employing the options -c 0.9. Cluster-representative reads were assembled with the Celera software runCA, CA version 8.3rc2 [http://sourceforge.net/projects/wgs-assembler/ [231]] using default options.

RNA-Seq library construction, sequencing and read processing

We depleted *D. papillatum* total RNA and mitochondrial RNA-enriched samples of cytosolic 5S, 5.8S, 18S and 28S rRNA using 5'-end biotinylated oligonucleotides complementary to these rRNA species. Two libraries (PA and DPA2) were prepared from total cellular RNA enriched for poly(A) RNA. The over-abundant mt-LSU rRNA was depleted by hybridization with the biotinylated form of oligonucleotide dp72 (dp72_5biosg) as described previously [227]. Library F1 was made from mitochondrial RNA, using the standard ScriptSeq RNA preparation kit and protocol, which involves RNA fragmentation. The same kit was used for library F2, only that the RNA fragmentation step was omitted, to allow detection of small RNA species. The DpGG library was made from a subcellular fraction enriched in mito-ribosomes (see above). Library preparation and sequencing was outsourced to commercial technology platforms; PA, F1, F2 and DPA2 to Macrogen (Korea) and DpGG to the Genome Quebec Innovation Center (Montreal). Libraries from the three other diplonemids were made from total RNA after depletion from cytosolic rRNAs, outsourced to the Montreal facility. Details on technology, read counts and length, see **Supplementary Table S2**. We removed reads corresponding to the bacteriophage phiX174 (internal control for library construction and sequencing) and cytosolic rRNAs by mapping reads against the corresponding sequences (GenBank acc. nos. CP004084, KF633466.1, KF633467.1, AY007785) using Bowtie 2 [130] (see below) with the options --local --no-unal, leaving between 33% (F1) and 95% (PA) of the reads. RNA-Seq reads were adapter-clipped and quality-trimmed as described for genomic reads. For the search of trans-acting RNA factors (see below), reads from F2 were processed in a second way: they were only adapter-clipped without quality trimming nor removal of reads based on length.

Mitochondrial reference genome sequence

From a Celera assembly of mitochondrial-genomic MiSeq Illumina reads (see above), we extracted contigs holding cassettes, i.e. those containing the distinctive left-hand and right-hand class-specific constant regions of chromosomes, but not the shared constant region (see **Supplementary Figure S1**). Contigs extracted this way represent all previously identified as well as new putative coding regions. As a final step, the contig (consensus) sequences were validated and polymorphisms were determined simultaneously by mapping back the MiSeq reads to the contigs with Bowtie 2 [130]. Genomic variants were called using GATK-UnifiedGenotyper [127] and FreeBayes [126], both suited for non-diploid genomes. Read mapping and variant calling are described in more detail in a following section.

Transcript *de novo* assembly and function annotation

RNA-Seq data from the PA and DPA2 libraries were assembled using SOAPdenovo-Trans [236] using kmers of 35, 55, 75, 95, 107, and 127. The resulting contig files were assembled again using SOAPdenovo [237] using a kmer size 127. We also used Trinity with default parameters (http://trinityrnaseq.github.io/ [274]). Mitochondrial rRNA sequences were assembled from GG reads (mito-ribosome library). Identification of mitochondrial transcripts was based on sequence identity with mitochondrial gene modules. Function assignment of mitochondrial transcripts was attempted by BLAST against the non-redundant nucleotide collection (nr), FASTA [275] against in-house taxonomically broad mitochondrial protein data collections, and HMMER (Eddy, S. 2008 http://hmmer.janelia.org). The search models for the latter were built from multiple alignments of proteins from excavates and eukaryotes with moderately derived sequences (see Methods section 'Determination of the degree of Nad4 conservation'). Identification of mt-rRNAs was attempted with Infernal 1.1 [147], Multilign, TurboFold [240] (http://rna.urmc.rochester.edu/), and RNAshapes [276] (https://bibiserv2.cebitec.uni-bielefeld.de/rnashapes). Again, the search model was built from multiple alignments of sequences from excavates and eukaryotes with moderately derived sequences.

Mitochondrial transcriptome reconstruction and assignment of orphan modules

In addition to *de novo* assembly described above, transcripts were also reconstructed via a split-read approach developed in-house. As a first step, RNA-Seq reads of the poly(A) library were mapped to the genome reference with Bowtie 2 in paired-end and local mode as specified above. Read pairs were identified whose two partners do not map to the same genomic module. Of these reads, we analysed the sequences that were soft-clipped during mapping in local mode. Notably, the custom python script findTransSplicedRNA.py inquired if the soft-clipped sequence overlaps with another genomic module. If so, then the two modules belong to the same gene and are neighbor modules in the trans-spliced transcript. In that way, transcripts were extended until no further neighbor module was retrieved. As an additional approach, gene module sequences were 'blasted' against the *de novo*-assembled mito-transcriptome (see above). The two procedures allowed to assign a dozen modules from unidentified genes (previously designated X1-m(k) etc. [174]) to transcripts, and previously unknown neighbor-module relationships were resolved. The newly discovered genes (of yet unknown identity) are designated *y1*, *y2*, etc. and their corresponding modules accordingly (**Table 1,** main text).

Mapping of Illumina reads to reference sequences

Illumina reads were mapped to the corresponding reference sequence with Bowtie 2 [130]. For genomic reads, we used the mito-genome reference, and for RNA-Seq reads the mito-transcriptome reference. We employed the options --local --no-unal (removing unaligned reads), and default values for the alignment and scoring parameters. Output files in sam format were subsequently transformed into '.bam' files with SAMtools v1.4 (http://samtools.sourceforge.net/). Alignments were visualised with the Integrative Genomes Viewer (IGV; https://www.broadinstitute.org/igv/) [233]. RNA-Seq Illumina reads were mapped in the same way to the mito-transcriptome reference.

Calculation of FPKM

To assess the read coverage of transcripts, we determined FPKM (Fragments Per Kilobase of transcript per Million of mapped reads) values. For that RNA-Seq reads of the library PA2 were mapped against the mito-transcriptome reference using TopHat v2.0.14 (https://ccb.jhu.edu/software/tophat/index.shtml) [277] or STAR version 2.4.2a (https://github.com/alexdobin/STAR) [124] with default parameters, followed by assembling mapped reads into contigs with Cufflinks (https://github.com/cole-trapnell-lab/cufflinks). FPKM values are reported in the file genes.fpkm_tracking.

In silico identification of polymorphic genomic sites

Bam files of reads aligning with the reference were processed with tools of the Picard suite (http://picard.sourceforge.net/), including marking of duplicates, addition of read group information, indexing, and creation of reference dictionary files, followed by generation of a reference index with samtools faidx. Variant sites in mtDNA were determined with the UnifiedGenotyper module of the Genome Analysis Toolkit (GATK) v3.3.1 (https://www.broadinstitute.org/gatk/ [127], [238]) and FreeBayes [126]. To our knowledge, these are the only tools suited for non-diploid genomes. For both tools we set the ploidy to 100 –the estimated copy numbers of mitochondrial chromosomes–, the minimum number of observed variants to 2, the minimum base quality and mapping quality to 30, and the minimum allele frequency to 0.01. In addition, reads marked as duplicates were used, because the true duplication rate as determined by prinseq-lite.pl v0.20.3 (http://sourceforge.net/projects/prinseq) was <10%; the tool 'MarkDuplicates.jar' of the Picard suite falsely labeled mitochondrial reads as duplicates due to the high sequencing depth. In summary, the options used for UnifiedGenotyper are -ploidy 100 -glm BOTH -stand_emit_conf 30 -stand_call_conf 30 -use-duplicates, and for FreeBayes --ploidy 100 --min-mapping-quality 30 --min-alternate-count 2 --min-alternate-fraction 0.01 --use-duplicate-reads. The output file of DNA-DNA comparison is referred to as DDd.vcf files. Only sites with at least 10% allele frequency were considered. We validated the obtained genome variant calls (based on Illumina MiSeq reads) in two ways. First, variants were inspected visually with IGV v2.3.40 (https://www.broadinstitute.org/igv/) [233]. For linked sites, we consolidated the allele frequencies reported by the variant caller software, by calculating the mean across all linked sites. We noted that variants located up to ~35-nt away from the ends of the reference sequences cannot be detected by variant callers, because these sites are included into the soft-clipped regions during read mapping. In addition, variants close to contig boundaries were missed, simply because mapping quality values of reads that extend beyond contigs are low. These variants were only reported when running FreeBayes and UnifiedGenotyper on alignment files whose mapping quality was corrected by setting it to 35. The obtained variants were also validated with reads generated by different technologies, notably Sanger and 454-FLX (Roche) obtained from random mtDNA libraries or PCR-amplified mtDNA regions. Among 35 tested variant sites predicted from MiSeq reads, 33 are supported by Sanger and Roche 454 reads. For two sites, the alternative allele (10% frequency) lacks support from the two latter technologies. Yet, coverage of Sanger and 454 reads was low. With only 9 and 16 reads, respectively, the probability of not observing the alleles is due to chance (p>0.1).

In silico identification of RNA editing sites

When mapping RNA-Seq reads against a genomic reference, Bowtie 2 and other mappers miss many reads that cover massively edited transcript regions. Conversely, when mapping RNA-Seq reads against a fully edited reference sequence, reads were missed that correspond to partially edited transcripts. Therefore, we chose to map reads against both the pre-edited and edited transcriptome references and to merge the resulting bam files. The pre-edited transcriptome reference is a virtual sequence; it was generated by concatenating genomic module sequences of each gene in the correct order and including 5′ and 3′ UTRs corresponding to those present in the transcripts. RNA-Seq reads were mapped to the reference sequence with Bowtie 2 version 2.2.3 [130] (in strand-specific mode with --norc). Throughout the alignments, read coverage (>500; except for the junctions of *y2*-m4, *y5*-m2/m3, and the unassigned module X12) and mapping quality (≥30) was sufficiently high, which significantly reduces false negatives (i.e. true sites not reported). The returned bam files were merged by the custom script mergeSAM.py, which is available in GitHub at https://github.com/SandrineMoreira/publishedscripts. To identify DNA-RNA differences (DRds), the merged bam file was used as input for the GATK UnifiedGenotyper [127] and FreeBayes [126], using the same parameters as for calling genomic variants described above. Note that both software tools are designed for detecting variants in genome sequence, and consequently, the ploidy variable has no tangible meaning for 'variants' in RNA. Our tests of a range of ploidy values (20, 50, 100, and 200) yielded the same result; the returned allele frequencies only differed in the number of digits after the decimal point. Results of DRd calling are written to DRd.vcf files. The FreeBayes-generated vcf file (Dp_mito_SNP-RNA_20160212.vcf) is available in Supplementary Data. To differentiate between genomic polymorphisms and RNA editing sites, the DRd.vcf and DDd.vcf files were compared with the tool vcf-isec of the VCFtool kit (https://vcftools.github.io/perl_module.html#vcf-isec), which intersects the two files and returns sites that are non-allelic in the genome sequence but differ in the transcriptome sequence. The results were inspected and validated by visualising RNA-Seq read-reference alignments with IGV v2.3.40 [233].

Analysis of RNA processing intermediates

Using the in-house script editpop.py, we analysed whether there is a correlation between RNA editing of the various sites on a given transcript, RNA editing and module-end processing, as well as RNA editing and trans-splicing. As a first step, reads were identified where one of the two partners maps to any of the following RNA processing intermediates: cassettes holding pre-edited modules, virtual pre-edited fully trans-spliced transcripts, and edited full-length transcripts. Then, reads overlapping the edited regions were extracted and examined for the presence of flanking regions or overlap with a neighbor module. For substitution editing, at least 50% of the region overlapped by a read must be edited. For editing by appendage of >2 Us, the read must overlap at least 2 Us. Then, reads having the same processing status (e.g. edited + trans-spliced) are grouped in the same class and the number of pairs for each class is counted. Sites of short U-appendages (1-2 nt) were analyzed by searching motifs in individual reads, using the Unix tool grep, requesting a full match of at least six terminal nucleotides in the module and six in the adjacent region. The search of terminal Us at a module's 5′ or 3′ ends was also performed with grep, inquiring forward reads for patterns of ≥2 Us preceded by 8 adjacent nucleotides from the corresponding module. Reverse reads were inquired for the reverse-complementary pattern at their 3′ end.

Analysis of partially edited *nad4* transcripts

The in-house script editedSitesStat.py (available from the GitHub repository at https://github.com/SandrineMoreira/publishedscripts), examined whether certain RNA editing sites remain pre-edited more often than others. For that, RNA-Seq reads from the DPA2 library (made from poly(A) RNA) were mapped with Bowtie 2 (see above) to the pre-edited (virtual) transcript sequences. The script parses the transcriptome-genome alignment file (in sam format) with a focus on reads aligning with *nad4*-module 1, while the corresponding vcf file (**Supplementary File 1**) was used to extract the positions to which the editing sites correspond in the transcript. In a first step, the different classes of editing patterns encountered in the corresponding RNA-Seq reads were established. In a second step, the number of reads falling in each class were counted, and classes sorted by the number of members. As a third step, we calculated the percentage of pre-edited versus post-edited state for each site.

Search for *cis* elements and RNA *trans*-factors that guide RNA editing

We searched common sequence and secondary structure (2D) motifs near RNA editing sites and clusters in the mitochondrial transcriptome. Sequence *cis*-motifs that potentially define individual substitution-editing sites were searched with the in-house script editbysite.py that looks for recurrent 2 to 17-nt long motifs up to 50 nt away from each site in the pre-edited and edited transcriptome sequence. For motifs potentially flanking substitution-editing clusters, we extracted 50 nt upstream and downstream of the cluster and submitted these sequences to the MEME web site [239] (http://meme-suite.org/tools/meme). We used the MEME algorithm for finding ungapped motifs, as well as the GLAM2 algorithm, which, in addition, recognizes gapped motifs. Common 2D motifs were searched in the same regions using RNAalifold [179], [240], [241], which identifies a common structure given the sequence alignment, and LocARNa [242], [243], RNAstructure and Multilign [240], [241], which build alignment and secondary structure simultaneously. Default parameters were used. We searched *trans*-acting guide RNAs in reads from the strand-specific library F2, which made from an RNA preparation that included small transcripts (**Supplementary Table S2**). The search was performed with GNU grep supplied with query motifs that are reverse-complements of the edited sequence. To allow anchoring of the hypothetical guide RNA, the query motif includes sequence adjacent to the editing site. For U-appendage guides, the motif includes 12 nt upstream (with respect to the transcript) of the editing site. Tests with a shorter anchor yielded numerous fortuitous hits. For substitution clusters, we added six adjacent nucleotides on both sides. To detect potential guide RNAs that do not correspond exactly to the reverse-complement of the edited sequence, we mapped reads of the F2 library against the sequences of full-length mature and pre-edited (virtual) transcripts. Bowtie2-mapping was performed using the options --local --nofw and default mismatch settings. Alignments (bam file) were inspected visually as described above.

Determination of the degree of Nad4 conservation

We determined whether the protein sequence deduced from edited *nad4* mRNA is phylogenetically more conserved than that inferred from the pre-edited mRNA. First, we built a multiple alignment of Nad4 protein sequences deduced from edited and pre-edited *nad4* genes from four diplonemids and 15 moderately divergent homologs. The chosen non-diplonemid taxa include four Excavata (*Jakoba libera, Malawimonas californiana, Reclinomonas americana, Naegleria gruberi*), three Archaeplastida (*Cyanophora paradoxa, Nephroselmis olivacea, Porphyra purpurea*), two stramenopiles (*Phytophthora infestans, Pylaiella littoralis*) two unikonts (*Monosiga breviata, Yarrowia lipolytica*), and *Rhodomonas salina* and *Bigellowiella natans*. All sequences were downloaded from NCBI's protein database (http://www.ncbi.nlm.nih.gov/protein/). The multiple protein alignment was constructed with MUSCLE using default parameters [234]. From this alignment, we extracted, via the alignment editor Genetic Data Environment (GDE; [278]), a sub-alignment including columns 1 to 130, which corresponds to the N-terminal region up to the last 'edited' amino acid in diplonemids. Based on this alignment, protein conservation was determined with MstatsX (https://github.com/gcollet/MststX) which uses the trident statistics [235] (**Supplementary Table S5**).

Protein secondary structure analysis of Nad4

To determine potential trans-membrane helices, the protein sequences of 'pre-edited' and 'edited' diplonemid Nad4 variants were scanned with TMHMM2.0 employing default parameters [244] (web service at http://www.cbs.dtu.dk/services/TMHMM/). We also used Phobius [245] at http://phobius.sbc.su.se/, and the TMpred tool from the ExPASy web suite [246] at http://www.ch.embnet.org/software/TMPRED_form.html; the results were essentially the same.

3. Supplementary Results and Discussion

Search for recurrent sequence and secondary structure elements defining RNA editing sites

We searched for recurrent sequence motifs in a window ±50 nt around RNA editing sites employing the discovery tools available at the MEME server (http://meme-suite.org/tools/meme), notably the MEME and GLAM2 algorithms for ungapped and gapped motifs, respectively [239]. The results are shown in **Supplementary Figures S3A-C**. We found several motifs that are shared by U-appendage sites, however, the same motifs are also present in unedited modules and are therefore considered unspecific.

The search of sequence motifs around substitution editing clusters also failed to return common patterns. Similarly, individual C-to-U and A-to-I sites seem not to be associated with a shared motif in the pre-edited sequence. Still, we observed that in the edited sequence, Is (but not editing-generated Us) occur in a distinctive -1/+1 context: G**I**U, U**I**G, or U**I**U (in cDNA: G**G**T, T**G**G, T**G**T). The implication of this finding is not clear. A major conceptual difficulty arises from the clustering of substitution sites in *Diplonema*. Since substitution RNA editing seems not to proceed directionally as does indel RNA editing in kinetoplastid mitochondria [279], a given pre-edited site can theoretically have numerous different sequence contexts through the various combinations of edited and pre-edited neighboring sites.

Search for potential *trans*-factors that guide RNA editing

In previous studies we searched for RNA species in *Diplonema* that may guide RNA editing. Most importantly, the experimental approaches did not detect gRNA-like transcripts known from kinetoplastids, i.e. molecules that are abundant and small (~50-100 nt), or that carry a 3′ oligo-U-tail or a 5′-tri-phosphate [174].

Further, we reported RT-PCR experiments and *in silico* searches that aimed at detecting guide RNAs that simultaneously instruct U-appendage editing and trans-splicing at the junctions *cox1*-m4/m5 and *rnl*-m1/m2. For the junctions *cox1*-m4/m5 where 6 Us are added, one candidate guide was identified [174]. However, due to the experimental design, only the sequence immediately around the junction/editing site could be resolved. Additional experiments to amplify and sequence the entire molecule or to identify it by Northern hybridization were unsuccessful. Computational searches for this candidate sequence found 10 distinct motifs in mtDNA and many more in nuclear EST and genome sequences [174].

Another study investigated the *rnl*-m1/m2 junction where ~26 Us are added [227]. RT-PCR experiments with various primer pairs yielded low amounts of antisense products, whereof one was confirmed by amplicon sequencing. In addition, we analysed *in silico* a strand-specific RNA-Seq library for potential editing and trans-splicing guides. The proportion of putative antisense reads was marginally yet significantly above background (2.5%), but extremely few reads covered the entire U tract, questioning the significance of the finding.

In the current study, we searched by *in silico* methods for potential guide RNAs for all identified editing sites. For that, we analysed reads of a strand-specific RNA-Seq library from *D. papillatum*. The library was prepared from RNA enriched in mitochondrial transcripts and without the commonly-applied step of insert fragmentation. Insert sizes are 10 nt and above (library F2, **Supplementary Table S2**). In these reads, we searched for sequence motifs that are reverse-complementary to each of the detected RNA editing sites. The anchor, by which the potential guide RNA binds to the pre-edited transcript, was set to 12 nt upstream of U-addition sites and 6 nt each upstream and downstream of substitution clusters. The hits (**Supplementary Figure S3D**) are considered spurious, since the number is small and decreases with the length of the searched motif. As a control, and to detect potential guide RNAs that do not correspond exactly to the reverse-complement of the edited sequence, we mapped reads of the F2 library against the sequences of full-length mature and virtual pre-edited transcripts, only allowing antisense reads to align (mode --very-sensitive-local --nofw; see Methods in main text). Visual inspection of read alignments does not reveal reads that fully cover U-appendage or substitution RNA editing sites in antisense orientation.

These results can be interpreted in different ways. If editing-guide RNAs do exist, then either their steady-state concentrations might be too low to be detected in RNA-Seq data, or their 5′ and/or 3′ nucleotides prevents adapter ligation (e.g. cyclic bases, 5′ OH or 3′ phosphate). Further, the RNA guides may have an unexpected sequence. For example, the reverse complement binding motif might not be contiguous in sequence but interrupted by intervening stretches that bulge out in the molecule's 2D structure. Finally, RNA editing could be guided by proteins rather than relying on RNA.

In sum, we have no convincing indication for RNA molecules guiding RNA editing in *Diplonema* mitochondria.

The role and 'reason' for RNA editing

RNA editing plays radically different roles in different systems. In the metazoan nucleus, it serves protein diversification and gene regulation, whereas in organelles, it compensates deleterious mutations [211]. In the latter case, one might wonder why RNA editing has persisted over long evolutionary periods, and why it has not been rendered obsolete by correcting 'errors' at the DNA level. A number of publication (ref. [280] and references therein) discuss this issue stressing that natural selection of the fittest is not the only principle in evolution, but that non-adaptive processes (genetic drift and mutation pressure [281]) are at work as well, giving rise to 'seemingly gratuitous complexity' in many forms.

Supplementary Table S1. Oligonucleotides used in this study

Primer	Sequence (5′ → 3′)	Target	Application
CDS-III	ATTCTAGAGGCCGAGGCGGCCGACATG$(T)_{30}$VN	poly-(A) RNAs	RT
dp24	GCATGGCATCCTCATGCTCTT	*cox1*-m2	PCR
dp28	CTCTGGGTGCCCGAAGACC	*cox1*-m4	PCR
dp235	GCTCTCCAGCTAGTGCTTAGCGGTACATGC	*rns*	hybridization
dp240	ATGTGTGGTTTGTGTTGGATGATTT	*nad4*-m1	RT-PCR
dp242	AGCATGGCGTGGAGTAGCAA	*nad4*-m1	sequencing
dp243	CATCATACTGCTACCTACGACGGT	*nad4*-m2	RT-PCR
dp244	AAAAAAAACUAGCUGGAGCUCUCCAG	*rns*	PCR
dp245	TAATACGACTCACTATAGGGTATGCTGTGGTACATGT	*rns*	PCR
dp250	GGCGTAACCGTACGGATACTGG	*rns*	primer extension
dp253	TAGCAGTATACGGTGTACCCAT	*nad4*-m5	PCR
dp256	CACTGTTCATGAGGAGTACCATGAG	*nad4*-m6	PCR
dp259	TCATGGCAGCAGAGCACCAC	*nad4*-m1	RT-PCR
dp268	GGGTATGTGATGTCCTGGTGGTG	*nad4*-m1	RT-PCR
dp72_5biosg	5′-biotin/CCATTAGCTCTACCGTACCTA	*rnl*-m2	pull-down

Supplementary Table S2. Libraries used in this study [a]

Library name	Material	Library preparation kit [b]	Sequencing technology [c]	Nr. of raw read pairs	Read length (nt)
Dp-nucDNA	*D. papillatum* total DNA	TruSeq DNA	MISEQ PE	25,880,204	250
PA	*D. papillatum* poly(A) RNA	ScriptSeq RNA	HISEQ PE SS	61,236,282	101
DPA2	*D. papillatum* poly(A) RNA	TruSeq RNA	MISEQ PE SS	261,280,954	250
F1	*D. papillatum* mt-RNA	ScriptSeq RNA	HISEQ PE SS	71,647,714	101
F2 [d]	*D. papillatum* mt-RNA	ScriptSeq RNA	MISEQ PE SS	53,564,184	101
DpGG	*D. papillatum* mt-rRNA	TruSeq RNA	MISEQ PE SS	3,039,880	250
DaT	*D. ambulator* total RNA	TruSeq RNA	MISEQ PE SS	3,038,621	250
DsT	*D.* sp. 2 total RNA	TruSeq RNA	MISEQ PE SS	3,150,711	250
ReT	*R. euleeides* total RNA	TruSeq RNA	MISEQ PE SS	3,385,789	250

[a] See Methods section of main text.

[b] TruSeq is a product from Illumina, ScriptSeq from Epicentre.

[c] PE, paired-end; SS, strand-specific.

[d] The standard ScriptSeq protocol was modified by not fragmenting RNA prior to the reverse transcriptase reaction. Minimum insert size is 10 nt.

Supplementary Table S3. Genomic sequence variants in cassettes of *D. papillatum* mtDNA

Cassette	Nt position in cassette	Reference nt(s)[a]	Variant nt(s)[a]	Variant in module (consequence)[b]	Variant frequency[c]	Variant type[d]
atp6-m3_casA-	16	T	G	/	0.25*	S
	19	G	C	/	0.25*	S
	20	C	CT	/	0.25*	I
	203	G	A	/	0.25*	S
cob-m2_casA-	8	TGC	TC	/	0.40*	I
	237	AG	ACG	/	0.40*	I
	255	AG	ACG	/	0.40*	I
cox1-m1_casB+	267	C	T	/	0.25	S
cox1-m2_casA+	131	CAT	CAAT	(frameshift[e])	0.50	I
	208	A	G	/	0.40	S
cox1-m7_casA+	20	T	G	/	0.20*	S
	68	CTAGCT	ATGT	/	0.20*	I,S
cox1-m9_casA-	85	G	A	(GAG→GAA; Glu)	0.20*	S
	263	T	C	/	0.20*	S
	265	C	A	/	0.20*	S
cox3-m3_casA-	154	G	A	(GTG→GTA; Val)	0.30*	S
	289	T	C	/	0.30*	S
nad1-m2_casA-	64	A	G	(GCA→GCG; Ala)	0.25*	S
	142	A	G	(TCA→TCG; Ser)	0.25*	S
	217	GG	CC	/	0.25*	SS
	224	C	T	/	0.25*	S
	244	G	A	/	0.25*	S
	249	ATCC	CTCT	/	0.25*	SS
nad1-m3_casA+	26	C	T	/	0.25*	S
	48	G	C	/	0.25*	S
nad1-m4_casA-	31	C	A	(TCC→TCA; Ser)	0.40*	S
	208	A	G	(GTA→GTG; Val)	0.40*	S
	263	CCACG	TCACC	/	0.40*	SS
	283	G	C	/	0.40*	S
nad1-m5_casA-	105	C	T	(CTC→CTT; Leu)	0.20*	S
	238	A	G	/	0.20*	S
nad4-m1_casA-	14	G	C	(GGA→GCA; Gly→Ala)	0.20*	S
	19	C	A	(CCG→ACG; Pro→Thr)	0.20*	S
	54	A	G	(GCA→GCG; Ala)	0.20*	S
	276	A	G	(CTA→CTG; Leu)	0.20*	S
nad4-m3_casA+	204	G	A	(GTG→GTA; Val)	0.45*	S

nad4-m8_casA-	17	A	G	(ATG→GTG; Met→Val)	0.20*	S
	173	A	G	(ACA→GCA; Thr→Ala)	0.20*	S
nad5-m1_casB+	16	C	G	/	0.40	S
nad5-m2_casA-	122	A	G	(TGA→TGG; Trp)	0.45*	S
	221	G	A	(GTG→GTA; Val)	0.45*	S
nad5-m3_casA-	7	G	A	/	0.40*	S
	15	C	T	/	0.40*	S
	78	C	T	(GCC→GCT; Ala)	0.40*	S
	112	G	A	/	0.40*	S
	139	GG	GAG	/	0.40*	I
	149	GACACA	GACA	/	0.40*	I
nad5-m7_casA-	285	C	G	/	0.35	S
nad5-m8_casA+	30	C	G	/	0.40	S
	156	G	A	(TGG→TGA; Trp)	0.40*	S
	175	T	G	(TCC→GCC; Ser→Ala)	0.40*	S
nad5-m10_casA-	94	G	A	(TGG→TGA; Trp)	0.35*	S
	109	T	C	(GCT→GCC; Ala)	0.35*	S
	115	C	G	(ACC→ACG; Thr)	0.35*	S
	216	G	C	/	0.35*	S
nad5-m11_casA+	7	C	T	/	0.25*	S
	233	C	T	/	0.25*	S
	235	G	C	/	0.25*	S
	249	T	C	/	0.25*	S
nad7-m2_casA-	37	G	C	/	0.15	S
	75	G	T	/	0.15	S
	248	CT	CAT	/	0.15	I
nad7-m3_casA+	57	GCTCTCTCTCTG	GCTCTCTCTG	/	0.35*	I
	129	C	T	/	0.35*	I
nad7-m4_casB-	140	GGCC	AGCA	/	0.15	SS
nad7-m5_casA+	168	T	C	(TCA→CCA; Ser→Pro)	0.10	S
nad7-m6_casA-	141	G	A	(ACG→ACA; Thr)	0.30*	S
	250	CA	CACCTA	/	0.30*	I
	266	C	G	/	0.30*	S
nad8-m1_casA+	16	A	GGT	/	0.20*	SI
	21	C	G	/	0.20*	S
	24	G	C	/	0.20*	S
	63	G	A	/	0.20*	S
	230	ACC	ACCC	/	0.20*	I
	264	AC	GG	/	0.20*	SS

rnl-m2_casB-	54	G	A	+	0.20	S
rns-m1_casB-	73	T	C	+	0.20	S
y1-m1_casA+	37	T	G	/	0.25*	S
y1-m2_casA+	44	GG	CC	/	0.25*/**	SS
	57	AGCAA	AA	/	0.15*	I
	111	GAGC	GC	/	0.1**	I
	118	GGT	GT	/	0.15*	I
y2-m4_casA-	26	GG	GTG	/	0.35*	I
	31	A	G	/	0.35*	S
	198	ACC	AC	(frameshift[e])	0.35*	I
	216	T	G	/	0.35*	S
	223	A	G	/	0.35*	S
y3-m1_casA-	26	CGTGGTG	TGGATGCA	/	0.25*	I
	234	AGGAC	AGGGAC	(GAG→GGG; Glu→Gly)[f]	0.25*	I (S)
y3-m2_casB+	131	T	C	/	0.1*	S
	147	G	T	/	0.1*	S
y4-m1_casA-	26	TGGATGCA	CGTGGTG	/	0.40	I
y5-m1_casA-	279	A	G	/	0.20	S
y6-m1_casB-	36	TGGTAT	AGGT	/	0.20	SI

[a] The reference sequence corresponds to the most abundant alleles.

[b] /, variant falls in the flanking region of the cassette. Parentheses indicate that the variant falls in a module of a protein-coding gene, and codon and amino acid changes are shown by arrows. +, variant falls in the module of a structural RNA gene.

[c] Variant frequencies of ≥0.1 are listed. Codon changes are only indicated for modules whose mature transcripts are known to be translated into proteins. Asterisks indicate linked variants.

[d] S, substitution, I, indel.

[e] The peptides detected in mass spectrometry correspond to the non-frame shifted protein.

[f] The G-insertion occurs at the module's 3′end, occupying the second codon position; the codon is completed by the first nucleotide in the downstream module. Thus, the insertion appears as a substitution in the trans-spliced transcript.

Supplementary Table S4A. Intermediates of substitution RNA editing, end-processing, and trans-splicing [a]

Library	Modules	Intermediate type [b] count (%)			
		-■ / ■-	-🟥 / 🟥-	□■ / ■□	□🟥 / 🟥□
PAs	*nad4*-m1	101 (6.8%)	4 (0.3%)	11 (0.7%)	1,359 (92%)
	y1-m2 [c]	779 (2.5%)	2,548 (8%)	294 (1%)	27,590 (88.5%)
	y3-m4	0	1 (0.3%)	2 (0.5%)	384 (99.2%)
F+T	*nad4*-m1	453 (90.6%)	1 (0.2%)	2 (0.4%)	44 (8.8%)
	y1-m2 [c]	454 (60%)	136 (18%)	0	167 (22%)
	y3-m4	0	0	0	40 (100%)

[a] Count of RNA-Seq read pairs supporting the various RNA processing intermediate types. Modules whose intermediate counts are below 5 are not shown. Pre-edited modules having a poly(A) or poly(U) tail have not been observed. The FPKM values of the transcripts *nad4*, *y1* and *y3* vary by a factor of 10.

[b] ■, pre-edited module; 🟥, edited module; -■ / ■-, module prior to end-processing; ■□ / □■, module trans-spliced with downstream or upstream neighbor. PAs, poly(A)-RNA libraries (PA and DPA2 combined). F+T, total RNA libraries (F1, F2, T3, and TG combined; see **Supplementary Table S2**). Yellow shading: among trans-spliced module transcripts, the edited forms outnumber the non-edited forms by a factor of >10–100. However, there is no consistent correlation between editing and end-processing of modules. Pink shading: among module transcripts whose ends are not fully processed, pre-edited modules outnumber edited

[c] 3′-terminal module. This explains the high number of unprocessed + edited forms in PAs libraries, which is due to an enrichment of all poly-adenylated forms.

Supplementary Table S4B. Intermediates of U-appendage RNA editing, end-processing, and trans-splicing [a]

Library	Module	U-tract length	■-	■uuuu / ■uuuu□	■□
PAs	*cox1*-m4	U6	70	195,193	3
	nad4-m4	U2	187	25,473	0
	rnl-m1	U26	57	2,607	1
	y2-m3	U18	311	266	0
	y3-m3	U27	54	9	0
	y3-m4	U17	18	3,328	0
	y4-m1	U28	115	220	0
F+T	*cox1*-m4	U6	104	7,471	0
	nad4-m4	U2	452	207	0
	rnl-m1	U26	118	69,862	0
	y2-m3	U18	1,148	157	0
	y3-m3	U27	2,052	17	0
	y3-m4	U17	11	67	0
	y4-m1	U28	121	13	0

[a] Count of RNA-Seq read pairs supporting the various RNA processing intermediate types. Module transcripts listed have >2 appended Us, are trans-spliced to a downstream module, and are supported by ≥5 reads. Yellow shading marks columns that show that U-appendage editing precedes trans-splicing.

[b] ■-, module transcript with unprocessed 3′ end. ■uuuu, module transcript edited by U-appendage. □, trans-spliced neighbor module transcript.

Supplementary Table S5. Conservation statistics of Nad4 [a]

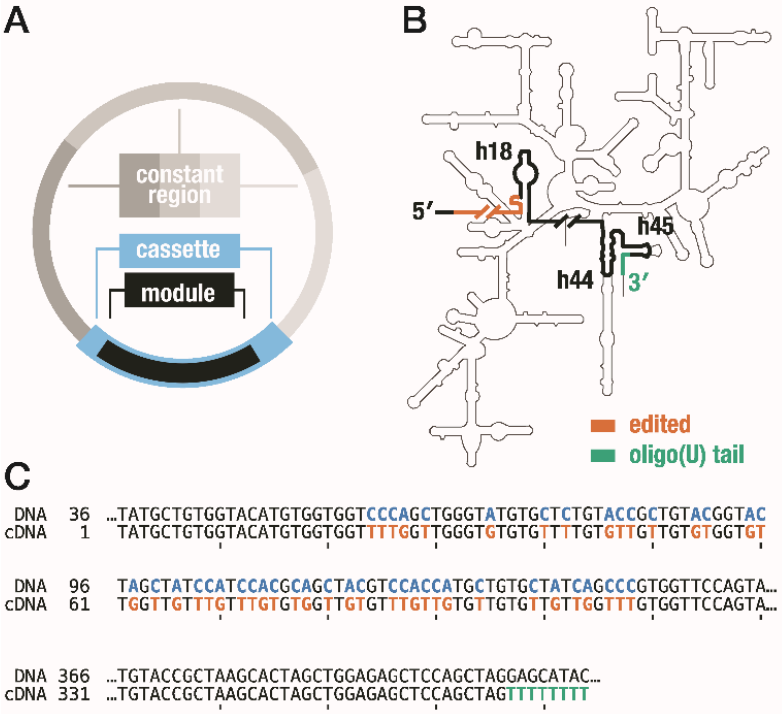
Multiple protein alignment	Dp_edited +others	Dp_pre-edited +others	Dipl_edited	Dipl_pre-edited	Dipl_edited +others	Dipl_pre-edited +others	Others
Conservation score	0.412	0.424	0.591	0.627	0.456	0.482	0.347

[a] Global Nad4 conservation scores based on multiple protein alignments including either the 'edited' or the 'pre-edited' proteins from diplonemids. Scores were determined by MstatsX (see Methods Section of main text). The smaller the score the higher the conservation. The multiple alignments from which the scores were calculated consist of the following sequences. Dp_edited, protein sequence inferred from fully edited *nad4* of *D. papillatum*. Dp_pre-edited, protein sequence inferred from pre-edited (genomic) *nad4* of *D. papillatum*. Dipl_, protein sequences inferred from *nad4* from all four diplonemids used in this study. others, moderately derived sequences from 13 other eukaryotes (see Materials and Methods Section of main text).

Supplementary Table S6. GenBank accession numbers of mitochondrial sequences

Organism	Sequence description	GenBank acc. nr.	Reference
D. papillatum	mtDNA cassettes	KU356490-570	This report
	mt-transcripts	KU341361-80	This report
	cob mRNA	HQ288819	Vlcek, 2011 [247]
	cox1 mRNA	EU123538	Vlcek, 2011 [247]
	cox2 mRNA	HQ288820	Vlcek, 2011 [247]
	cox3 mRNA	HQ288821	Vlcek, 2011 [247]
	nad7 mRNA	HQ288822	Vlcek, 2011 [247]
D. ambulator	*nad4* mRNA (edited + pre-edited)	KU341385-86	This report
D. sp.2	*nad4* mRNA (edited + pre-edited)	KU341387-88	This report
R. euleeides	*nad4* mRNA (edited + pre-edited)	KU341389-90	This report

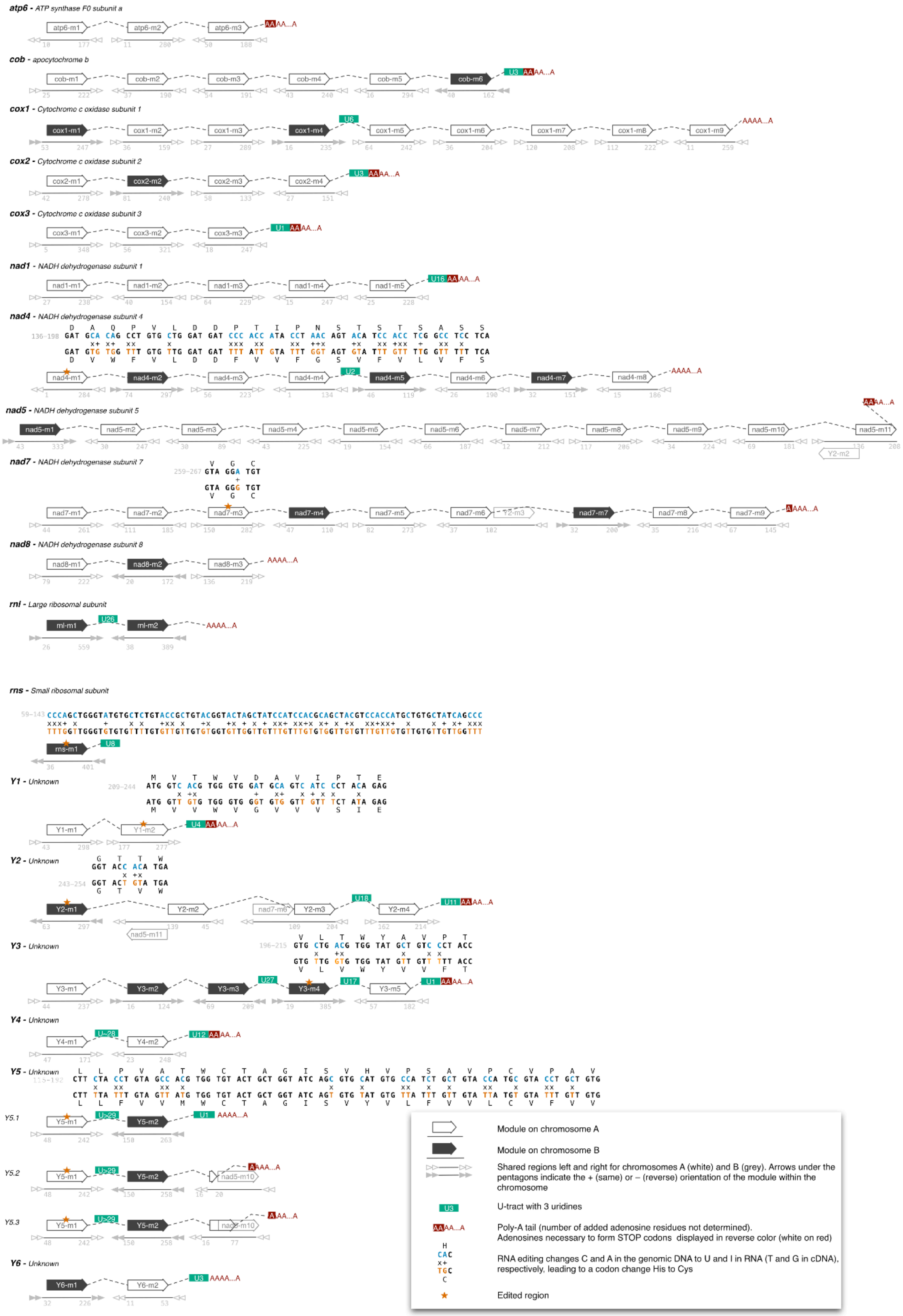
Supplementary Figure S1



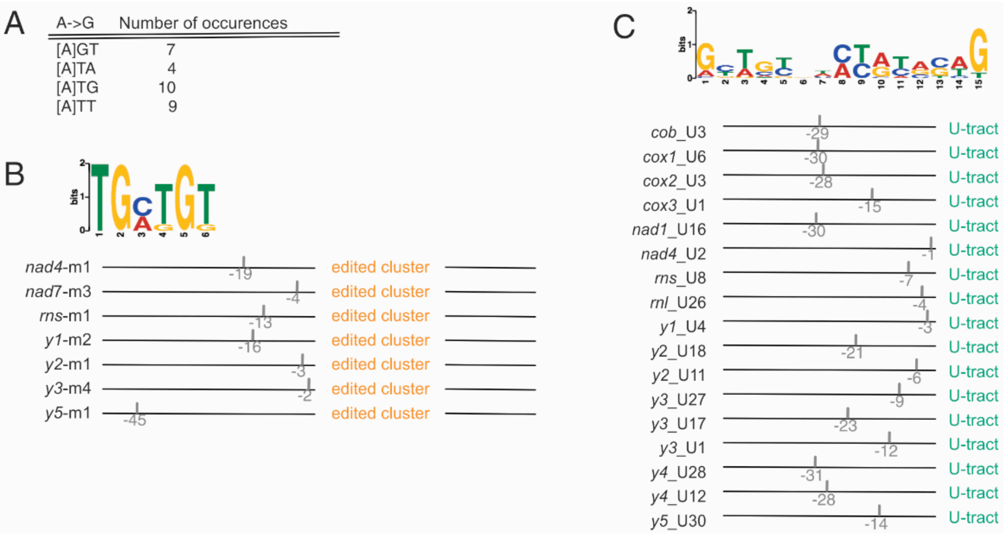
```
 DNA  36  ...TATGCTGTGGTACATGTGGTGGTCCCAGCTGGGTATGTGCTCTGTACCGCTGTACGGTAC
cDNA   1     TATGCTGTGGTACATGTGGTGGTTTTGGTTGGGTGTGTGTTTTGTGTTGTTGTGTGGTGT

 DNA  96     TAGCTATCCATCCACGCAGCTACGTCCACCATGCTGTGCTATCAGCCCGTGGTTCCAGTA...
cDNA  61     TGGTTGTTTGTTTGTGTGGTTGTGTTTGTTGTGTTGTGTTGTTGGTTTGTGGTTCCAGTA...

 DNA 366  ...TGTACCGCTAAGCACTAGCTGGAGAGCTCCAGCTAGGAGCATAC...
cDNA 331  ...TGTACCGCTAAGCACTAGCTGGAGAGCTCCAGCTAGTTTTTTTT
```

Supplementary Figure S1. Mitochondrial chromosome structure and tentative mitochondrion-encoded SSU rRNA of *D. papillatum*. (**A**) Structure of the chromosome encoding the tentative *rns* gene. The gene (the only one consisting of a single module; black) is framed on both sides by a flanking region. The cassette (gene module plus flanking regions; light blue) is bounded by left-hand-side and right-hand-side class-specific constant regions (dark and light grey). Opposite the cassette is the shared constant region that is common to both A-class and B-class chromosomes (grey). (**B**) Secondary structure elements of the tentative mt-SSU rRNA mapped onto the bacterial SSU rRNA model. Positions of the editing cluster and the 3′ oligo(U)-tail are indicated in orange and green, respectively. Elements other than helices h18 and h44+h45 could not be identified with confidence. (**C**) Gene (DNA)/transcript (cDNA) sequence alignment. Substitutions are highlighted in blue and orange, and U additions in green. Ticks indicate nucleotide positions in increments of 10.

Supplementary Figure S2. Module composition and RNA editing of *D. papillatum* mitochondrial genes.


atp6 - ATP synthase F0 subunit a
cob - apocytochrome b
cox1 - Cytochrome c oxidase subunit 1
cox2 - Cytochrome c oxidase subunit 2
cox3 - Cytochrome c oxidase subunit 3
nad1 - NADH dehydrogenase subunit 1
nad4 - NADH dehydrogenase subunit 4
nad5 - NADH dehydrogenase subunit 5
nad7 - NADH dehydrogenase subunit 7
nad8 - NADH dehydrogenase subunit 8
rnl - Large ribosomal subunit
rns - Small ribosomal subunit
Y1 - Unknown
Y2 - Unknown
Y3 - Unknown
Y4 - Unknown
Y5 - Unknown
Y5.1
Y5.2
Y5.3
Y6 - Unknown
Module on chromosome A
Module on chromosome B
Shared regions left and right for chromosomes A (white) and B (grey). Arrows under the pentagons indicate the + (same) or – (reverse) orientation of the module within the chromosome
U3
U-tract with 3 uridines
AAAA...A
Poly-A tail (number of added adenosine residues not determined). Adenosines necessary to form STOP codons displayed in reverse color (white on red)
H
CAC
x+
TGC
C
RNA editing changes C and A in the genomic DNA to U and I in RNA (T and G in cDNA), respectively, leading to a codon change His to Cys
★
Edited region

Supplementary Figure S3



A->G	Number of occurences
[A]GT	7
[A]TA	4
[A]TG	10
[A]TT	9

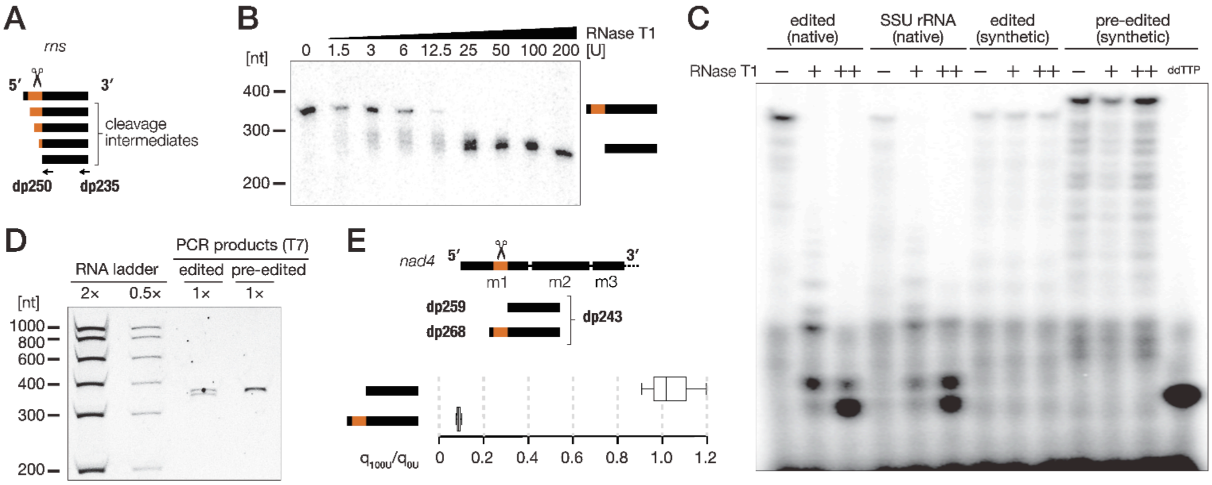
D

U-appendage site	Searched sequence motif representing antisense of RNA editing site	Nr. of reads with motif in	
		antisense	sense
cob-m6-U3	**AAAA**CAGTATATGGG	2	12
cox1-m4-U6	**AAAAAA**GTCCTCCTCGAT	4	1181
cox2-m4-U3	**AAA**CTGTGACCATGC	0	4
cox3-m3-U1	**A**CATGACTAGCTG	0	17
nad1-m5-U16	**AAAAAAAAAAAAAAAA**GGTGGTCAGCGT	0	2
nad4-m4-U2	**AA**CCTCTTGGCATC	3	33
rnl-m1-U26	**AAAAAAAAAAAAAAAAAAAAAAAAAAAA**GAGCTGTAGTA	36	3439
rns-m1-U8	**AAAAAAAA**CTAGCTGGAGCT	0	0
y1-m2-U4	**AAAAA**GCTGCTGAGTG	0	3
y2-m3-U18	**AAAAAAAAAAAAAAAAAAA**CCGAGGTGCTG	0	6
y2-m4-U11	**AAAAAAAAAAA**CGTAGCCCTTCG	0	2
y3-m3-U27	**AAAAAAAAAAAAAAAAAAAAAAAAAAA**CTTAGCAGTAGG	0	3
y3-m4-U17	**AAAAAAAAAAAAAAAAA**TGTGGTAGGTAT	0	8
y3-m5-U1	**A**GGTGATCAGGAG	22	9
y4-m1-U28	**AAAAAAAAAAAAAAAAAAAAAAAAAAAAA**CCTGCAGGATG	0	2
y4-m2-U12	**AAAAAAAAAAAAA**GCTCGTGCTGT	0	0
y5-m1-U29	**AAAAAAAAAAAAAAAAAAAAAAAAAAAAA**CGGTGTACTGGT	0	2
y5-m2-U1	**A**GTACATGCTACT	15	55
Y6-m2-U3	**AAA**CTACGGGTAGAT	0	0
Substitution cluster			
nad4-m1	CCATGA**AAAAACCAAAACAAATACACTACCAAATACAACAAAATCATCCAACACA** **AACCACA**CATCCC	0	9
nad7-m3	TAGACA**C**CCTACA	0	107
rns-m1	AACCAC**AAACCAACAACACAACACAACAAACACAACCACACAAACAAACAACC** **AACACCACACAACAA** **CACAAAACACACACCCAACCAAA**ACCACC	3	1572

y1-m2	TCCTCT**ATAGAAACAACCACACCCACCCACACA**ACCATC	0	42
y2-m1	GATCAT**ACA**GTACCG	0	43
y3-m4	TTGGTA**AAAACAACATACCACACCAA**CACAGC	0	40

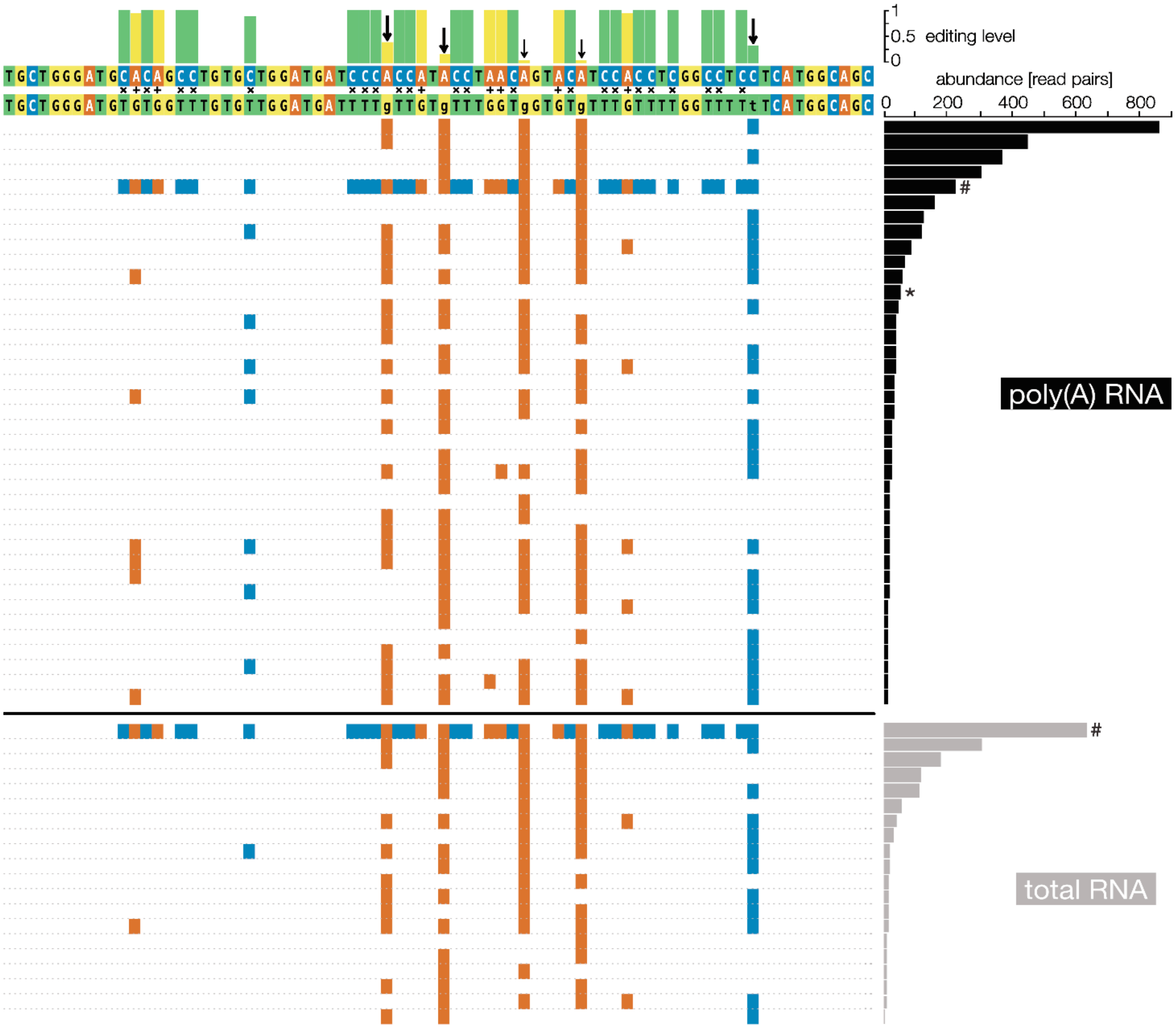
Supplementary Figure S3. Search for potential *cis*-elements (**A**-**C**) and RNA *trans*-factors (**D**) that recognize RNA editing sites. *Cis*-elements **(A-C)** are expected to represent consensus sequence motifs in the vicinity of substitution RNA editing sites, across multiple genes. **(A**) Consensus motifs around individual substitution-editing sites. Motifs were searched 50 nt upstream and downstream of each site, in either the edited or pre-edited sequence. Certain dinucleotides downstream of A-to-I sites are over-represented in the edited sequence. (**B)** Consensus motif bounding substitution-editing clusters. The most significant motif TG[C,A]TGT is located at various distances upstream of clusters, however, it also occurs in non-edited transcripts. (**C**) Consensus motif detected at various distances upstream of U-appendage sites. However, the motif also occurs in non-edited transcripts. (**D**) Search for RNA *trans*-factors characterized by a sequence that is reverse complementary to the RNA editing site plus an anchor. The sequence of the searched motif is given in the second column. For U-appendage editing the anchor is 12 nt long and upstream of the site (on the pre-edited transcript). For substitution editing clusters, the anchor includes 6 nt each upstream and downstream of the cluster. The read counts refer to library F2 including short transcripts (see Methods).

Supplementary Figure S4



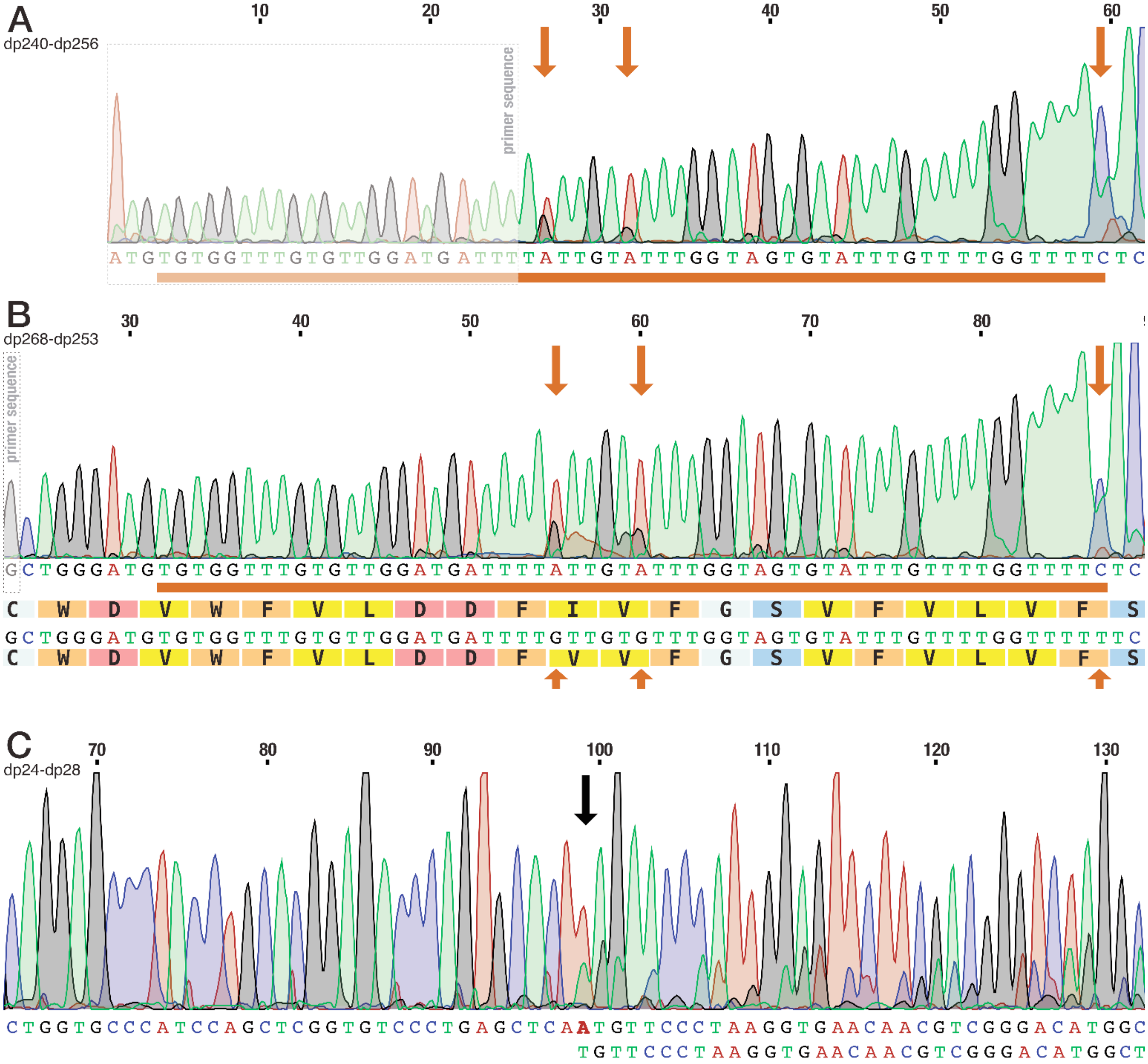
Supplementary Figure S4. Optimization of the glyoxal/borate/RNase T1 assay for the detection of inosines in RNA. (**A**) Design of the primer extension (dp250) and Northern blot hybridization (dp235) assays for mt-SSU rRNA (*rns*) after partial RNase T1 cleavage. (**B**) Northern blot hybridization visualizes the progressive cleavage with increasing RNase T1 concentrations in the editing cluster of mt-SSU rRNA. Five µg of glyoxal/borate-protected total RNA were used per sample. (**C**) Primer extension by reverse transcriptase (RT) using dp235 that anneals with mt-SSU rRNA. Several RNA samples were treated with glyoxal/borate/RNase T1: total RNA labeled 'edited (native)'; mt-SSU rRNA extracted from purified mitoribosomes ('SSU rRNA (native)'), *in vitro* transcribed substitution-edited mt-SSU rRNA ('edited (synthetic')), and *in vitro* transcribed pre-edited mt-SSU rRNA ('pre-edited (synthetic')). The sizes of extension products in 'edited (native)' and 'SSU rRNA (native)' are consistent with RNase T1 cleavage at predicted Is. In turn, 'edited (synthetic)' and 'pre-edited (synthetic)' templates give rise to the same pattern of extension products irrespective of presence of absence of RNase T1, because the templates lack Is and therefore are not cleaved. In the lane labelled 'ddTTP', 'pre-edited (synthetic)' RNA was extended in the presence of ddTTP (no dTTP added) to visualize the position of the 3′-most A residue in the editing cluster. Note that the longest RT-extension product for 'pre-edited (synthetic)' migrates slower than the product obtained with 'edited (synthetic)', although the product length is identical. This difference in migration is due to a different nucleotide composition of the products, with A+C content of 87.5% in the edited *vs* 54.5% in the pre-edited version. (**D**) Electrophoretic separation of denatured PCR products used to synthesize the *in vitro* transcripts 'edited (synthetic)' and 'pre-edited (synthetic'). While both DNA strands of the pre-edited-like transcription template have almost identical apparent mobility, the PCR product used as a template for 'edited (synthetic)' separates into the A+C-rich and G+T-rich strand. The amount of PCR product loaded per lane was 30 ng; bands of the RNA ladders '2x' and '0.5x' represent 60 ng and 15 ng respectively. (**E**) Upper part, design of the RT-PCR assay for detection of RNase T1 cleavage at unprotected Is in the *nad4*-m1 edited cluster. The forward primer binds downstream (dp259) or upstream (dp268) of the substitution editing cluster, and the reverse/RT primer binds within module 2 (dp243). Lower part, quantification of RT-PCR amplicon yield based on band intensity (see **Figure 2D**). q100U/q0U, ratio of products obtained from RNase T1-treated (100 U) vs untreated from three independent experiments. The product covering only non-edited sequence of *nad4* shows no effect of RNase treatment (ratio ~1), whereas the product including the substitution editing cluster is drastically reduced after RNase digestion (ratio ~0.1).

Supplementary Figure S5



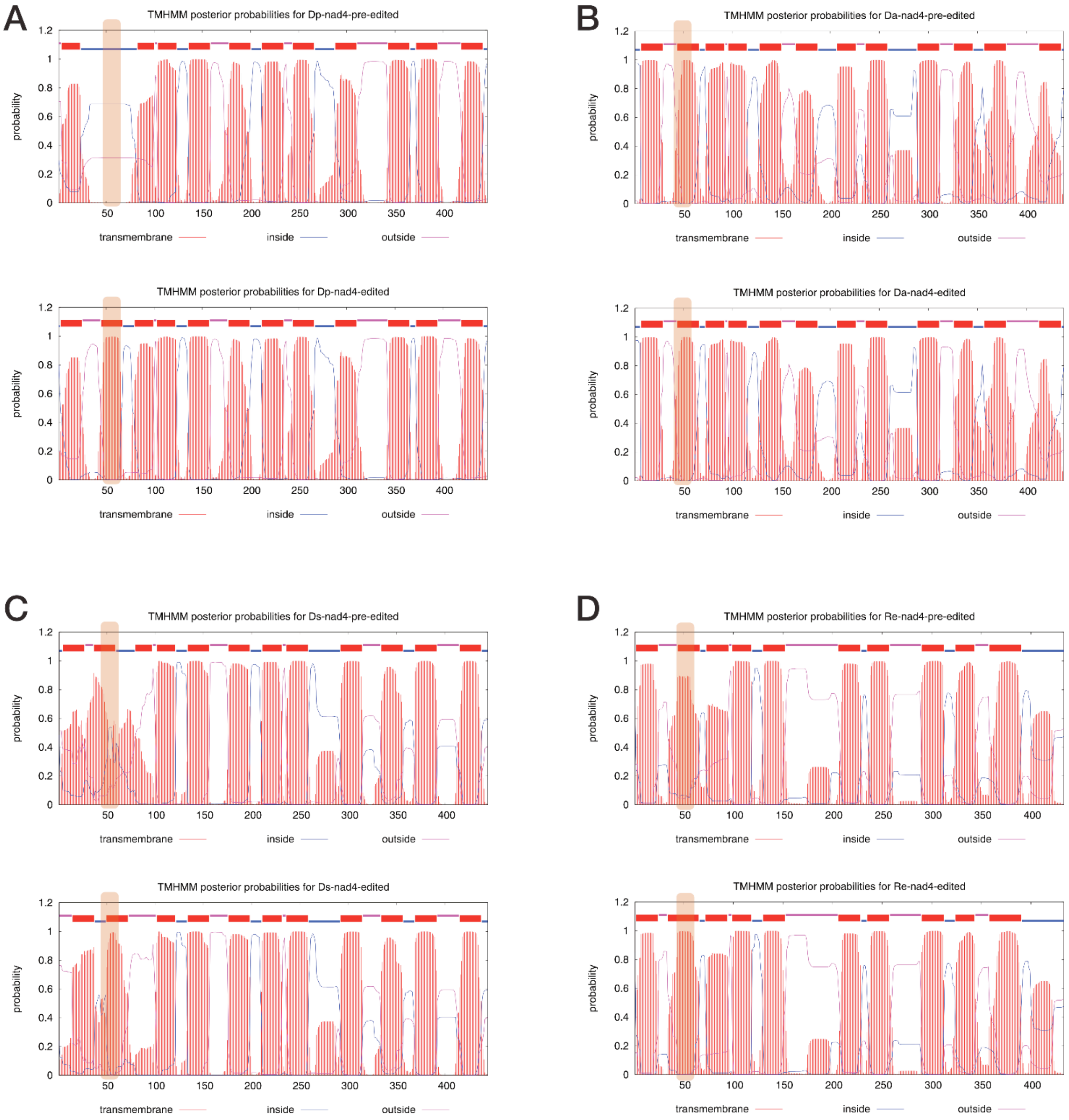
Supplementary Figure S5. The pattern of partial substitution RNA editing of *nad4*-m1 is essentially the same in poly(A) and total RNA. RNA-Seq reads from two poly(A) libraries combined (DPA2 and PA) and from four total RNA libraries combined (F1, F2, T3, TG) were analyzed for the frequency of RNA editing in the substitution editing cluster (positions 129-204 in the corresponding cassette). The histogram on the top summarizes the editing level within the cluster, with green and yellow bars representing the portion of Is and Us generated by deamination, respectively. Arrows point to moderate-frequently (5-38%) edited sites. Thick arrows indicate the three sites, where both editing states were confirmed to co-exist in *nad4*-mRNA. Top-most nucleotide sequence, genomic sequence. Nucleotide sequence below, cDNA sequence showing the edited state of all sites observed to be edited above the 5% exclusion threshold. Lower-case letters, sites edited below 50%. +, x: predominant A-to-G and C-to-T substitution sites, edited above 50%. Charts below sequences: editing patterns observed in RNA-Seq reads of poly(A) RNA (black) vs total RNA libraries (grey). Red and blue squares indicate pre-edited As and Cs, respectively. The patterns shown are supported by >10 read pairs (among 3,475) in poly(A) RNA and >5 read pairs (among 1,609) in total RNA. The right-hand histograms represent the number of reads per editing pattern. The predominant edited form of *nad4*-mRNA, which is shown in all other figures, ranks as first in poly(A) RNA and second in total RNA. #, the entirely pre-edited version, ranking fifth in poly(A) RNA, but first in total RNA. *, the maximally edited transcript (i.e. where all, high- and low-frequency sites are in the edited state), ranking 12th in poly(A) RNA, but below the exclusion threshold in total RNA.

Supplementary Figure S6



Supplementary Figure S6. Trans-spliced polyadenylated transcripts contain partially edited sites. (**A, B**) Sequence traces of RT-PCR products amplified with primers targeting the substitution editing cluster (orange bar) of the *nad4* mRNA. Orange arrows above the traces indicate the partially edited sites seen in RNA-Seq data (nucleotide positions 143, 148, and 185 in cassette *nad4*-m1; see Figure 3). Partial editing leads to two silent mutations and an Ile-to-Val change. (**C**) Sequence trace of the *cox1* RT-PCR product confirms the presence of two mRNA variants that differ by a single A-indel (bold; black arrow). Reverse transcription was performed with the CDS-III primer on poly(A) RNA. PCR products were amplified with indicated primers and sequenced with a primer within ~100 nt from the analyzed site (dp242 for the dp240+dp256 and dp268+dp253 products; dp24 for the dp24+dp28 product).

Supplementary Figure S7



Supplementary Figure S7. Predicted trans-membrane (TM) helices of diplonemid Nad4 proteins. Predictions by TMHMM2.0 are shown. Similar results were obtained with TMpred and Phobius. The position of the editing cluster is highlighted by the orange overlay. Predictions for the pre-edited and edited proteins are depicted in upper and lower panels, respectively. (**A**) *D. papillatum*. (**B**) *D. ambulator*. (**C**) *D.* sp. 2. (**D**) *R. euleeides*. The effect of editing on protein structure is most conspicuous in *D. papillatum*.

Supplementary Figure S8

```
>Naeg.grub.mt
-FDISLIIFFIFNIVWVLFDSFEPSFQFI----QYINLFEYS-------FLFGIDGISIF-FIYLSTFLIPLCLLFS
>Bige.nata.mt
-VALSSAIFVTSLSLWVFFDTLDTRYQFT----DALPGLGSLGVP----LQFGVDGLSLF-FILLTTLITPICLAMS
>Phyt.infe.mt
-IFMSFFIFLMSLPLWILFDKSTSNFQYL----FKIEWISQFNIN----FYLGLDGISLF-FIILTTFLIPICMLIS
>P.litt.mt
-LGVTSLVFVCSLFLWLGFDHSTSKFQYV----VDWLWVPFANIN----LVLGVDGISLF-FVILTTLIFPLCLLAS
>Cyan.mero.mt
-LIISSLTFLITIYLWILFDESYFYFQFV----ISKKILNFELIQ----YFLGVDGISLF-FIILTSLLIFVCFLLN
>Rhod.sali.mt
-FVCSVFTFVISLLLLLFFDSSSCKFQFV----EWLELLIFFNFN----LFIGVDGISIF-FILLTTFLIPICLLVE
>Porp.purp.mt
-FGTAQLTFISSILLWICFDPTISLFQFV----CTINWFPSYNIY----YTIGIDGISLF-FIILTTWLVTTCVLIS
>Chon.cris.mt
-LWISCLTFLFSLLLWIQFNSGTSFFQFS----TTFLWFPFFNLY----YTIGIDGISLF-FILLTTLLIISCILVS
>Mala.jako.mt
-LNTTIITFILSIVLWILFDNTTSKYQFL----LKIELIPFLNLN----YYIGIDGISIF-FLILTTFLVPICILAS
>Marc.poly.mt
-IWTSLITFLYSLFFWIRFENDTAKFQFV----ETIRWLPYSNIN----FYIGIDGISLF-FVILTTFLTPICILVG
>Neph.oliv.mt
-LSVSLINFIASLFLWIQFDHSTPQFQFV----TKITWLPISNLH----IYMGIDGISLF-FVILSTFLVPICILVG
>Recl.amer.mt
-LYASLLTFILSLFLWIFFDKSTSKFQFV----YEVNWIETLNIH----FSLGIDGISLF-LVILTTFLIPLCILTG
>Jako.libe.mt
-LLFSLLTFVFSLWLWILFDESTSKFQFQ----VSYDWFSTSNVS----ATLGVDGISLF-LVLLTTFLIPLCLLVS
>Dipl.papi.Nad4-edited
------------MDAHLSRVSAGTHSGYV------MSWWCWD-------VWFVLDDFVIFGSVFVLVFSWQQSTTVS
>Dipl.ambu.Nad4-edited
---------MAGIDAYLAYTLPAYRTAAV------FSWWHTE-------YCIIVDGLYVFVFVFILVFAMAASAIAH
>Dipl.sp.Nad4-edited
--MACGCAVALLLDCMLSYTSMSMSVVVC------IYWWAHC-------CVFLMDVFFMFIMIFVCTALVATSAMHT
>Rhin.eule.nad4-edited
MVSVTVSAIAYVVLYLLSVDVLSTRDVVCMKDSVLIVWWALP-------LVLVYDGFVCFVGLFVFVYLLAVSCTLH
1        11        21        31        41        51        61        71
```

Supplementary Figure S8. Multiple alignment of the N-terminal region of Nad4 from various eukaryotes. The region includes all residues encoded by the substitution editing clusters in diplonemids. The clusters correspond to amino acids 25-43 (*D. papillatum*), 32-46 (*D. ambulator*), 37-51 (*D.* sp. 2), and 42-54 (*R. euleeides*). The amino acid isoleucine (I), marked in red font color and yellow background, is encoded by the AUU codon whose first codon position is an A-to-I site that remains pre-edited in ~62% of transcripts. When edited, the codon specifies valine (V).

Chapitre 4: Three-dimensional structure model and predicted ATP interaction rewiring of a deviant RNA ligase 2

Sandrine Moreira, Emmanuel Noutahi, Guillaume Lamoureux, and Gertraud Burger

Référence: S. Moreira, E. Noutahi, G. Lamoureux, and G. Burger. 2015. "Three-Dimensional Structure Model and Predicted ATP Interaction Rewiring of a Deviant RNA Ligase 2." *BMC structural biology* 15(1):20.

Contribution des auteurs:

SM conceived the project, and all authors participated in the design of the study. EN performed the 3D model determination and refinement. SM performed the comparative analysis, phylogeny, molecular dynamics simulation, and electrostatic calculation, and evaluated the results. GL provided guidance on structural biology and molecular dynamics simulations. GB was involved in the design of the phylogenetic analysis and the interpretation of results. SM and GB wrote the manuscript. All authors read and approved the final manuscript.

I.Abstract

1. Background

RNA ligases 2 are scarce and scattered across the tree of life. Two members of this family are well studied: the mitochondrial RNA editing ligase from the parasitic trypanosomes (Kinetoplastea), a promising drug target, and bacteriophage T4 RNA ligase 2, a workhorse in molecular biology. Here we report the identification of a divergent RNA ligase 2 (DpRNL) from *Diplonema papillatum* (Diplonemea), a member of the kinetoplastids' sister group.

2. Results

Using homology modeling and molecular dynamics simulations, we established a three-dimensional structure model of DpRNL complexed with ATP and Mg^{2+}. This model was compared with available crystal structures from *Trypanosoma brucei*, bacteriophage T4, and two archaeans. Interaction of DpRNL with ATP is predicted to involve double π-stacking, which has not been reported before in RNA ligases. This particular contact would shift the orientation of ATP and have considerable consequences on the interaction network of amino acids in the catalytic pocket. We postulate that certain canonical amino acids assume different functional roles in DpRNL compared to structurally homologous residues in other RNA ligases 2, a reassignment indicative of constructive neutral evolution. Finally, both structure comparison and phylogenetic analysis show that DpRNL is not specifically related to RNA ligases from trypanosomes, suggesting a unique adaptation of the latter for RNA editing, after the split of diplonemids and kinetoplastids.

3. Conclusion

Homology modeling and molecular dynamics simulations strongly suggest that DpRNL is an RNA ligase 2. The predicted innovative reshaping of DpRNL's catalytic pocket is worthwhile to be tested experimentally.

II.Background

RNA ligase from phage T4, the work horse of molecular biology research, is the best known member of a large protein family encompassing RNA and DNA ligation enzymes [282]. RNA ligases fall into three classes: (i) RNA ligases type 1, (ii) RNA ligases type 2, and (iii) capping enzymes. All nucleic acid ligases share a characteristic nucleotidyltransferase domain in their N-terminal part with five conserved motifs (I, III, IIIa, IV and V) [283]. Two other classes of enzymes that have RNA ligase activity but lack the above structural features are the LigT phosphoesterases involved in RNA splicing [284]-[286] and the recently identified RtcB proteins [287], [288]. In the following, the term "RNA ligase family" will refer to the two former classes that contain a nucleotidyltransferase domain.

RNA ligase 1 enzymes are mainly present in viruses, mammals and fungi [289]. This enzyme class is typically involved in defense as exemplified by its founding member, the phage T4 RNL1, which is deployed in the counter-attack against antiviral strategies of bacteria [290], but is also involved in tRNA intron splicing [284] and in the unconventional splicing initiating the unfolded protein response of the endoplasmatic reticulum. RNA ligases 2 have a broad but punctuated distribution across the tree of life [289]: they are found mainly in viruses -with the archetypical example of T4 RNA ligase 2 [291]- and bacteria, while only a few examples are known in archaea and eukaryotes. The biological role of RNA ligases 2 is unknown, except for the members of kinetoplastids [292].

Kinetoplastids (Euglenozoa) are a group of protozoans, some members of which are causing life-threatening human diseases (leishmaniasis, Chagas disease, sleeping sickness) [293]. These species also display a unique mitochondrial genome structure composed of an intricate network of large and small circular chromosomes [294]. Large chromosomes encode typical mitochondrial protein-coding genes. Small circles specify guide RNAs that serve as proofreading templates for editing pre-mRNAs of mitochondrial genes [295]. Editing proceeds by cutting the pre-mRNA molecule at the place of the mismatch, then adding or removing uridines, and finally religating the two parts of the RNA molecule. It is this last step that is performed by RNA ligase 2. Specifically, two different RNA ligases 2 are involved, one dedicated to adding and the second to deleting uridines as exemplified by the ligases TbREL1 and TbREL2 respectively for *Trypanosoma brucei* [296].

Here we report the identification of a putative new member of the RNA ligase 2 family in *Diplonema papillatum*, a member of diplonemids (Euglenozoa), which are the sister group of kinetoplastids. The corresponding gene was discovered in our search of a candidate enzyme involved in the eccentric post-transcriptional processing in *Diplonema* mitochondria [44], [227]. This protist harbors a highly complex mitochondrial genome sharing certain similarities with that of kinetoplastids. First, the *Diplonema* mitochondrial DNA (mtDNA) is also multi-partite, as it is composed of hundreds of circular chromosomes of two size classes. The difference and uniqueness of the diplonemid mtDNA is that each chromosome contains one short coding region specifying a fragment of a gene. Each gene module is transcribed separately and then trans-spliced to form full-length mRNAs or structural RNAs. The second resemblance with kinetoplastid mitochondria is RNA editing [44], [45]. Uridine insertion and deletion editing in kinetoplastids involves an RNA ligase 2 to reseal the transcript. In *Diplonema*, RNA editing proceeds by uridine appendage at certain module ends, prior to trans-splicing. We hypothesize that an ancestral molecular machinery containing RNA ligase 2 has led to the editosome in kinetoplastids, while it has evolved to perform trans-splicing in the diplonemid branch.

RNA ligases 2 consist of two discrete portions: the N-terminal nucleotidyltransferase domain (amino acids 1 to 234 in T4) and a C-terminal domain (amino acids 244 to 329 in T4) responsible for substrate specificity. The ligation reaction of RNA ligase 2 is ATP and Mg^{2+} dependent [284], [297], [298] and proceeds, like all members of the DNA/RNA ligase family, in three steps. During the first step, ATP adenylates the enzyme on the lysine residue of the conserved KxxG tetramer in motif I of the nucleotidyltransferase domain. In step 2, the covalently linked AMP is transferred to the 5′P of the 'downstream' RNA molecule to be ligated. Finally, the 3′OH of the 'upstream' RNA molecule attacks the 5′P of the 'downstream' RNA by releasing AMP and joining the two RNA molecules (**Supplementary Figure S1**). The crystal structure has been determined for only a few family members, notably T4 RNA ligase 2 [299], [300] and one of the two paralogous mitochondrial RNA ligases 2 from *Trypanosoma brucei*, notably in apo form as well as complexed with a magnesium ion and ATP [301].

In this study, we devise a strategy based on hidden Markov models (HMMs) and structural comparisons to identify proteins of large evolutionary distance to well-studied counterparts in model organisms. Comparative analysis of highly diverged homologs is particularly informative for identifying functionally and structurally important residues that are under elevated selective pressure. Employing this analytic strategy, we identify the gene and model the structure and ligand interactions of a putative RNA ligase 2 from *Diplonema*. The model predicts intriguing innovations in the interaction network between ATP and the residues of the catalytic pocket, which are worthwhile to be tested experimentally by resolving the crystal structure. We discuss possible evolutionary scenarios that led to these innovations.

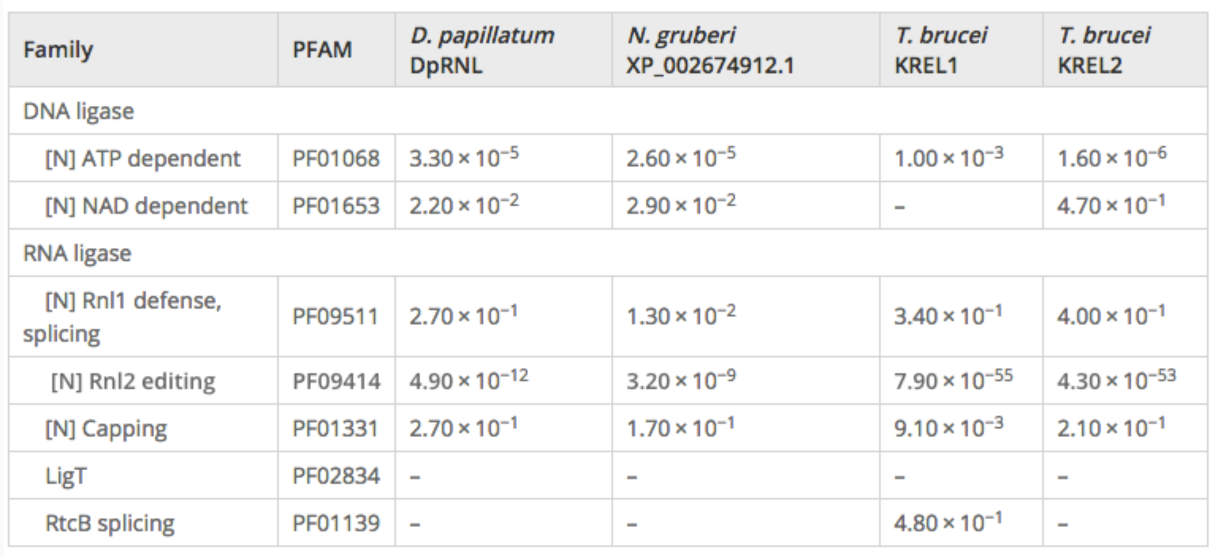
III.Results

1. HMM-based detection of a divergent RNA ligase 2 in *Diplonema*

In general, proteins of *D. papillatum* display a low level of sequence similarity with homologs of other taxa, and are difficult to identify with tools based on sequence similarity such as BLAST [302]. Therefore we employed more sensitive methods based on Hidden Markov Models (HMMs). We used the HMM PF09414.4 from the Protein FAMily database (PFAM) [143], a model that was built based on RNA ligases 2 from all domains of Life including mitochondrial RNA ligases 2 of kinetoplastids. We identified one candidate protein, Dp28902_3, in the conceptual translation of the *Diplonema* draft genome assembly (version no. 2). Expression of this open reading frame was confirmed by RNAseq experiments. The corresponding transcript is poly-adenylated and its steady-state level is about 1/10 compared to the expression of Aspartyl tRNA synthase.

For comparison, we also used HMMs for other RNA and DNA ligase super-families in searches against Dp28902_3 and RNA ligases 2 of *Trypanosoma* (TbREL1, positive control) and the heterolobosean *Naegleria gruberi*. *Naegleria* was chosen because heteroloboseans are the sistergroup of Euglenozoa, and because sequences of this taxon have not been used in building the PFAM HMM. **Table 1** summarizes the corresponding *E*-values. Dp28902_3 has the lowest *E*-value with the PF09414 model, a value that is 10^7 times smaller than the second-best match, which was obtained with the HMM of ATP-dependent DNA ligases. Models for proteins that have a different fold (PF02834-LigT, PF01139-RtcB) did not yield significant *E*-values (>0.05) for either Dp28902_3 or the RNA ligases 2 of *Trypanosoma*. Therefore, Dp28902_3 most likely belongs to the RNA ligase 2 family and will be referred to as DpRNL.

Table 1 Identification of the ligase family to which belongs DpRNL[a]



Family	PFAM	*D. papillatum* DpRNL	*N. gruberi* XP_002674912.1	*T. brucei* KREL1	*T. brucei* KREL2
DNA ligase					
[N] ATP dependent	PF01068	3.30×10^{-5}	2.60×10^{-5}	1.00×10^{-3}	1.60×10^{-6}
[N] NAD dependent	PF01653	2.20×10^{-2}	2.90×10^{-2}	–	4.70×10^{-1}
RNA ligase					
[N] Rnl1 defense, splicing	PF09511	2.70×10^{-1}	1.30×10^{-2}	3.40×10^{-1}	4.00×10^{-1}
[N] Rnl2 editing	PF09414	4.90×10^{-12}	3.20×10^{-9}	7.90×10^{-55}	4.30×10^{-53}
[N] Capping	PF01331	2.70×10^{-1}	1.70×10^{-1}	9.10×10^{-3}	2.10×10^{-1}
LigT	PF02834	–	–	–	–
RtcB splicing	PF01139	–	–	4.80×10^{-1}	–

[a]Family names preceded by an [N] are those containing a Nucleotidyltransferase domain. Each model was searched with HMMer against all the proteins of *Diplonema papillatum*, *Naegleria gruberi* and *Trypanosoma brucei* TREU927. This table presents the *E-value* for the RNA ligases 2 proteins only. The line for the PFAM domain specific for RNA ligases 2 is in bold

2. DpRNL contains a nucleotidyltransferase domain typical for RNA ligases 2

The RNA/DNA ligase super-family is characterized by a nucleotidyltransferase domain including five subdomains (motifs I, III, IIIa, IV, V) [283] located in the N-terminal portion of the protein. We demonstrate the presence of these motifs in DpRNL by three different methods: sequence alignment against PFAM HMM (**Supplementary Figure S2**); multiple sequence alignment of DpRNL and RNA ligases 2 from kinetoplastids, enterobacteriphage T4, and *Naegleria* (**Figure 1**) ; and structural alignment of DpRNL with RNA ligases 2 for which the three-dimensional (3D) structure has been experimentally determined, notably from *Trypanosoma brucei*, the phage T4, and the archaean *Pyrococcus abyssii* (**Figure 2)**.

While the five subdomain motifs are well conserved across all RNA ligases 2 and readily recognizable in DpRNL, the rest of the N-terminal portion of the *Diplonema* protein shows only low sequence similarity to established RNA ligases 2 (e.g., ~18% identity with TbREL1). DpRNL lacks portions of two loops between domains III and IIIa [TbREL1 amino acid (aa) 163-166 and aa 176-205] that are distinctive for kinetoplastid RNA ligases 2, and that have been shown to interact with RNA [301]. Also missing from DpRNL is the loop between domains IIIa and IV of TbREL1 (aa 262-282), a loop that has been predicted to interact with other proteins of the editosome [301]. Finally, the C-terminal portion of DpRNL (aa 178-203) has no recognizable resemblance with, and its length is also shorter than the corresponding region of other RNA ligases 2.

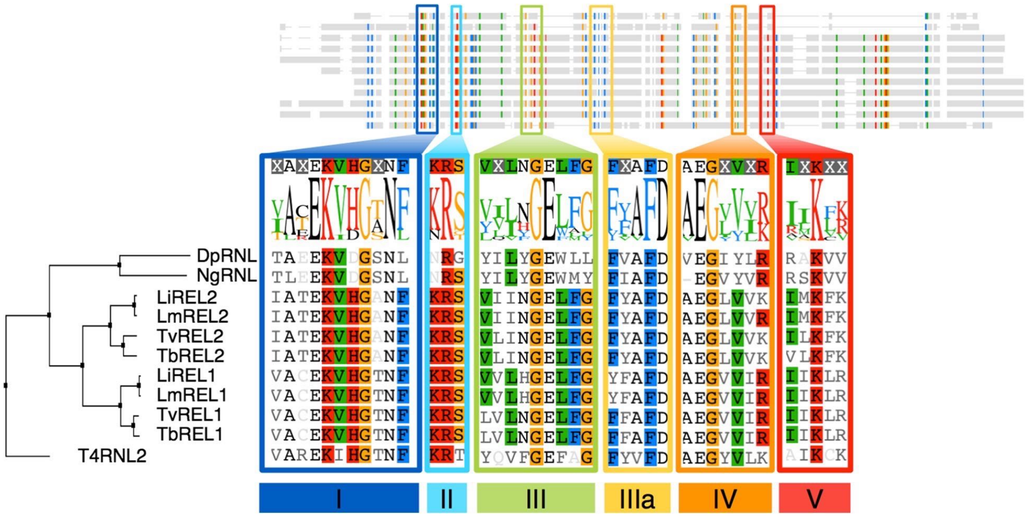


Fig. 1 Delineation of the Nucleotidyltransferase domain. Multiple alignment of RNA ligases 2 from *Enterobacteriophage* T4 (T4RNL2), *Diplonema papillatum* (DpRNL), *Naegleria gruberi* (NgRNL), and four kinetoplastids, *Leishmania infantum* JPCM5 (LiREL1, LiREL2), *L. major Friedlin* (LmREL1, LmREL2), *Trypanosoma vivax* Y486 (TvREL1, TvREL2), and *T. brucei* TREU927 (TbREL1, TbREL2). The six sub-domains (I, II, III, IIIA, IV and V) highlighted in orange, cyan, green, blue, yellow and red, respectively are clearly detectable in DpRNL

```
              I        II      III       IIIa       IV        V
             20 23     41        81       101     158       175
DpRNL      TAEEKVDG    R   YILYGEWLL  FVAFDVYD  EGIYLRS   QRAKVV
             87 90    111       159       209     283       307
Tb 1XDN    VACEKVHG    R   LVLNGELFG  FFAFDIKY  EGVVIRH   TIIKLR
             35 38     55        99       119     204       225
T4 1S68    VAREKIHG    R   YQVFGEFAG  FYVFDIIV  EGYVLKP   VAIKCK
             95 98    115       147       172     227       241
Pa 2VUG    YVEEKVDG    R   LILVGEMAG  FFLFDVQE  EGIIMKS   KIVKYV
Aa 3QWU    WVEEKVDG    R   LVICAEVAG  LFVFDFMK  EGVVFKE   KRAKYI

DpRNL      EEEEELLL    L   EEEEEEEEL  EEEEEEEE  LEEEEEE   LEEEEE
Tb 1XDN    EEEEELLL    L   EEEEEEEEE  EEEEEEEE  LEEEEEE   LEEEEE
T4 1S68    EEEELLLL    L   EEEEEEEEL  EEEEEEEE  LEEEEEE   LLEEEE
Pa 2VUG    EEEEELLL    L   EEEEEEEEL  EEEEEEEE  LEEEEEL   LEEEEE
Aa 3QWU    EEEEELLL    L   EEEEEEEEL  EEEEEEEE  LEEEEEE   LEEEEE
```

Fig. 2 Structural alignment with DALI [303] **of the *Diplonema* model (first line) and four structures: 1XDN (*Trypanosoma brucei*), 1S68 (*Enterobacteriophage* T4), 2VUG (*Pyrococcus abyssi*), and 3QWU (*Aquifex aeolicus*)**

3. The 3D model of apo-DpRNL possesses all structural features typical for RNA ligases 2

The global three-dimensional (3D) model of DpRNL was predicted by I-Tasser [304] (**Figure 3**) and validated with SAVes (http://services.mbi.ucla.edu/SAVES/). Nearly all (96.1%) amino acids have a stereochemical conformation in the "favored" or "allowed" regions of the Ramachandran plot. Only the seven most C-terminal residues are in an unfavorable environment according to the assessment by the tool Verify-3D [305]. While the per-residue analysis of ModFold [306] also found lower quality scores for the C-terminal region, the overall p-value of the model (1.547×10^{-3}) is highly confident. The estimated TM-Score obtained from the standard output of I-Tasser was 0.70±0.12. A TM-score >0.5 usually indicates a model of correct topology, and a TM-score <0.17 means a similarity no better than random. As a whole, the topology of the I-Tasser model of DpRNL is of good quality.

The 3D model of DpRNL is characterized by a core of anti-parallel-twisted β sheets decorated with apical α helices. Two structural sub-domains with similar composition are facing one another. One contains the two extremities of the molecule and consists of an anti-parallel β sheet of four β strands and four helices. The other sub-domain, corresponding to the middle part of the protein, has six β strands and three α helices. The interface between these two sub-domains forms the catalytic pocket of the protein, with the residues of the five nucleotidyltransferase motifs pointing to the pocket's cavity. From the inside to the outside are located motifs I, IV and V on one side, and motifs IIIa, III and II on the other, the two sides facing each other.

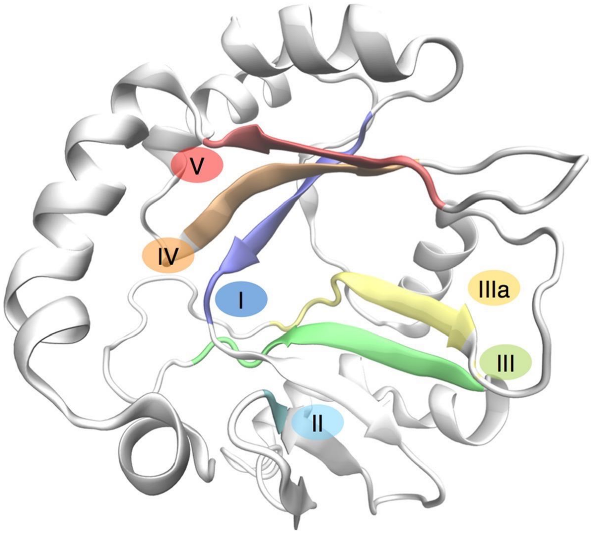


Fig. 3 Three-dimensional model of DpRNL inferred by I-Tasser. The five Nucleotidyltransferase sub-domains are represented in color

4. Molecular dynamics simulations confirm the stability of the DpRNL 3D model

To assess the stability of the proposed DpRNL model and the relative flexibility of the structural domains, we performed a 50-ns molecular dynamics (MD) simulation. The Root Mean Square Deviation (RMSD) of the backbone α-carbon atoms remained stable after 10 ns of simulation with a mean of 4.2 Å (**Supplementary Figure S3A**).

When monitoring the secondary (2D) structure conservation during the simulation (**Supplementary Figure S4**), we observed that the β sheets, which are buried inside the protein, are more stable, whereas the α helices and loops, which are peripheral, are more flexible as reflected by the high Root Mean Square Fluctuation (RMSF) values of the corresponding residues. Specifically, certain residues of the α helices (aa 54-73 and aa 139-154) transiently adopted a 3-10 helix conformation. Flexible α helices and loops are also observed in TbREL1 of *Trypanosoma*, where the exposed regions of the protein interact with the RNA substrate and with other proteins of the editosome [307]. Therefore, the flexible peripheral regions of DpRNL presumably play a functional role as well.

The C-terminal region of DpRNL is linked to the rest of the molecule by a flexible loop, but this region displays less motion than expected. This is because the C-terminal domain is entangled in a network of hydrogen bonds with more N-terminal amino acids. Most stable are the interactions between the carboxyl group of tyrosine at position 203 (DpRNL_Y203, the last residue in the protein) and the lateral chain of two other residues (DpRNL_R41 with 86% occupancy and DpRNL_S24 with 46% occupancy), as well as between the lateral chains of DpRNL_Y203 and DpRNL_Q52. Additional stabilization of this domain comes from a hydrogen bond involving the carbonyl group of DpRNL_K202 in the main chain and the hydroxyl of DpRNL_S49. In conclusion, the 3D model of DpRNL is stable both at the 2D and 3D level. The observed flexibility parallels that of other RNA ligases 2 [300], [307], providing strong support for DpRNL being a functional member of this protein family.

5. Comparative 3D structure analysis of DpRNL with well characterized RNA ligases 2

Compared to recognized RNA ligases 2, DpRNL is more conserved in 3D structure than in sequence. Nevertheless, the β strands of DpRNL are generally shorter than those of its counterparts, resulting in a 15%-30% shorter nucleotidyltransferase domain compared to the enzymes of *Trypanosoma* or phage T4. Pairwise structural comparison with experimentally confirmed structures (**Supplementary Table S1**) reveals only a moderate fit of DpRNL with TbREL1 (RMSD of 3.4 Å), although kinetoplastids are the sister group of diplonemids. The fit is slightly better with the RNA ligases of T4 (T4RNL2; RMSD of 3.2 Å), and *Pyrococcus abyssii* (PAB1020; PDB id 2VUG; RMSD of 2.3 Å), and the putative DNA ligase from *Aquifex aeolicus* (aq_1106; PDB id 3QWU; RMSD of 2.3 Å). Note that PAB1020 was initially annotated as DNA ligase, but more recent experimental studies shown that it catalyzes the ligation of RNA [308].

The proteins from *Pyrococcus* and *Aquifex* are both homodimeric with subunits being held together through the interaction of two peripheral α helices [308]. As DpRNL has no region whose sequence resembles that of these interacting helices, we investigated if the two most C-terminal helices of DpRNL allow dimerization through typical hydrophobic interface contacts [309]. The hydrophobicity map of exposed residues (**Figure 4D** and **Supplementary Figure S5**) shows that the C-terminal helices of DpRNL do not have the propensity to form an hydrophobic surface comparable to that of the archaean ligases. This suggests that DpRNL is active in a monomeric state as are TbREL1 and T4RNL2.

To determine if the nucleotidyltransferase domain of DpRNL contains deviant residues otherwise not found in RNA and DNA ligases, we computed a score of «exceptionality» along the structural multiple alignment from selected enzymes including archaeal and kinetoplastid homologs. Each amino acid in *Diplonema* was assigned an exceptionality score based on the proportion of residues in the corresponding alignment column having common physicochemical properties in other ligases (**Figure 4C**). The amino acid with the highest score is the tyrosine DpRNL_Y161, a position occupied in all other cases by a different, generally aliphatic residue. The second most deviant amino acid is the valine DpRNL_V177, whose position is generally occupied by a basic residue that covalently binds AMP in reaction step 1 [310]. Further exceptional residues in DpRNL are S49, G50, W60, W82, D96, Y104 and R173. The consequences of these substitutions for interactions with RNA and ATP will be discussed in a later section.

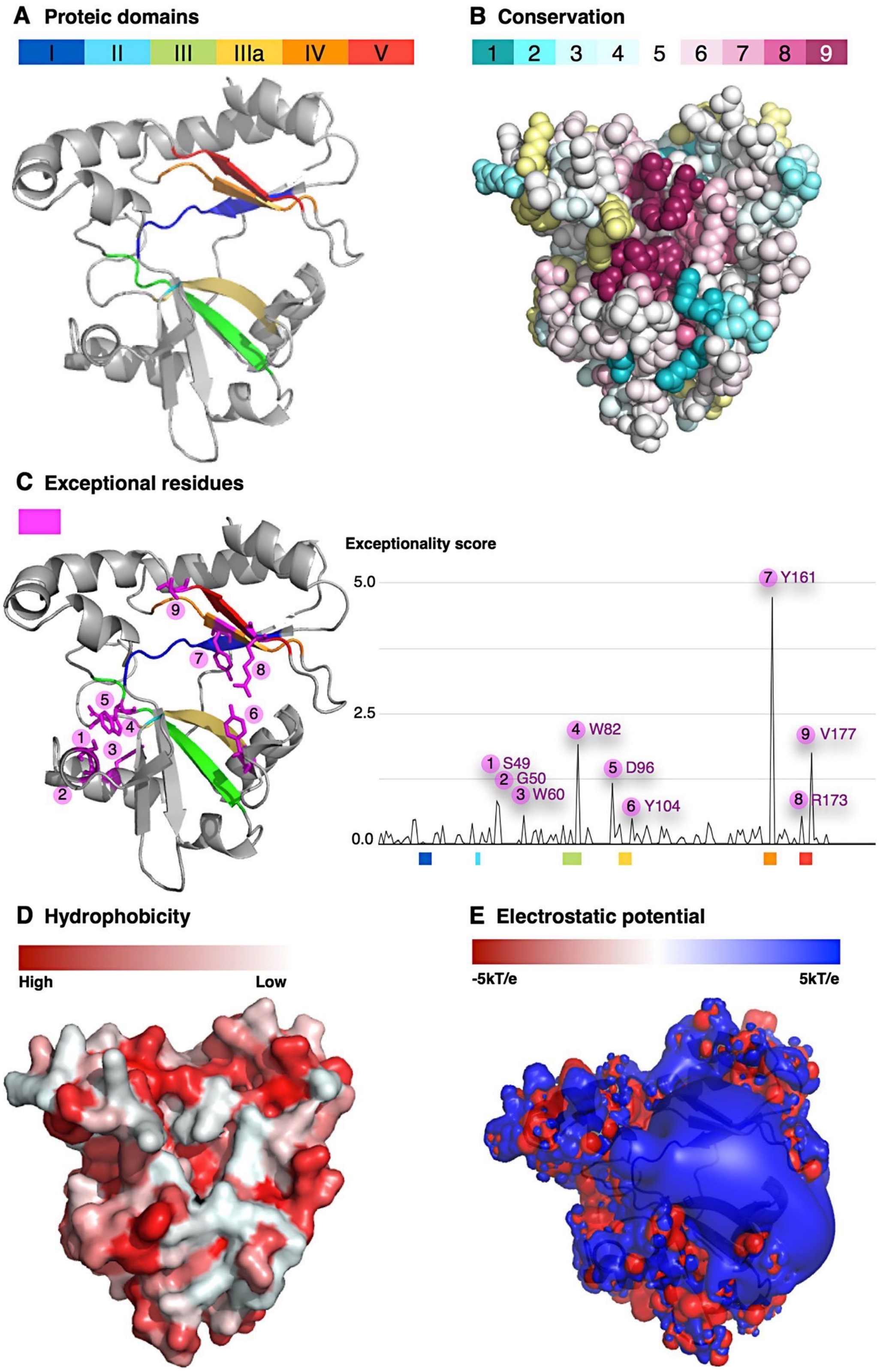

A Proteic domains
I
II
III
IIIa
IV
V
B Conservation
1
2
3
4
5
6
7
8
9
C Exceptional residues
1
2
3
4
5
6
7
8
9
Exceptionality score
5.0
2.5
0.0
1 S49
2 G50
3 W60
4 W82
5 D96
6 Y104
7 Y161
8 R173
9 V177
D Hydrophobicity
High
Low
E Electrostatic potential
-5kT/e
5kT/e

Fig. 4 Protein properties mapped onto DpRNL. a Localisation of the five Nucleotidyltransferase sub-domains. **b** Amino acids conserved across the RNA ligase 2 family. The value 9 (dark purple) represents highest conservation. **c** Exceptional residues as determined in this work. **d** Hydrophobicity. **e** Electrostatic potential

6. Phylogeny of RNA ligases 2

The moderate structural similarity of DpRNL with RNA ligases 2 from the diplonemid sister group raised questions about the phylogenetic relationship of these proteins. We focused our analyses on Excavate taxa, because a broader taxonomic sampling would have resulted in sequences too diverse for meaningful phylogenetic reconstruction. The inferred tree (**Supplementary Figure S6**) shows well supported grouping of kinetoplastid RNA ligases 2, which are split into two subgroups corresponding to the two paralogs (e.g. TbREL1 and TbREL2 in *T. brucei*). The subgroup clustering strongly suggests a duplication of RNA ligases 2 in the kinetoplastid branch prior to the speciation of *Leishmania* and *Trypanosoma*. In contrast, the phylogenetic position of DpRNL in the tree has virtually no support, and the observed affiliation with a homolog from *Naegleria* (heterolobosean) might be an artifact known as long-branch attraction [311], [312]. The phylogenetic reconstruction in this instance suffers from lack of taxa within Euglenozoa (only one diplonemid, no euglenid, and no basal kinetoplastids), and from low sequence conservation. Nevertheless, the tree indicates that DpRNL diverged prior to the gene duplication event seen in kinetoplastids, and that this protein has no specific relationship to the kinetoplastid RNA ligases 2 that take part in mitochondrial RNA editing.

7. DpRNL is predicted to interact with RNA in a T4-like fashion

RNA ligases 2 interact with their substrate via two regions of the protein, the C-terminal domain and regions of the N-terminal nucleotidyltransferase domain that have a positive electrostatic potential. Substrate interaction of the C-terminal domain in kinetoplastid RNA ligases 2 is indirect: the four helices bind a protein partner carrying an OB-fold that, in turn, interacts with the substrate. For example, TbREL1 recruits KREPA2, and TbREL2 associates with KREPA1 [313]. In contrast, the C-terminal domain of T4RNL2 alone suffices for efficiently binding the substrate. In DpRNL, the C-terminal domain carries only two short helices making a TbREL_KREPA-like interaction unlikely. In having a positive electrostatic potential and being rich in residues able to interact with RNA, the C-terminal domain of

DpRNL resembles that of T4RNL2 [300] (**Figure 4E**), and probably also interacts directly with the RNA substrate.

We mentioned earlier that the nucleotidyltransferase domain of DpRNL lacks the two substrate-binding loops of kinetoplastid RNA ligases 2. RNA interaction of loop 1 (TbREL1 aa 167-177) and loop 2 (TbREL1 aa 190-200) had been predicted based on the crystal structure [301] and the calculation of the ensemble averaged electrostatic potential [307], and has been confirmed by an RNA ligation assay with an N-terminal fragment of TbREL1 containing these two loops [301]. The same study also shows that the equivalent N-terminal portion of T4-RNL2, which lacks these loops, does not display this activity. Again, substrate interaction in the *Diplonema* protein must be different from that in kinetoplastid RNA ligases 2 and rather similar to that of T4RNL2.

In the nucleotidyltransferase domain of the phage T4RNA ligase 2, RNA interaction is achieved by a patch of positively charged residues located in the exposed region of central beta sheets, as revealed by the crystal structure of T4RNL2 bound to a nicked nucleic acid duplex (PDB id 2HVR). To identify such regions in DpRNL, we computed the electrostatic potential at the solvent-accessible surface of the protein (see Methods). We found a large region in DpRNL`s nucleotidyltransferase domain with strong positive potential [299] (**Figure 4E**). Superposition of the DpRNL 3D model onto the T4RNL2 structure with bound RNA duplex shows that the potential is distributed in a pattern similar to that in T4RNL2, and in addition, that the duplex broadly overlaps the positively charged regions of DpRNL (**Figure 4E**). However, this region in DpRNL is not completely covered by the duplex. Either the substrate is slightly shifted and|or the unoccupied region interacts with another partner. Still, in this superposition, the two C-terminal helices of the *Diplonema* protein wrap themselves around the nucleic acid like a hook, corroborating the predicted position of the RNA substrate in the DpRNL model.

8. Refinement of the DpRNL structural model by molecular dynamics simulations

RNA ligases 2 typically bind ATP in a covalent fashion during the first step of the catalysis resulting in a ligase-AMP complex (**Supplementary Figure S1**). In a previous section we reported that certain conserved residues otherwise involved in the covalent attachment of AMP, are substituted by different amino acids in DpRNL. To investigate how DpRNL might interact with ATP, we performed an MD simulation after introducing an ATP molecule together with a magnesium ion into the catalytic pocket of the 3D model to mimic the situation at the beginning of the first step of the enzymatic reaction. Our approach has been validated by a control simulation with TbREL1, where ATP and Mg^{2+} assumed stable positions in the catalytic pocket that correspond to those in the crystal structure [301].

MD simulations were performed for 50 ns and 45 ns. We restrained the position of ATP in the catalytic pocket during the first 15 ns (thereafter called the ATP-restrained production phase) followed by four replicates of unrestrained MD simulation during 35 ns. Second, we conducted three independent ATP-restrained productions of 15 ns, each followed by 30 ns unrestrained MD simulation in order to test whether ATP adopts each time the same position (see Supplementary figure S9). We observed that the most important fluctuations during the entire simulation period took place in peripheral helices and loops, while the core β strands stabilized already during the first 10 ns (see lower RMSF values, **Supplementary Figure S7**). However, the conformation of the catalytic pocket was primarily influenced by the subtle motion of lateral chains in the core β strands that took place during the first 10-ns pre-production phase. In particular, the motion of the residues DpRNL_F101 and DpRNL_Y161, which are among the five residues with the lowest RMSF, had the strongest impact, reshaping the whole interaction network with ATP. Interestingly, DpRNL_Y161, which in the initial structure was perpendicular to ATP, turned around to face both the adenine ring and DpRNL_F101. This rotation occurred already during MD the equilibration phase, and the new position of this residue was retained for the rest of the simulation time in six of the seven replicates. A distinct conformation was adopted by the last replicate for which the number of distance violations during the ATP-restrained production phase was much higher (18%), and ATP is more distant from both aromatic residues (5.54 Å from DpRNL_Y161 and 5.79 Å from DpRNL_F101) with a mean angle of 52° (SD = 8.8 Å) with DpRNL_F101 (**Supplementary Figure S8, Supplementary Figure S11, Supplementary Table S2**). Such a conformation is incompatible with π-stacking. The conformation of the six consistent simulations will be referred to as the predominant conformation and analyzed in the following sections, while the deviant conformation will be addressed in the Discussion. To summarize, the predominant 3D model of DpRNL, the pre-production phase locked the catalytic core of the protein in a stable conformation that favors interaction with ATP.

9. Predicted interactions of DpRNL with adenine and ribose of ATP

We compared the predicted interaction network of ATP in the DpRNL model with that inTbREL1, which is the only enzyme for which both the crystal structure of the protein bound to ATP (1XDN) and detailed molecular dynamics simulations are available [307]. ATP interactions of T4RNL2 are similar to that of TbREL1 (homologous residues are listed in **Figure 6**) [299], [300], [307].

The phenylalanine (DpRNL_F101) and tyrosine (DpRNL_Y161), which together sequester the adenine base of ATP in the DpRNL model, establish a π-π stacking interaction with the substrate. This contrasts with the TbREL1 structure, where the base is enclosed by a sandwich composed of the aromatic ring of a phenylalanine (TbREL1_F209, motif IIIa), and a valine (TbREL1_V286). In the *Diplonema* protein the valine is replaced by a tyrosine (DpRNL_Y161), a residue determined as highly exceptional by comparative analysis (see above). This stabilizing interaction reduces greatly the degrees of freedom of the ATP molecule, and gives a significant turn to the interactions in the catalytic pocket by shifting the position of the ligand in DpRNL compared to well characterized RNA ligases. Additional ATP stabilization in DpRNL comes from two hydrogen bonds implicating the amine group of ATP. One hydrogen contacts the carbonyl group of DpRNL_E19 (equivalent to TbREL1_E86) and the other the lateral chain of DpRNL_E18 (which has no equivalent in TbREL1).

In TbREL1, the ribose of the ATP is bound by five residues (TbREL1_I59, TbREL1_K87, TbREL1_N92, TbREL1_R111, TbREL1_E159) allowing the sugar moiety only little mobility. Four out of these five residues (except TbREL1_I59) are conserved in the *Diplonema* protein (**Figure 3**), but only two of the counterparts (DpRNL_N25 and DpRNL_E81) interact with the ribose of ATP (**Figure 5**). Interactions in the DpRNL model take place indirectly through water molecules, and are weaker than the direct salt bridges in TbREL1, thus allowing the larger motions of the sugar that we observed. The two conserved residues that are not involved in stabilizing the sugar (DpRNL_R41 and DpRNL_K20) play an equally important role as detailed in the following.

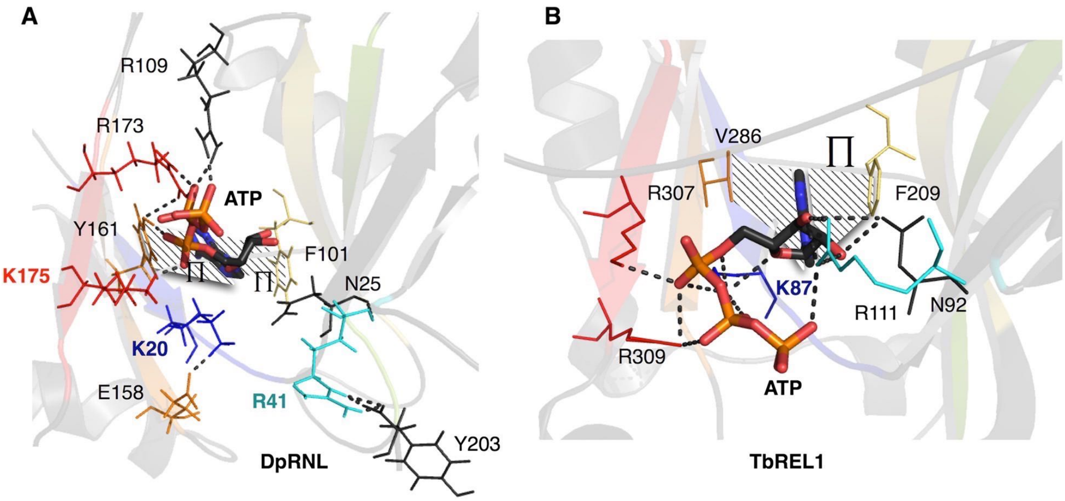
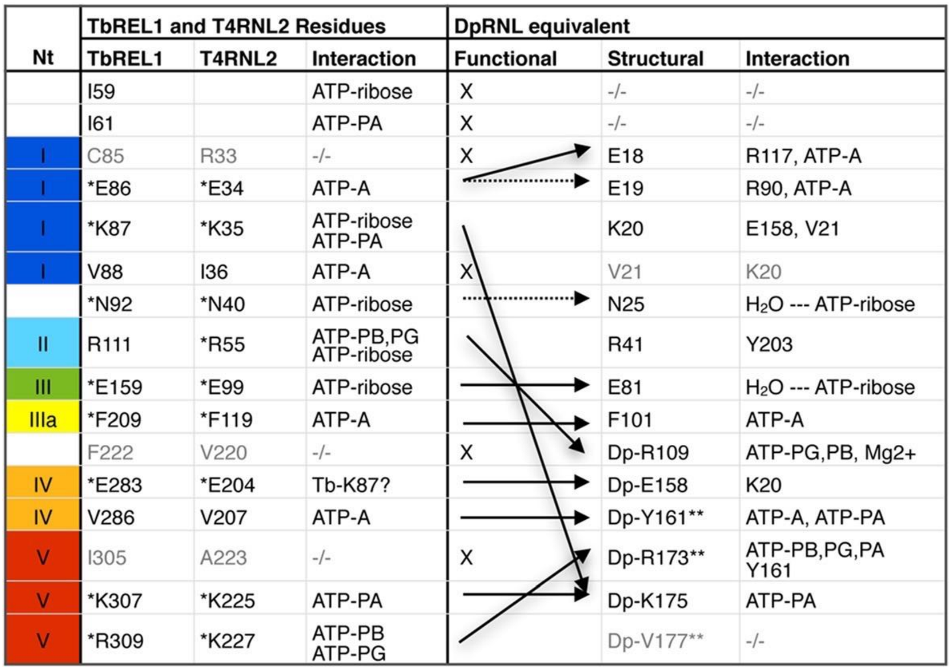


Fig. 5 Catalytic pocket of DpRNL and TbREL1. **a** DpRNL. **b** 1XDN. Dashed lines represent interactions (π-stacking and hydrophobic) with the adenine ring. Important residues are in color



	TbREL1 and T4RNL2 Residues			DpRNL equivalent		
Nt	TbREL1	T4RNL2	Interaction	Functional	Structural	Interaction
	I59		ATP-ribose	X	-/-	-/-
	I61		ATP-PA	X	-/-	-/-
I	C85	R33	-/-	X	E18	R117, ATP-A
I	*E86	*E34	ATP-A		E19	R90, ATP-A
I	*K87	*K35	ATP-ribose ATP-PA		K20	E158, V21
I	V88	I36	ATP-A	X	V21	K20
	*N92	*N40	ATP-ribose		N25	H_2O --- ATP-ribose
II	R111	*R55	ATP-PB,PG ATP-ribose		R41	Y203
III	*E159	*E99	ATP-ribose		E81	H_2O --- ATP-ribose
IIIa	*F209	*F119	ATP-A		F101	ATP-A
	F222	V220	-/-	X	Dp-R109	ATP-PG,PB, Mg2+
IV	*E283	*E204	Tb-K87?		Dp-E158	K20
IV	V286	V207	ATP-A		Dp-Y161**	ATP-A, ATP-PA
V	I305	A223	-/-	X	Dp-R173**	ATP-PB,PG,PA Y161
V	*K307	*K225	ATP-PA		Dp-K175	ATP-PA
V	*R309	*K227	ATP-PB ATP-PG		Dp-V177**	-/-

Fig. 6 Structurally and functionally equivalent residues in DpRNL, TbREL1 and T4RNL2. Residues on the same line are structural equivalents (at the same position in a structural alignment). Residues having the same functional role are connected with an arrow. Dotted arrows indicate partial functional equivalence. X, no functional equivalent was identified. Residues in grey seem not to play a functional role. ATP-A: adenine of ATP; ATP-ribose: ribose of ATP; ATP-PA, PB, PG: phosphate alpha, beta, and gamma, respectively of ATP; *: essential residue; **: exceptional residue; –/–, no structural equivalent identified

10.The triphosphate tail of ATP engages in a rich network of stabilizing interactions

In the predicted predominant conformation of DpRNL, the triphosphate tail of ATP is stabilized by a network of interactions with three basic residues (DpRNL_R109, DpRNL_R173, and DpRNL_K175). In TbREL1, the triphosphate tail is held in place by five residues, TbREL1_I61, TbREL1_K87, TbREL1_R111, TbREL1_K307 and TbREL1_R309 (see **Figure 5**). Among these latter residues, only TbREL1_K307 has the same 3D position and plays the same role as predicted for DpRNL-K175, while TbREL1_I61 has no positional counterpart in DpRNL. The remaining three amino acids have a positional homolog in the DpRNL model, but apparently a different function compared to the *Trypanosome* protein (**Figure 6**).

TbREL1_K87 is the catalytic lysine that in reaction step 1 will covalently bind ATP. This reaction is favored by strong salt bridges between ATP and several other amino acids. DpRNL_K20, the structural equivalent to TbREL1_K87, forms several salt bridges with residues DpRNL_E158, DpRNL_G159 and DpRNL_V21. But instead of promoting the covalent attachment of ATP, the interactions of DpRNL_K20 appear to rather pull this residue away from ATP, the computed distance between DpRNL_K20-Nz and Pα being on average 7.7Å (**Supplementary Table S2**). A candidate residue for covalently binding ATP could be DpRNL_K175, owing to its position apical to the Pα at an average distance of 4.3Å. This distance is comparable to that observed in TbREL1 between K87 and Pα. We propose that the unusual position of ATP in the DpRNL model, as well as the posited substitution of the catalytic lysine, are due to DpRNL_Y161, which, by transforming a simple to a double π-stacking interaction, shifts the position of the ligand.

TbREL1_R111 interacts with the triphosphate tail of ATP, and therefore, the functional homolog of this residue is thought to be DpRNL_R109. However, the positional counterpart of the former residue in our model (DpRNL_R41) plays a radically different role, rather forming hydrogen bridges with residues in the C-terminal region of the protein (maintained for 75.3% of the frames). It should be stressed that simulations with TbREL1 have always been performed with a sequence lacking the C-terminal domain (because the crystal structure was determined with the N-terminal fragment of the protein), so that interactions with the C-terminal domain are not known. In T4, the crystal structure of the adenylated full-length enzyme revealed a salt bridge between two residues of the C domain, R266 and D292, probably reinforcing its structural integrity [300].

Finally, TbREL1_R309 as well interacts with the triphosphate tail of ATP, and in the homologous position of this residue, we find in the DpRNL model a valine (DpRNL_V177). However, this valine seems not to interact with ATP or any amino acid of the catalytic pocket. The functional homolog of TbREL1_R309 is rather DpRNL_R173. Note that both DpRNL_V177 and DpRNL_R173, are "exceptional" residues, and that a non-basic residue at the position corresponding to V177 in T4RNL2 was demonstrated to prevent ligation of ATP [310]). The implications of these findings will be considered in the Discussion section.

IV.DISCUSSION

In the search of an enzyme responsible for the unique trans-splicing in mitochondria of diplonemids, we identified a candidate RNA ligase 2 in the *D. papillatum* genome sequence. Detection of this candidate required the most sensitive HMM search method, because molecular sequences of diplonemids are in general highly divergent [314].

To confirm the sequence-based gene assignment, we constructed a preliminary 3D model of DpRNL that we aligned with RNA ligase 2 family members. Based on the structural sequence alignment, we delineated the boundaries of the predicted functional domains of the *Diplonema* protein. To pinpoint deviant amino acids in the 3D model of DpRNL, we computed a score of exceptionality for each residue. The preliminary structural model was refined by first, eliminating structural inconsistencies and second, performing molecular dynamics simulation. The final model was compared with well-characterized RNA ligases 2.

Available information on how RNA ligases 2 interact with their substrate and ATP comes from crystal structure analysis and enzymatic assays of trypanosome TbREL1 and bacteriophage T4RNL2. In contrast, the presented ligand-binding mode of DpRNL was inferred from molecular dynamics simulations that were based on an *in-silico* modeled 3D structure of the protein. Homology models built from a template that is very distant in sequence space are usually less reliable and tend to be biased toward the template. Even if the main chains of residues interacting with ATP are correctly placed in the DpRNL model, misplacement of their side chains may influence the simulation of ligand binding. To alleviate these difficulties, we have refined the homology model using extensive MD simulation, and have tested the resulting structure using several metrics (e.g. SAVES, ModFold). The predicted unusual ATP-binding mode in the *Diplonema* protein must be considered with this precautionary note in mind.

1. How the postulated rewiring of ATP interactions in DpRNL may have evolved

The present model of DpRNL indicates a reorganization of residue-residue and residue-ATP interactions in the catalytic pocket compared to other ligases, entailing that (i) the ribose is less firmly stabilized than in TbTEL1 and T4RNL2, (ii) the conserved lysine DpRNL_K20 in motif I is pulled away from ATP, and (iii) ATP is now contacted by the conserved lysine DpRNL_K175 in motif V. Such a reshaping would most likely impact steps 1 and 2 of the catalysis (Supplementary Figure S1; Supplementary Figure S10, see legend for detailed description of the hypothesis).

Evolution of such reorganization in the catalytic pocket of DpRNL would require at least two consecutive steps. We speculate that initially, the nearly neutral mutation of a valine to tyrosine DpRNL_Y161 (at the position corresponding to residue 207 in T4RNL2) was made possible by the subsidiary presence of the lysine DpRNL_K175, which incidentally replaced the original catalytic lysine (DpRNL_K20). In this intermediary step, the system could have reverted back to its previous organization. Yet, the accumulation of mutations in a second step (DpRNL_V177, DpRNL_R173 by genetic drift) led to a state with no way back, in the manner of a ratchet [315]. Such a two-step scenario is archetypal of the constructive neutral evolutionary process [268].

As mentioned before, two residues highly conserved at the structural level are predicted to have a different function in DpRNL compared to orthodox RNA ligases 2. These are the ubiquitous lysine (TbREL_K87|T4RNL2_K35) and arginine (TbREL_R111|T4RNL2_R55), which correspond in the structure alignment to DpRNL_K20 and DpRNL_R41, respectively (**Figure 6**). Conservation of the residues in *Diplonema* but not their predicted function raises the question about the underlying selection pressure. Interestingly, the catalytic lysine of proven RNA ligases 2 (e.g., TbREL_K87), has been suggested to also interact with the RNA substrate, notably in the reaction step 3 [316] (**Supplementary Figure S1**). Therefore, we speculate that both DpRNL_K20 and DpRNL_R41, may be subject to a negative selection in favor of conserving a second yet unrecognized role. The key message is that the observation of constant sites across an otherwise diverse family is not necessarily indicative of an identical molecular function of the corresponding residues, as residues can play multiple (structural and catalytic) roles in the corresponding protein [317].

2. The biological process involving DpRNL

We found that sequence- and structure-wise, mitochondrial RNA ligases 2 of kinetoplastids are not the closest homologs of DpRNL. Specifically, the 3D-structure model of DpRNL does not fit better the structure of TbREL compared to that of RNA ligases 2 from a bacteriophage or an archaean. Further, phylogenetic analysis of RNA ligases 2 did not group together the kinetoplastid and diplonemid proteins, but placed DpRNL without support next to a member of the heteroloboseans, a group that emerged prior to Euglenozoa. The large distance between kinetoplastid RNA ligases and DpRNL is probably due to a divergent, accelerated evolution and hyper-specialization of both the kinetoplastid and *Diplonema* proteins. Therefore, we cannot extrapolate from TbREL the biological process in which DpRNL may be involved.

At present it is unknown whether or not DpRNL acts inside mitochondria. There is no recognizable signal in the inferred protein sequence indicative for import into mitochondria or any other subcellular localisation. After translation, DpRNL may either remain in the cytoplasm or be imported into mitochondria by one of the cryptic signals reported for proteins of several other eukaryotes [318]. If DpRNL indeed ends up in mitochondria, then its interaction partner must be fundamentally different to those of the kinetoplastid TbREL, because of significant structural differences between the two proteins (e.g. characteristics of the C-terminal domain, the pattern of electrostatic surface potential, and the absence of interacting loops). Our *in silico* analyses have prepared the ground for determining experimentally the location of DpRNL in the cell, the protein and RNA partners with which it may interact, and ultimately, via 'guilt by association', the biological process in which it participates.

V.Conclusion

RNA ligase 2 from bacteriophage T4 is widely used as a tool in molecular biology, in particular for massively parallelized RNA sequencing technologies. Enzyme versions have been engineered with higher efficiency and fidelity than the natural protein. Specifically, the truncated version of the RNA ligase from phage T4 produces less concatemer side products and is 10 times more active than the natural enzyme [299]. Further, attempts have been undertaken to abolish concatemer formation of T4 RNA ligase by directed mutation of specific amino acids (substitution of T4RNL2-K227 by glutamine abolishes reversibility of the second step of the reaction) [310]. Comparative analysis with divergent RNA ligases such as DpRNL are bound to reveal unrecognized evolution-born innovations and to pinpoint residues otherwise not expected to be relevant enzymatically. Our *in silico* analysis suggests that DpRNL activity relies on structure-function innovations not present in the commonly used RNA ligases, which might reveal suitable for future applications in biotechnology.

VI.Methods

1. Identification of RNA ligase 2

We identified RNA ligase 2 in the draft version of the *D. papillatum* nuclear genome obtained from a Mira V3.4.1.1 [319] assembly of 7.5 million 454 reads at a coverage of ~ 10X. The search was performed with PFAM [143] domain PF09414 present in kinetoplastid RNA ligase employing HMMer 3 [320], [321] using the maximum sensitivity option (parameter --max). We found a single significant hit (E-value = 1.3e-06) in the *Diplonema* sequence matching a hypothetical protein (DpRNL). The identification of the domains characteristic for the RNA ligase 2 family was first performed by analysing the alignment of DpRNL with the PF09414 HMM domain in the HMMer result file, then by a multiple alignment of the two ligase paralogs from four *Leishmania* species (*L. braziliensis*: LbrM.20.5890 and LbrM.01.0620; *L. mexicana*: LmxM.01.0590 and LmxM.20.1730; *L. major Friedlin*: LmjF.20.1730 and LmjF.01.0590; *L. infantum*: LinJ.01.0610 and LinJ.20.1700) and six *Trypanosoma* species (*T. brucei TREU927*: Tb09.160.2970 and Tb927.1.3030; *T. brucei Lister strain 427*: Tb427.01.3030 and Tb427tmp.160.2970; *T. brucei gambiense*: Tbg972.1.1840 and Tbg972.9.2300; *T. cruzi CL Brener Esmeraldo-like*: Tc00.1047053506363.110 and Tc00.1047053511585.20; *T. cruzi CL Brener Non-Esmeraldo-like*: Tc00.1047053506975.9 and Tc00.1047053510155.20; *T. congolense*: TcIL3000.1.1450 and TcIL3000.9.1420; *T. vivax*: TvY486_0101350 and TvY486_0901490).

The specificity of PF09414 in detecting RNA ligases 2 was evaluated by comparing the score of all PFAM domains of DNA and RNA ligases against (i) the *Diplonema* candidate RNA ligase, (ii) the two well characterized RNA ligases from *Trypanosoma brucei* TREU927 (TbREL1, Gene ID = Tb927.9.4360 and TbREL2, Gene ID = Tb927.1.3030) downloaded from TriTrypDB [322], and (iii) the RNA ligase from *Naegleria gruberi* (XP_002674912.1), a protist diverging basally to Euglenozoa.

2. Three-dimensional structure modeling

The three-dimensional model of DpRNL has been determined by I-Tasser (the Iterative Threading Assembly Refinement program) web server (http://zhanglab.ccmb.med.umich.edu/I-TASSER/) [323] using default parameters (no restraints, no guide or exclusion template). I-Tasser selected the structure of the DNA ligase of *Aquifex aeolicus* (PDB ID = 3QWU) as the closest structural homolog of DpRNL and proposed five candidate models. Then, we refined the models with ModRefiner [324], and evaluated the quality of the models with tools available from the SAVeS Web server (Structural Analysis and Verification Server http://nihserver.mbi.ucla.edu/SAVES/) and ModFold [306]. From the five models proposed by I-Tasser, we selected the one having the lowest structural variations compared to the template, and the best structural qualities according to SAVeS.

3. System preparation for molecular dynamics simulations

Two different molecular dynamics simulation protocol were used for DpRNL. To investigate the stability of our model, we used the apo form of the protein (apo-DpRNL). To examine the interactions between the ligand and the protein, we used DpRNL with bound ATP and Mg^{2+} (DpRNL_ATP+Mg^{2+}). In this experiment, we superimposed DpRNL onto TbREL1, the *Trypanosoma* homolog of DpRNL crystallised with ATP (PDB ID = 1XDN), and manually copied the ATP and Mg^{2+} residues from 1XDN to the corresponding position in DpRNL. We added hydrogens when needed with WHATIF [325] and rendered the file CHARMM compatible by employing the PDB Reader of CHARMM-GUI [326].

4. Molecular dynamics simulations

All molecular dynamics (MD) simulations were performed with the Gromacs 4.0.5, 4.6.5, 5.0.1 and 5.0.2 software [327] and CHARMM27 force field [328]. We modified the charmm27.ff force field [329] files in Gromacs to add topology and parameter information for ATP from toppar_all36_na_nad_ppi.str by following the procedure specified in the Gromacs manual (http://www.gromacs.org/Documentation/How-tos/Adding_a_Residue_to_a_Force_Field). Proteins and ligands were solvated in a cubic box of TIP3P water molecules at a distance of 3 nm (30Å) from the solute. The net charge of the system was neutralized by addition of six chloride ions for the DpRNL apo system, four chloride ions for DpRNL+ATP+Mg^{2+} and five sodium ions for TbKREL1+ATP+Mg^{2+}. The cut-off for short-range van der Waals and electrostatic interactions was 1.0 nm (default values), and PME (Particle Mesh Ewald) was used for long-range interactions in all simulations. First, we performed an energy minimisation by steepest descent to remove possible spurious contacts until convergence to a maximum force of 1000 kJ/mol/nm on any atom of the system (850 steps). For all MD simulations, the leap-frog formula was used to integrate the equations of motion. Then two MD equilibrations of 100 ps each (25,000 steps with 2 fs timesteps) were performed with restrained positions of protein and ligand. For the first NVT (constant number of particles, volume, and temperature) equilibration, the temperature was set to 300 K using the V-rescale thermostat [330] with separate baths for protein and non-protein atoms. Then, for the subsequent NPT (constant number of particles, pressure, and temperature) equilibration, the Parrinello-Rahman barostat [331], [332] was used in addition to the V-rescale thermostat in order to couple the pressure to 1 bar. Following these pre-production steps, MD simulation productions were performed on apo-DpRNL and on holo-DpRNL loaded with ATP and Mg^{2+} .

For DpRNL apo, we performed a 50 ns simulation with 2 fs timesteps. The 2D structure conservation during the simulation period was measured using the timeline plugin of VMD [333]. For DpRNL loaded with ATP and Mg^{2+}, we performed MD simulations of 50 and 45 ns in total. During a preliminary simulation, ATP escaped from the catalytic pocket. Therefore, as a precaution, we restrained its position during the initial 15 ns of the production simulation (referred to as ATP-restrained phase), to let the protein equilibrate around the ligand, and after lifting the restriction, the simulation was continued. First, we used the same 15 ns ATP-restrained simulation (15R0) that we extended by four independent 35-ns MD simulations (replicates 15R0+35_1 to 15R0+35_4). Second, we ran three independent 15-ns restrained simulations (15RI, 15RII and 15RIII) followed by 30 ns MD simulations. When measuring the distance between the two molecules during the initial time interval, we noted that the restraint was used in less than 1% of the frames for all the predominant replicates (15R0, 15RI, 15RII) and in 18% of the frames for the deviant replicate (15RIII) (**Supplementary Figure S8**). As an anchor of the restraint, we chose DpRNL_F101 because first, this residue is highly conserved among ligases; second, it is positioned deeply inside the catalytic pocket; and third, in *Trypanosoma* TbREL1, the adenine of ATP has been shown to make π-stacking interactions with the homologous position, TbREL1-F209 [301]. We set a distance restraint of 0.3 nm around the initial distance *ri* between each pair of atoms from the phenyl group of DpRNL_F101 and the pyrimidine ring of the adenine, meaning that there is a component for the restraint added to the potential energy function for *ri* > *ri* + 0.30 nm and *ri* < *ri* - 0.30 nm. Three distances are set: *r0* = *ri* - 0.30 nm, *r1*= *ri* + 0.30 nm and *r2* = *r1* + 1 nm. The potential for the distance restraints is quadratic below *r0* and between *r1* and *r2*, and linear above *r2*.

To test whether inserting ATP + Mg^{2+} in the catalytic pocket of DpRNL leads to a realistic positioning of the ligands, we performed a control experiment on TbREL1. To prepare the system, we replaced the selenomethionine used for crystallisation with methionine, then we ran an MD simulation first on the apo protein for 15 ns. Then, we used the structure from the last frame of the previous simulation as a starting point, inserted ATP + Mg^{2+} into the molecule, and ran a simulation for 30 ns.

5. Exceptional residues

In order to identify exceptional residues in the candidate RNA ligase of *Diplonema*, we computed a score measuring how unexpected each residue of the protein is. Using the I-Tasser model of DpRNL as the query structure, we searched for structural "neighbors" with DALI (http://ekhidna.biocenter.helsinki.fi/dali_server/, [303]): we selected 23 unique RNA and DNA ligases whose structure have the highest percentage of identity and the lowest Root Mean Square Deviation (RMSD), and performed a multiple structural alignment including DpRNL. For subsequent computations, we used the alignment without expanding the gaps, meaning that inserted segments relative to DpRNL are hidden. For each position in the 23 proteins, we computed the entropy s as given by [334] which represents the diversity of amino acids for a given position. The entropy s at position l is $s(l) = -\Sigma_{i=1}^{6} P_i(l) \log P_i(l)$ where i is the category of amino acid (1: aliphatic, {AVLIMC} 2: aromatic {FWYH}, 3: polar {STNQ}, 4: positive {KR}, 5: negative {DE}, 6: special {GP}), and $P_i(l)$ is the proportion of amino acids belonging to category i at position l. At a given position, amino-acid categories for which $P_i(l)$ is null are ignored. If the entropy is low, then the position is conserved among family members. The entropy is set arbitrarily to 0 when the position in the multiple alignment contains more than 50% gaps. We designed an exceptionality score S at position l for amino acids of DpRNL as $Sl = (P_{max}(l) - P_i(l)) / s(l)$ where $P_i(l)$ is the proportion in the previously computed multiple alignment of the amino acid observed at position l for DpRNL, and $P_{max}(l)$ is the proportion of the most abundant amino-acid category (the category that we expect).

6. 3D model analyses

Trajectory analyses were performed with R [335], VMD [336] and PyMOL [337]. Hydrogen bonds were computed using VMD with a distance cutoff of 3.0 Å and an angle cutoff of 30°. The evolution of the secondary structure [333] was computed *via* the timeline plugin of VMD based on the STRIDE algorithm [338]. The conservation surface was colored with the web server ConSurf [339] using the structural multiple alignment performed by DALI as input and with the Bayesian method for computing the evolutionary rate [340].

The electrostatic potential of the molecule was computed by the classical calculation using the last frame of the simulation, employing the APBS web server (http://www.poissonboltzmann.org/) [341]-[343] and visualized using the dedicated APBS plugin of PyMOL. The isovalue cut-off for the analyses was set to $+5k_BT/e$ (blue) and $+5k_BT/e$ (red). For DpRNL, this procedure was sufficient to reveal a large region with positive potential, having the propensity to bind RNA. In contrast, for TbREL1, the classical potential calculation (using Delphi [301]) identified only small positive patches. To find a positive region sufficiently large for RNA binding in TbREL1, the authors had to calculate an ensemble average on their 70-ns simulation [307].

7. Expression

The expression of the gene coding for DpRNL was assessed by mapping RNA-seq reads from a total-RNA library of *D. papillatum* onto the contig carrying the gene. Library construction and read processing have been described earlier [227]. Cutadapt version 1.2.1 [138] was used to remove adapters at 5' and 3' termini of reads with an error rate of 0.1 and to clip low-quality sequences with a threshold of 20. Reads < 20 nt were discarded, leaving 29 million paired reads, which were mapped with Bowtie2 [122] onto the 1314-nt long contig containing the DpRNL reading frame. Output files in sam format were subsequently transformed into 'bam' files with SAMtools version 1.4 [128]. Alignments were visualized with tablet version 1.13.05.17 [177].

8. Phylogenetic reconstruction of RNA ligases 2 from Excavata

We identified RNA ligase 2 proteins in Excavata species by searching with the same PFAM HMM PF09414 as used for *Diplonema*. Sequences were aligned using MAFFT with option «--localpair» (for distantly related species with a single alignable domain). The multiple alignment was refined by successive re-alignment of the sequences on a guiding hmm model built from the alignment with HMMer 3 [320], [321]. The best scoring alignment according to HMMer was selected and filtered with an in-house script to retain positions with less than 30% gaps and a conservation score greater than 8 as given in the stockholm format. We reconstructed the phylogeny with RaXMLHPC v.7.2.6, a maximum likelihood method, using a gamma distribution to model the heterogeneity of substitution rate over sites and the WAG substitution matrix. A Bootstrap analysis of 100 runs was performed to assess the significance of each node.

Availability of supporting data

The sequence of DpRNL is available under Genbank accession number KT828338. The 3D model is included as additional files in PDB format. Alignment is available on request.

Additional files

Additional file 1: Supplementary data. Supplementary result, figures and tables for DpRNL identification, phylogenetic study, 3D model properties, and complementary analyses of MD simulations (DOC 3297 kb)

Additional file 2: DpRNL model. Atomic coordinates of DpRNL model. (PDB 258 kb)

Additional file 3: ITasser threading of DpRNL on 3QWU. Atomic coordinates of DpRNL and 3QWU model. (PDB 11 kb)

Abbreviations

2D: Secondary structure; 3D: Tertiary structure; ATP: Adenine triphosphate; DpRNL: RNA ligase 2 from *Diplonema* papillatum; HMM: Hidden Markov model; MD: Molecular dynamics; PFAM: Protein FAMily database;

T4RNL2: RNA ligase 2 from bacteriophage T4; TbREL1: RNA ligase 2 from *Trypanosoma brucei*, paralog of KREL1, Tb927.9.4360.

VII.Additional file 1

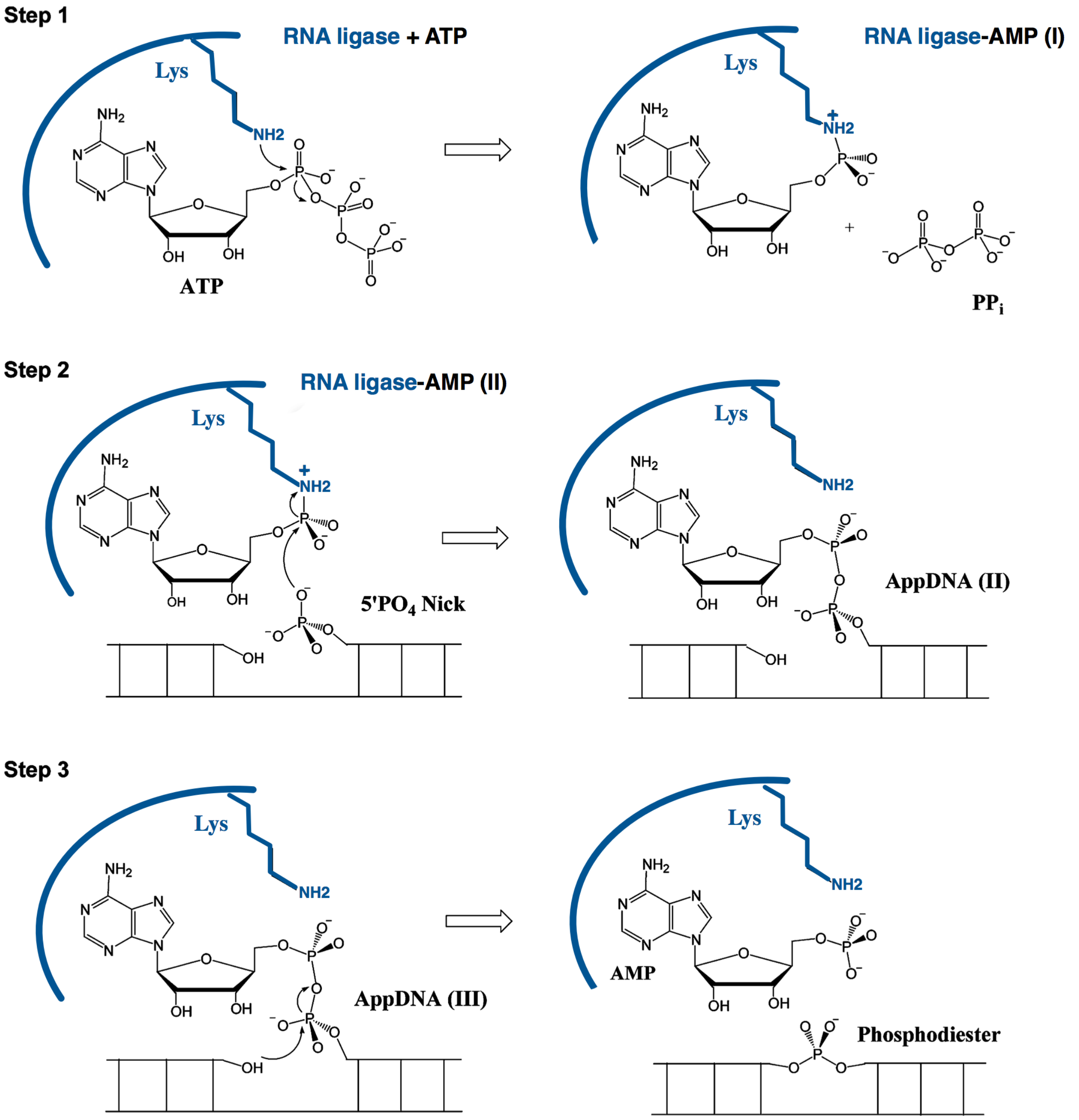
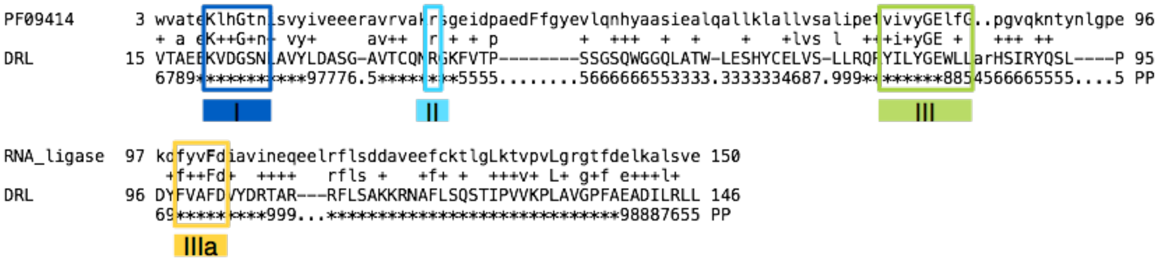


Figure S1. Catalytic steps of the ligation reaction performed by an RNA ligase 2 adapted from [300]



```
PF09414      3 wvateKlhGtnlsvyiveeeravrvakrsgeidpaedFfgyevlqnhyaasiealqallklallvsalipetvivyGElfG..pgvqkntynlgpe 96
               + a eK++G+n+ vy+      av++  r + + p         +  +++  +  +    +    +lvs l  +++i+yGE +    +++ ++
DRL         15 VTAEEKVDGSNLAVYLDASG-AVTCQNRGKFVTP--------SSGSQWGGQLATW-LESHYCELVS-LLRQFYILYGEWLLarHSIRYQSL----P 95
               6789*************97776.5*********5555........5666666553333.3333334687.999*********8854566665555....5 PP

RNA_ligase  97 kdfyvFdiavineqeelrflsddaveefcktlgLktvpvLgrgtfdelkalsve 150
                +f++Fd+  ++++   rfls +   +f+ +  +++v+ L+ g+f e+++l+
DRL         96 DYFVAFDVYDRTAR---RFLSAKKRNAFLSQSTIPVVKPLAVGPFAEADILRLL 146
               69**********999...****************************98887655 PP
```

Figure S2. Alignment of the PF09414-Rnl2 HMM model with DpRNL. The alignment was produced by HMMer [320], [321]. Four Nucleotidyltransferase sub-domains are highlighted.

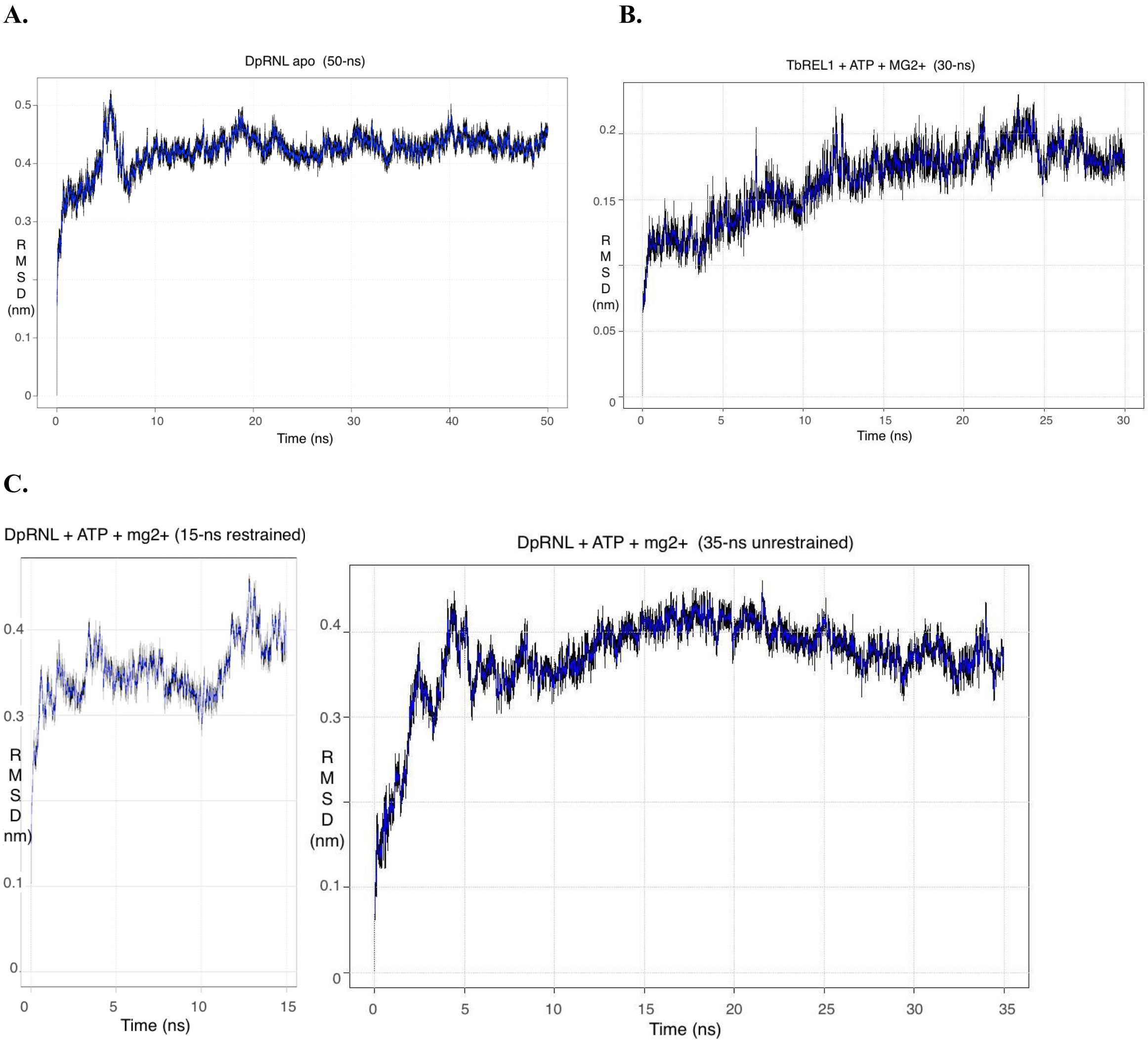


Figure S3. RMSD (Root Mean Square Deviation). RMSD-Backbone after least-square fit to backbone A. DpRNL apo backbone, 50 ns simulation, B. TbREL + ATP + Mg^{++}, 30 ns simulation, C. DpRNL + ATP + Mg^{++}, 50-ns (replicate 15R0+35_1) .

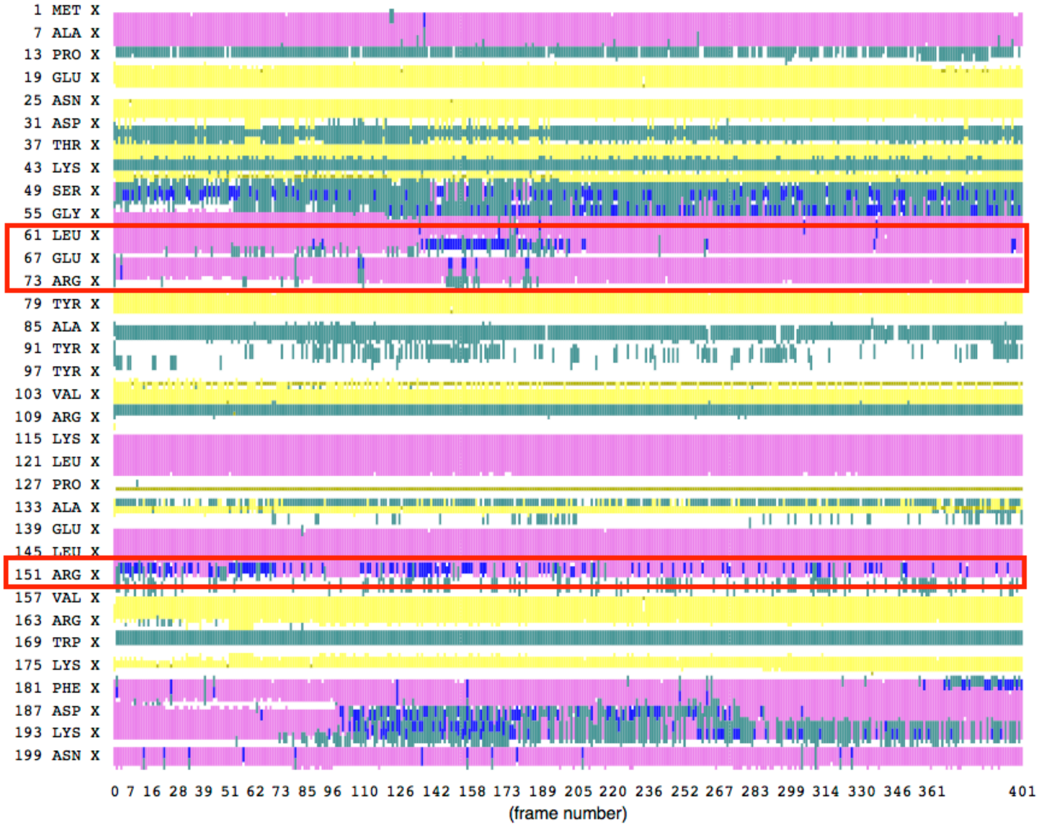
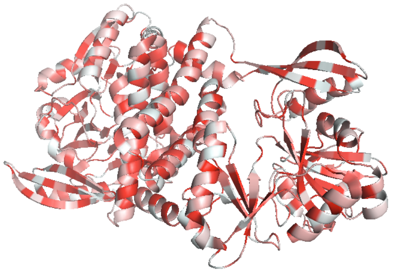
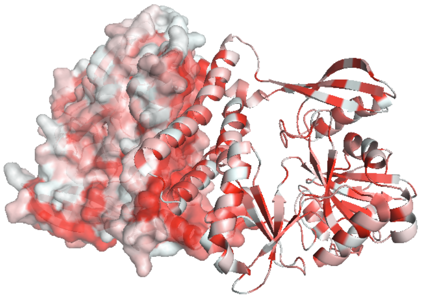


Figure S4. Define Secondary Structure of Proteins (DSSP) calculated for DpRNL (replicate 15R0+35_1) with the tool VMD. Each amino acid of the ligase (vertical) is assigned a secondary structure state for every frame of the simulation (horizontal). Pink, α helices; blue, 3-10 helices; yellow, β strands; green, turns; and white, other loops. Flexible helices are highlighted in red.

	Cartoon representation	Cartoon and surface
2VUG *Pyrococcus abyssi*		

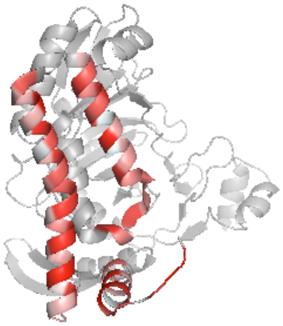
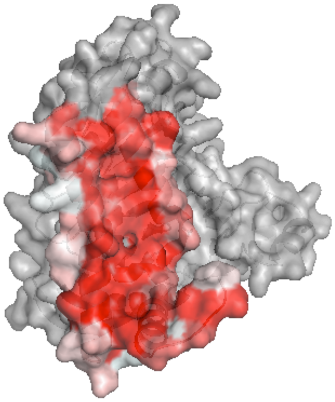
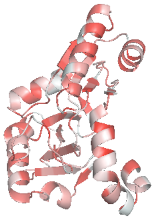
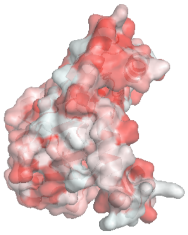
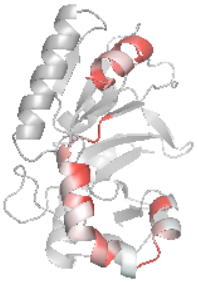
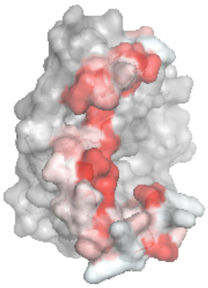
2VUG - Interaction helices *Pyrococcus abyssi* Monomer with only interaction helices highlighted		
DpRNL *Diplonema papillatum*		
DpRNL - C-terminal helices *Diplonema papillatum*		

Figure S5. Hydrophobicity of residues represented by different shades of red according to the Eisenberg hydrophobicity scale [344] **on *Pyrococcus abyssi* and *Diplonema* 3D representation.** The hydrophobic alpha helices allowing dimerisation are visible in the archaean structure whereas the corresponding region in DpRNL's model does not display such an hydrophobic surface.

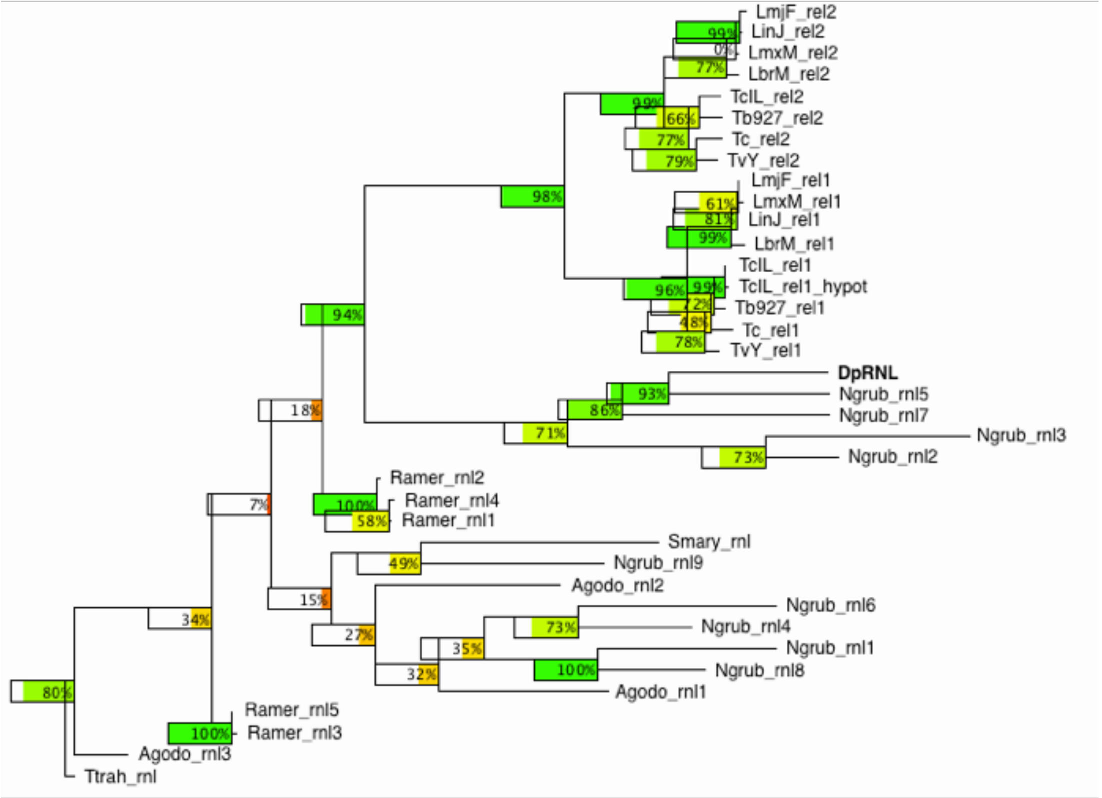


Figure S6. Phylogenetic tree of RNA ligases 2 from Excavata. *Leishmania major* strain Friedlin (LmjF, rel1: LmjF. 01.0590, rel2: LmjF.20.1730), *Leishmania infantum* strain JPCM5 (LinJ, rel1: LinJ.01.0610, rel2: Linj.20.1700), *Leishmania mexicana* strain MHOM/GT/2001/U1103 (LmxM, rel1: LmxM.01.0590, rel2: LmxM.20.1730), *Leishmania braziliensis* strain MHOM/BR/75/M2904 (LbrM, rel1: LbrM.01.0620, rel2: LbrM.20.5890), *Trypanosoma congolense* strain IL3000 (TcIL, rel1: TcIL3000_9_1420, rel1_hypot: TcIL3000_0_36090, rel2: TcIL3000_1_1450), *Trypanosoma brucei* strain TREU927 (Tb927, rel1: Tb927.9.4360, rel2: Tb927.1.3030), *Trypanosoma cruzi* strain CL Brener Non-Esmeraldo-like (Tc, rel1:TcCLB.510155.20, rel2: TcCLB.506975.9), *Trypanosoma vivax* strain Y486 (TvY, rel1: TvY486_0901490, rel2: TvY486_0101350), *Diplonema papillatum* (DpRNL), *Naegleria gruberi* (Ngrub, rnl1: XP_002669886.1, rnl2: XP_002670045.1, rnl3: XP_002672316.1, rnl4: XP_002672691.1, rnl5: XP_002674912.1, rnl6: XP_002676530.1, rnl7: XP_002678251.1, rnl8: XP_002679286.1, rnl9: XP_002679917.1), *Reclinomonas americana* (Ramer, rnl1: Contig155.33 [1608 - 145], rnl2: Contig383.16 [7789 - 6296], rnl3: Contig411.2 [830 - 2062], rnl4: Contig54.25 [1893 - 3302], rnl5: Contig7385.1 [4474 - 3215]), *Sawyeria*

marylandensis (Smary, rnl: SML00003242), *Andalucia godoyi* (Agodo, rnl1: comp2286_c0_seq2_1, rnl2: comp3909_c0_seq1_5, rnl3: comp8362_c0_seq3_5), *Thecamonas trahens* (Ttrah, rnl: AMSG_03775). RNA editing ligases are abbreviated as «rel» and hypothetical RNA ligases as «rnl».

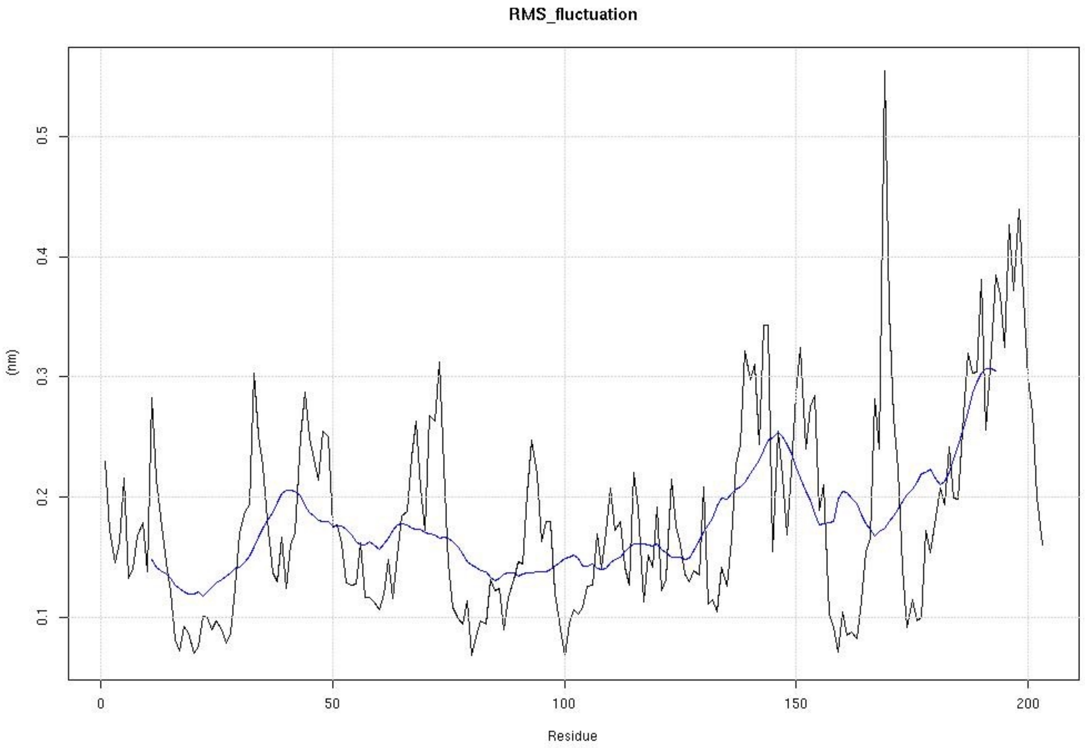
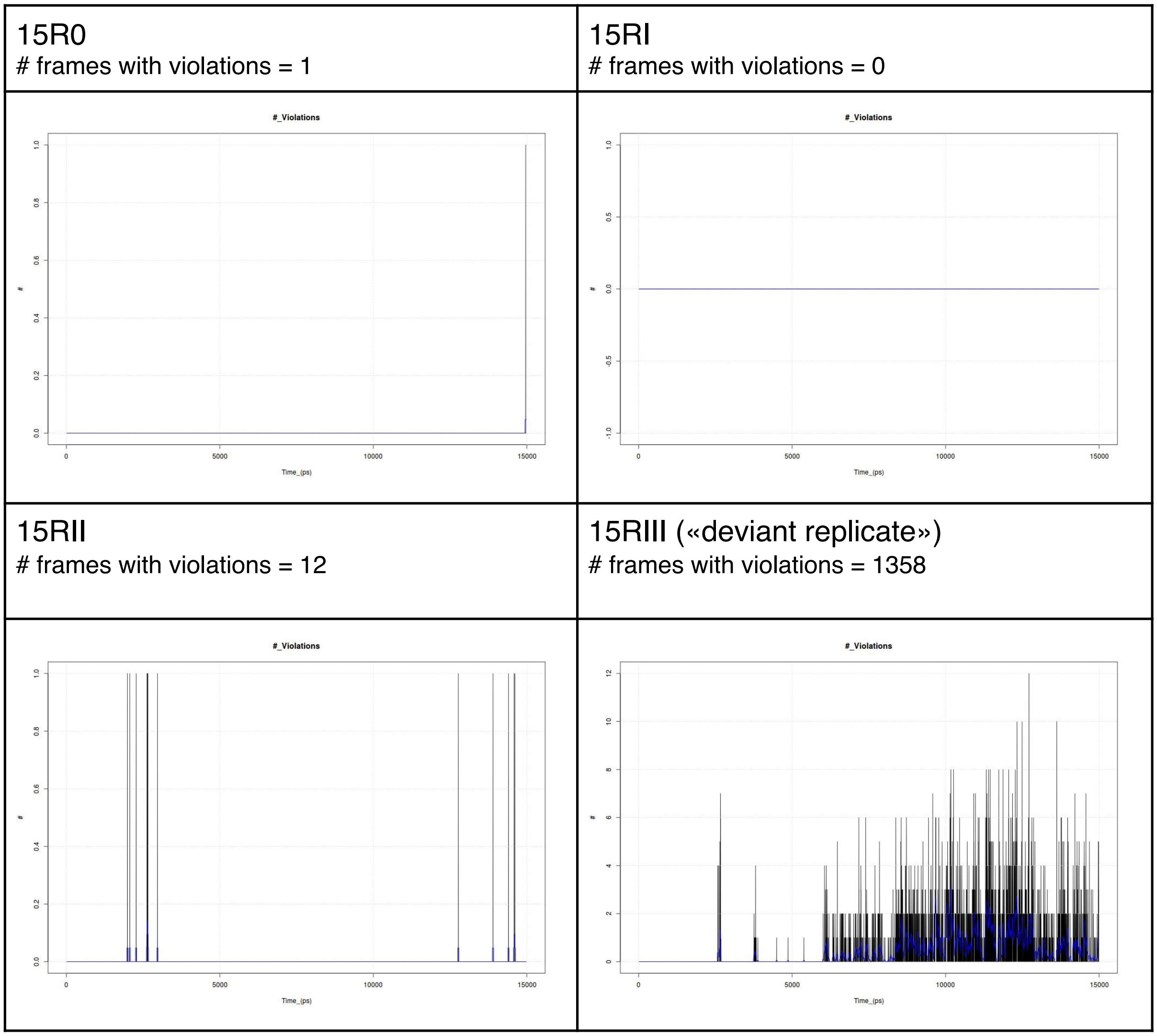


Figure S7. RMSF (Root Mean Square Fluctuation) of DpRNL apo backbone, 50-ns simulation. Blue line: smoothed average.



15R0
# frames with violations = 1

15RI
# frames with violations = 0

15RII
# frames with violations = 12

15RIII («deviant replicate»)
# frames with violations = 1358

Figure S8. Number of violations of distance restraints during the 15-ns of the four ATP-restrained simulation replicates. No force is applied to the system to restrain the distance between the centroid of the phenyl group of DpRNL-F101 and the pyrimidine group of ATP between the green line (r0) and the red line (r1). During this simulation, the restriction did not come into effect.

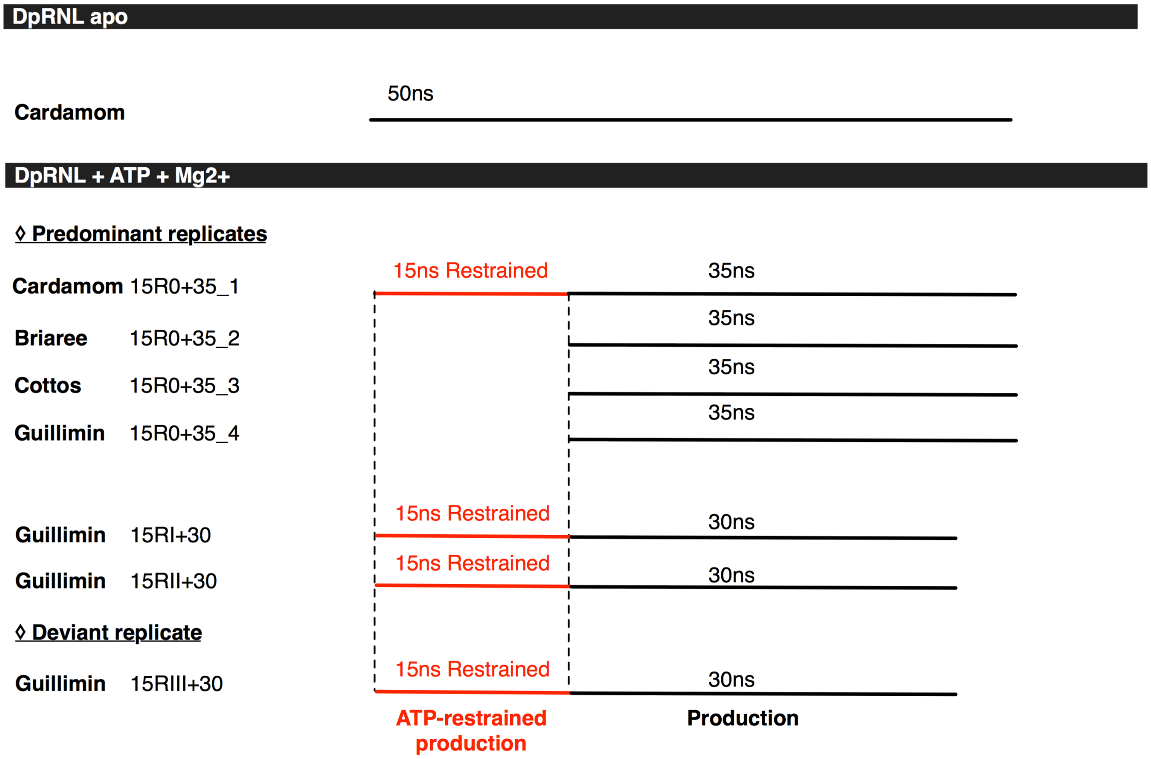


Figure S9. Design of MD simulations performed. A single MD simulation was performed on the apo enzyme. For DpRNL loaded with ATP and Mg2+, four independent 15-ns ATP-restrained simulation (in red, ATP-restrained production phase) were performed. Following the first (15R0-*) ATP-restrained simulation, 4 independent MD simulations (production phase) of 35 ns were run. For the other ATP-restrained replicates (15RI-*, 15RII-*, 15RIII- *) a single production MD simulation of 30 ns was performed. Cardamom, Briaree, Cottos, Guillimin designate the server used which corresponds to different versions of Gromacs: Gromacs 4.0.5 for Cottos, 4.6.5 for Cardamom, 5.0.1 for Briaree and 5.0.2 for Guillimin.

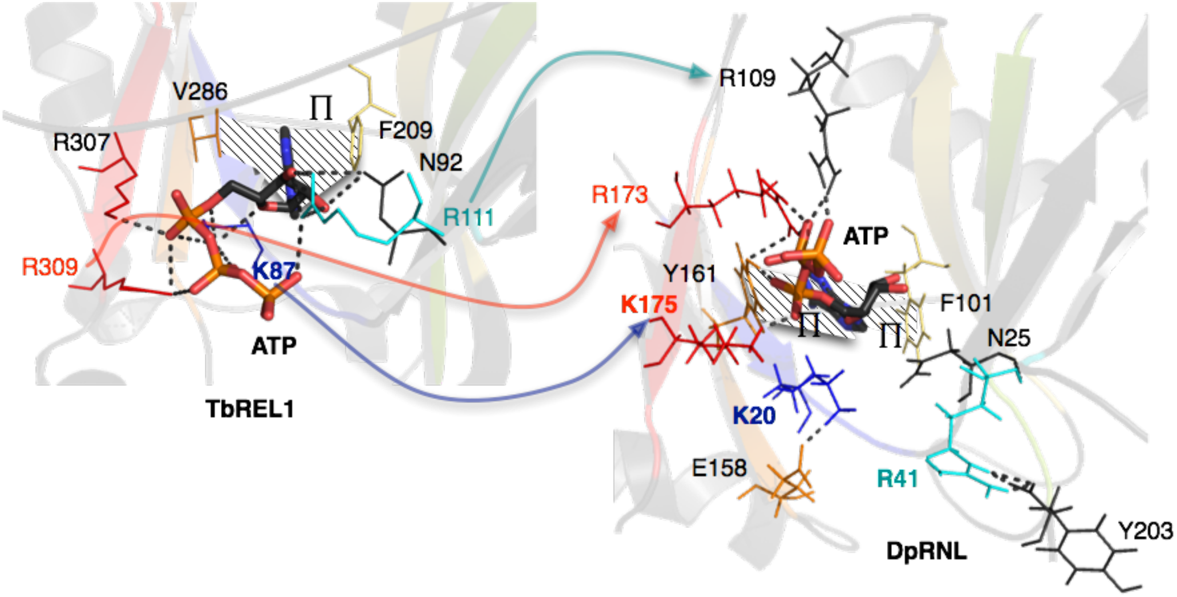


Figure S10. Proposed model for ATP interactions rewiring in DpRNL. Change in the function of a residue is depicted by an arrow. According to our predictions, the most peculiar feature of DpRNL is the reorganization of residue-residue and residue-ATP interactions in the catalytic pocket compared to other ligases. In fact, one of the rare residue in the ATP interaction network of DpRNL that is both structurally and functionally conserved appears to be DpRNL_F101, on which the adenine base is stacked canonically. Reshaping of DpRNL is thought to impact steps 1 and 2 of the catalysis (Supplementary Figure S1). We have shown that in the DpRNL model, the adenine of ATP is sandwiched between two aromatic residues (the highly conserved DpRNL_F101 and the 'exceptional' DpRNL_Y161) via double π-stacking, and that the orientation of the base in the catalytic pocket has been shifted as a result. This shift entails several consequences. First, the ribose is less firmly stabilized than in TbTEL1 and T4RNL2, by only two indirect interactions (with DpRNL-N25 and DpRNL-E81). Second, the conserved lysine DpRNL_K20 in motif I is contacted by H-bonds and salt bridges and pulled away from ATP, so that now, ATP could be contacted by the conserved lysine DpRNL_K175 in motif V which is not contacted by any other noncovalent interaction. This latter residue is therefore proposed to attack the α phosphorus of ATP and establish the covalent bound with AMP in reaction step 1 (blue arrow). Third, the triphosphate tail of ATP is stabilized by the exceptional residues DpRNL_R173 rather than by the amino acid in DpRNL position 177, which is usually basic (TbTEL1_R_309|T4RNL2_K227, red arrow) but a valine in the *Diplonema* protein. Finally, the conserved arginine DpRNL_R41, instead of forming H-bonds with the sugar and the polyphosphate tail of ATP as do the structural counterparts of this residue in other RNA ligases 2 (TbREL1-R111|T4RNL2_55, cyan arrow), interacts with the most C-terminal residue DpRNL_Y203, thus restraining the motion of the C-terminal domain. Note that in the structure of the full-length T4RNL2, interactions between the N-terminal and C-terminal domains have not been observed [1].

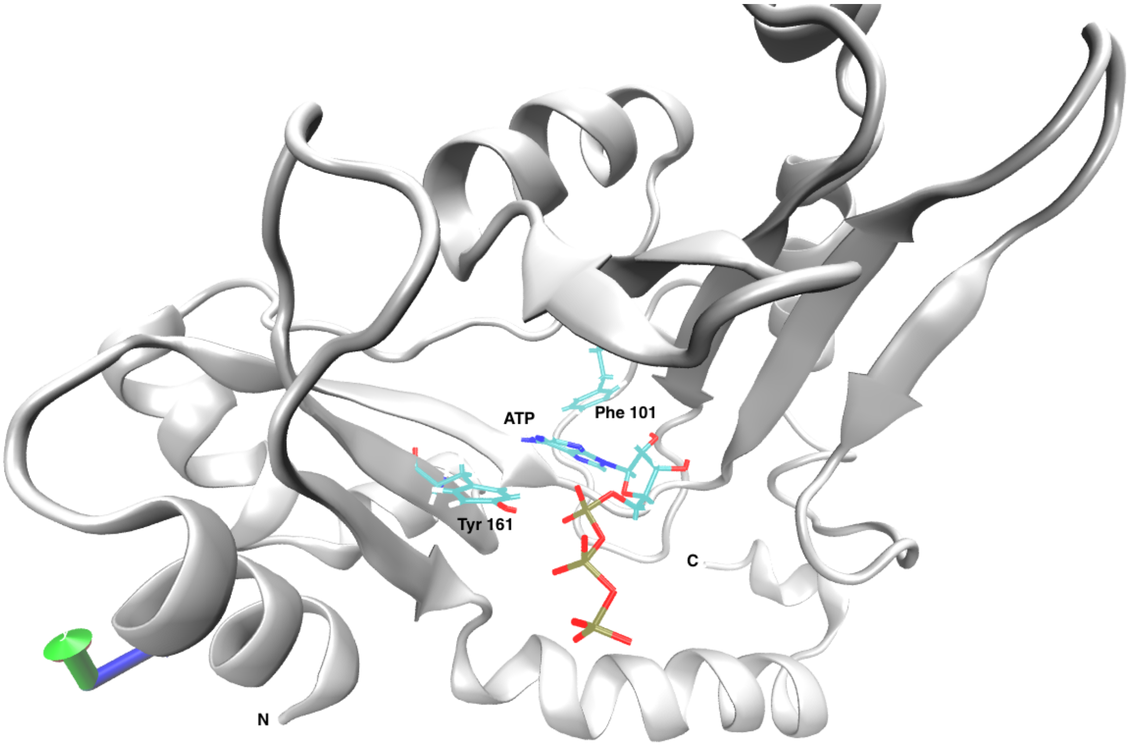


Figure S11. Catalytic pocket for the deviant MD simulation

RNA/DNA ligases		ITasser threading template (LOMETS)			ITasser (TM-align)	Refined DpRNL			
	Length	Ident. %	Cov. %	Z-score	TM-score	Aligned length	RMSD	Ident. %	TM-score norm. to DpRNL
3QWU - *Aquifex aeolicus* putative DNA ligase	366	19	96	3.13	0.878	189	2.19	19.0	**0.842**
2VUG - *Pyrococcus abyssi* RNA ligase homodimeric	373	-/-	-/-	-/-	0.854	187	2.40	21.4	0.823
1S68 - *Enterobacteriophage T4* 2-233 + AMP	233	21	90	1.72	0.751	175	3.21	18.9	0.681
1XDN - *Trypanosoma brucei* RNA ligase	262	24	89	**3.52**	0.745	176	3.38	19.3	0.678

Table S1. Comparison of DpRNL with close structural homologs of interest from the output of ITasser and by structural comparison of the refined model of DpRNL. When many threading templates for the same template were provided by ITasser, we selected the one with the highest Z-score. Ident. : percentage of identity in the aligned region, Cov: Coverage.

	Motif	Protein TbREL1	Mean (Std Dev)	Protein DpRNL	Mean (Std Dev) For all replicates	Mean (Std Dev) For deviant MD
ATP:PA	I	K87:NZ	4.04 (0.09)	K20:NZ	**7.7 (0.5)**	4.4 (0.8)
ATP:Adenine	IIIa	F209	3.87 (0.23)	F101	**4.1 (0.01)**	5.8 (0.3)
ATP:Adenine	IV		-	Y161	**4.1 (0.00)**	5.5 (0.4)
ATP:PA	V	K307:NZ	-	K175:NZ	**4.3 (0.01)**	3.8 (0.2)

Table S2. Comparison of ATP interaction with DpRNL and TbREL1. Mean distances between residues are reported in angstrom for the six replicates and the deviant replicate.

Chapitre 5: Résultats non publiés

Contributions : Sandrine Moreira a réalisé l'assemblage du génome, participé à toutes les annotations du génome, et supervisé les analyses du génome. Jean Coquet a analysé les répétitions dans le génome. Annie Lebreton a réalisé l'apprentissage des prédicteurs de gène et établi un modèle de gène pour *Diplonema*. Anzhelika Butenko (AB) et Pavel Flegontov (PF) ont assemblé des répétitions du génome. Matus Valach AB et PF ont annoté manuellement plusieurs contigs du génome afin de permettre l'apprentissage des prédicteurs de gène. Matthew Sarrasin a développé un pipeline d'annotation ayant permis de générer l'annotation finale du génome.

Les composants de la **machinerie de traitement post-transcriptionnel des ARNs mitochondriaux** sont très probablement encodés par le génome nucléaire. C'est le cas de la majorité des protéines du protéome mitochondrial, [62]. Afin d'identifier ces composants, le génome nucléaire de *D. papillatum* a été séquencé, puis je l'ai assemblé et annoté. Mes analyses de l'ARN ligase 2 (chapitre 4) sont basées sur l'annotation résultante. Nous préparons une publication du génome, avec nos collaborateurs. Ce chapitre résume mon travail sur cet aspect.

I.Assemblage et annotation du génome nucléaire

1. Assemblage

La procédure d'assemblage du génome nucléaire de *D. papillatum* est décrite dans la figure 19. Il s'agit de la stratégie qui a donné les meilleurs résultats, évalués selon le nombre de contigs, la taille médiane des contigs, et le nombre de transcrits s'alignant sur l'assemblage.

J'ai utilisé trois types de données: des lectures Illumina appariées avec chevauchement (MiSEQ, "Paired-ends overlap"), des lectures Illumina appariées avec de longs inserts (HiSEQ, "Mate-pairs"), et des très longues lectures PacBio. Un quatrième type de données, les lectures Roche 454, n'a pas été utilisé car il aboutissait systématiquement à un assemblage de moins bonne qualité (plus fragmenté et avec des contigs en moyenne plus courts).

Une étape cruciale pour l'assemblage des lectures a été le pré-traitement des données qui comprend l'excision des adapteurs de séquençage et des regions de basse qualité, le filtrage des séquences non-nucléaires, et la correction des erreurs de séquençage. Durant l'étape de filtrage, j'ai éliminé les séquences mitochondriales par alignement contre les séquences des chromosomes pour éviter de passer du temps de calcul sur ces données, et parce que j'ai constaté que l'assemblage d'un mélange de génomes aboutit à un assemblage de moins bonne qualité que leur traitement indépendant. Jusqu'à 75% des lectures (pour la librairie HiSEQ) ont été éliminées lors de cette étape.

La correction des séquences était indispensable pour les lectures PacBio réputées pour leur très haut taux d'erreur (11% ; valeur médiane selon http://www.pacb.com/uncategorized/a-closer-look-at-accuracy-in-pacbio/ datant du 22 janvier 2013). Pour notre librairie, le taux d'erreur observé est d'environ 7%).

Seules les librairies Illumina ont été utilisées pour l'assemblage, ce qui représente une couverture d'environ 39X. Les lectures PacBio ont été utilisées pour faire la jonction entre les *scaffolds* ou remplir les brèches.

L'assemblage final mesure environ 188 Mb, répartis en 4,305 *scaffolds*. Il est quasiment complet car 98.5% des transcrits reconstitués à partir de séquençage ARN s'alignent sur le génome.

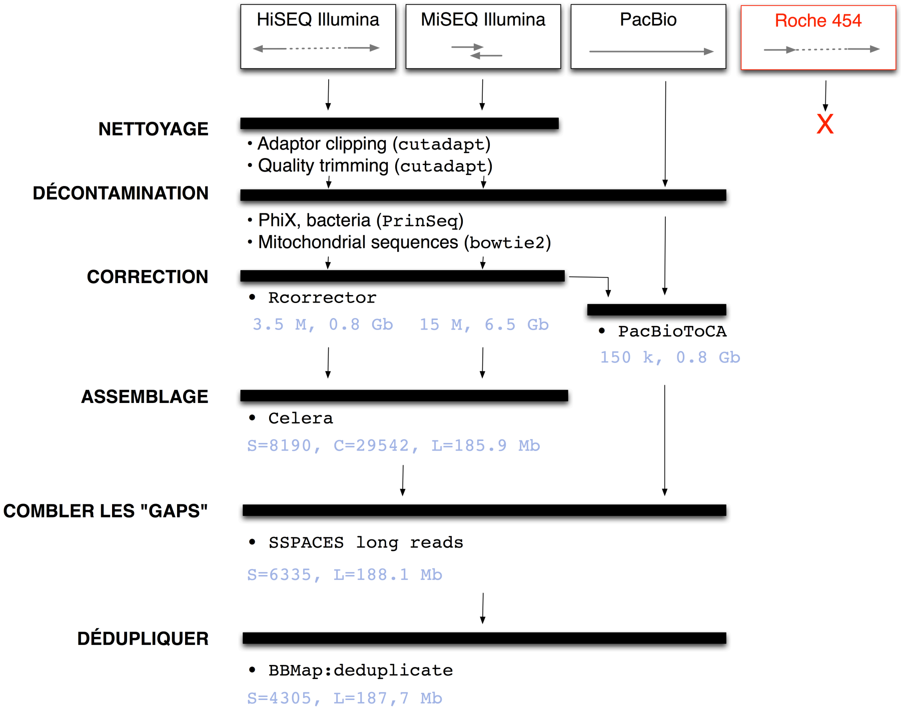


Figure 19. Procédure pour l'assemblage du génome nucléaire de *D. papillatum*. S, nombre de *scaffolds* ; C, nombre de contigs ; L, longueur.

2. Annotation

La procédure d'annotation du génome nucléaire de *D. papillatum* est décrite dans la figure 20. Lors de l'annotation syntaxique (qui détermine la localisation et la structure des gènes), la détection des gènes est réputée meilleure si le modèle utilisé est adapté à l'espèce. Dans le cas d'espèces divergentes comme *D. papillatum*, cette étape est d'autant plus importante que les autres indices comme les similarités avec les protéines d'autres espèces ou la présence de domaines protéiques sont rares. Nous avons donc annoté manuellement ~200 gènes choisis arbitrairement sur les contigs contenant les gènes détectés par le logiciel CEGMA [133]. Après élimination des gènes potentiellement incomplets aux extrémités des contigs et des gènes dont l'annotation était ambigüe, notre jeu d'apprentissage est constitué de 99 gènes. Nous disposons également de plusieurs librairies de séquençage ARN utilisées par les prédicteurs de gène. Malgré la grande couverture en lectures ARN qui pourrait sembler suffisante pour la prédiction des gènes, nos tests comparatifs ont montré de meilleurs résultats en utilisant les modèles entrainés sur les 99 gènes annotés manuellement. Notre pipeline d'annotation automatique a détecté 25 718 gènes (tableau 20). La densité en gène est d'environ 47,8% (90 Mb codant / 188 Mb total), ce qui est du même ordre de grandeur que ce qui est observé chez *Trypanosoma brucei* TREU927 par exemple (40,56% densité de codage, source http://tritrypdb.org/ accédé le 2 août 2016). Les gènes de *D. papillatum* sont en moyenne pauvre en intron avec 1,21 intron par gène.

L'annotation fonctionnelle nous a permis d'attribuer des indices sur la fonction des gènes (domaines protéiques ou similarité) pour presque la moitié des gènes.

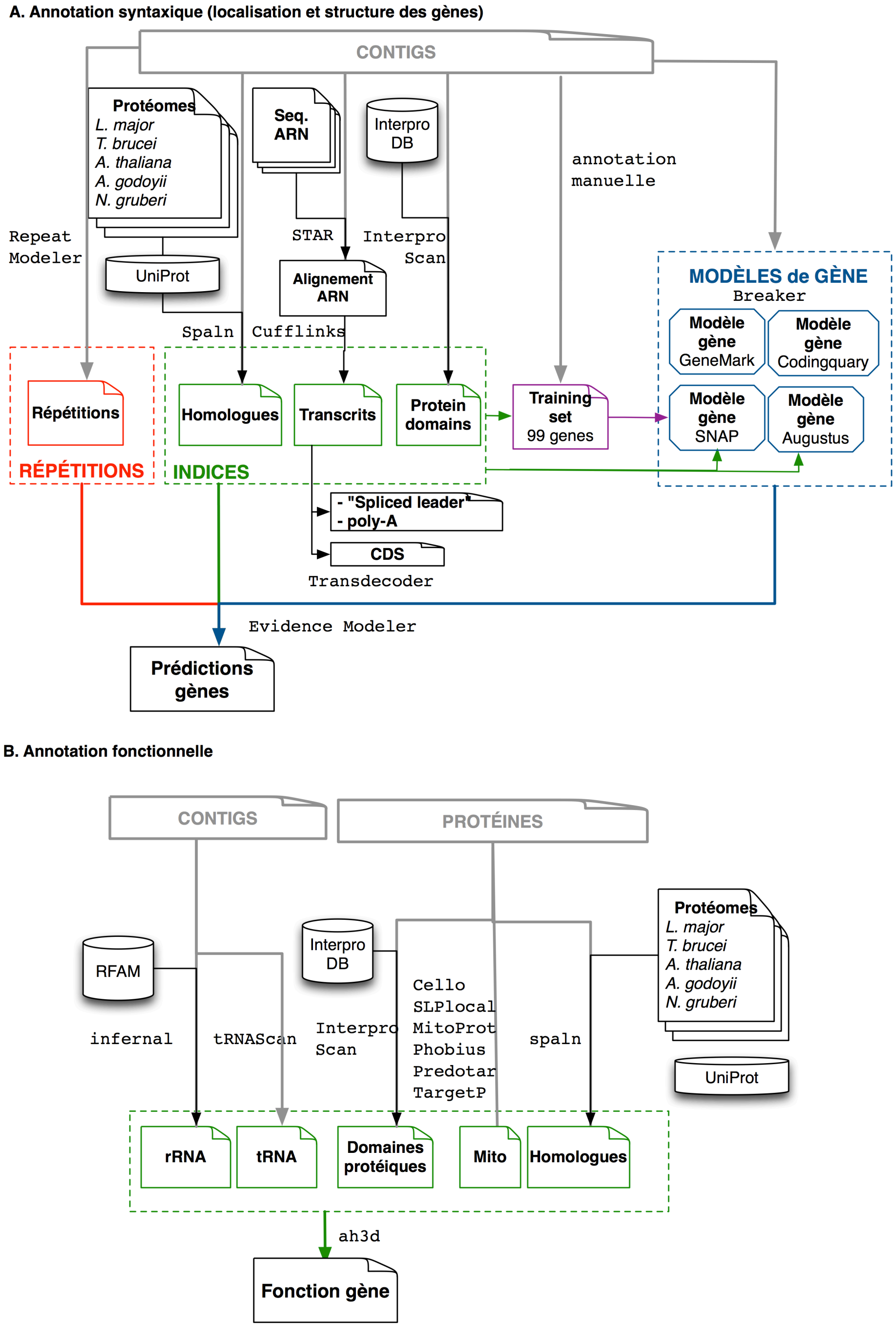


Figure 20. Procédure d'annotation du génome nucléaire de *D. papillatum*.

Tableau II. Statistiques des annotations syntaxiques et fonctionnelles de *D. papillatum*.

Annotation syntaxique	
Nombre de gènes	25 718
Nombre de transcrits	26 479
Taille moyenne transcrits (pb)	3 429,46
Nombre d'exons par transcrit	2,28
Nombre d'introns par transcrit	1,21
Annotation fonctionnelle	
Similarité banque uniprot (BLASTP)	4505 (17%)
Domaines protéiques PFAM	11 721 (44,3%)
Adressage mitochondrial	1981 (7.4%)

3. Recherche de gènes candidats

J'ai recherché des gènes candidats pour la machineries de traitement post-transcriptionnel des ARNs mitochondriaux en me basant sur les prédictions de domaines protéiques et d'adressage mitochondrial. Les gènes trouvés sont listés dans le tableau 3. Dans le partie Discussion, je détaillerai la motivation de recherche pour chacune des catégories et la pertinence des résultats obtenus en relation avec ce que l'on connait des processus dans les autres espèces.

Tableau III. Protéines candidates impliquées dans le traitement post-transcriptionnel des ARNs mitochondriaux.

Edition par substitution	**Nombre de gènes (localisation prédite dans la mitochondrie)**
A-deaminase (métabolique)	4(3)
ADAR/ADAT	1(0)
C-deaminase (métabolique)	6(1)
APOBEC	1(0)
PPR	73(36)
Édition par polyuridylation	
TUTase	5(1)
Épissage en trans ("trans-splicing")	
RNA ligase 2	3(2)
RNA ligase type RtcB	3(1)

Discussion

Le décryptage de l'information génétique par édition d'ARN et épissage en trans

L'objectif de ma thèse était d'élucider comment la mitochondrie d'un protiste marin peut produire des ARNs matures contigus et fonctionnels à partir de gènes en morceaux, éparpillés sur des chromosomes différents, et cryptés. Il s'agissait d'identifier les étapes de maturation, de comprendre les processus biochimiques en jeux, et d'en identifier les protagonistes moléculaires.

Durant ma thèse nous avons identifié une **nouvelle forme d'encryptage** des gènes mitochondriaux par édition de A-vers-I et C-vers-U. Le séquençage des ARNs à haut débit m'a permis d'identifier, de quantifier et pour certaines étapes, d'ordonner les **intermédiaires de maturation** et de mettre en évidence une importante parallélisation des processus post-transcriptionnels [345]. Nous avons également identifié les gènes codant pour l'ARN de la petite et de la grande sous-unité ribosomique. Cette dernière est non seulement fragmentée en deux modules distincts mais aussi éditée par **polyuridylation** à la jonction des deux fragments avant **épissage en trans**. Cette combinaison de processus de maturation est unique [227]. Enfin, j'ai exploré le contenu génomique nucléaire et identifié des candidats prometteurs des complexes de maturation des ARN mitochondriaux. Leur validation pourra utiliser une combinaison d'approches bio-informatiques [150] et expérimentales.

Beaucoup d'aspects pourraient être discutés dans un système aussi déviant que celui des diplonémides. J'ai choisi d'en aborder deux. Le premier concerne la relation entre la structure fragmentée des gènes et la structure des protéines. Pour reprendre l'analogie du crime de l'avant-propos, je pose l'hypothèse que les victimes (les gènes mitochondriaux des diplonémides) n'ont pas été frappées (fragmentés) au hasard mais sur des "points faibles", dont je discuterai. Je discuterai exclusivement le cas des gènes protéiques. Le second aspect est le "profilage" de suspects impliqués dans le décryptage de l'information génétique.

I.Relation structure-fonction: y a-t-il un lien entre la fragmentation des gènes, l'édition par polyuridylation et la structure des protéines ?

> *"Tout celà se résume d'une façon particulièrement frappante :* ***le plan de la cellule est dans le[s] chromosome[s]****. Les gènes ne sont pas distribués au hasard dans le texte génomique, mais leur position est en rapport avec leur mode d'expression selon la nature de l'environnement, et avec leur localisation dans les divers compartiment de la cellule." - Antoine Danchin, 1998* [346]

Fonction et structure sont intimement liées en génomique. Nous l'avons vu au niveau de la réplication du génome mitochondrial où les ADNs polydispersés sont indicatifs de la réplication par cercle roulant par exemple. De la même manière, l'organisation des gènes est un indice de leur mode d'expression. Qu'en est-il de l'organisation baroque des gènes mitochondriaux chez *Diplonema* ? Y aurait-il un lien entre la fragmentation des gènes, les modifications post-transcriptionnelles, en particulier l'édition par polyuridylation qui rajoute en quelque sorte des exons virtuels, et la structure des protéines ?

1. Les sites de fragmentation des gènes sont-ils aléatoires ?

Les gènes mitochondriaux de *D. papillatum* sont intensément fragmentés. Jusqu'à 11 modules composent le gène *nad5*. Nous avons proposés le modèle évolutif selon lequel des éléments génétiques mobiles se seraient insérés dans le génome mitochondrial, éventuellement entre les gènes. La recombinaison entre ces éléments aurait aboutit à la structure multi-chromosomique actuelle [347].

On peut se demander si ce processus est toujours actif et si les génomes des diplonémides sont encore en perpétuel réarrangement. Par comparaison des gènes *cox1* et *nad4* chez trois autres diplonémides nous avons observé au contraire, que le nombre de modules et les zones de coupures sont conservés malgré de grandes distances évolutives entre ces espèces. Cette conservation des modules, donc de la fragmentation génique, pourrait être le signe d'une recombinaison peu active. Il est possible que l'insertion d'éléments génétiques dans un ancêtre des diplonémides ait été suivie d'une période de stabilité génétique. Ceci peut être testé par comparaison de la structure des génomes de différents diplonémides. Nos premières analyses chez *D. ambulator* montrent plutôt une architecture chromosomique très variable avec jusqu'à huit modules présents sur le même chromosome (non publié). L'observation de modules très conservés malgré une architecture très plastique est un indice d'une forte pression de sélection sur le placement des jonctions entre modules.

Il est possible que la fragmentation des gènes ne soit possible que dans des régions particulières: dans des séquences présentant un biais de composition ou dans des motifs spécifiques par exemple. Pour la protéine Nad4, on remarque que les sites de jonction se situent préférentiellement entre les hélices trans-membranaires. Pour un ARN parfaitement épissé, il est difficile d'imaginer en quoi la position de la jonction dans la structure protéique pourrait être une contrainte pour l'insertion de l'élément génétique dans le gène. Or, si la machinerie d'épissage est imparfaite, en excisant de manière imprécise les régions flanquantes par exemple, elle défavoriserait les jonctions dans les régions structurellement ou fonctionnellement importantes. En ce sens, les boucles entre les hélices trans-membranaires sont de très bon candidats pour la fragmentation. On s'attendrait également à un mécanisme pour compenser cette excision imparfaite qui pourrait décaler le cadre de lecture. L'ajout de nucléotides à l'extrémité des modules, telle que la (poly)uridylation, pourrait remplir ce rôle.

2. La polyuridylation est liée à la fragmentation des gènes

Les évênements d'édition par polyuridylation se produisent exclusivement à l'extrémité 3' de certains modules. Par rapport au transcrit, l'édition survient entre deux modules (8 cas) ou en 3' du dernier module (11 cas). Comme nous l'avons discuté dans le paragraphe précédent, la (poly)uridylation pourrait compenser un épissage imparfait. L'ajout d'uridines pourrait également être une stratégie de la cellule pour compenser la perte de modules, un évènement qui pourrait arriver dans un scénario d'insertions d'éléments génétiques où de courtes portions de gènes pourraient être perdus, (voir figure 4 dans l'appendice p.199) [347].

La machinerie d'uridylation peut agir par synthèse d'un nombre exact d'uridines guidée par un gabarit tel qu'un ARN anti-sens par exemple. Dans nos analyses de séquençages brin-spécifiques de petits ARNs, nous n'en avons cependant pas trouvé. Toutefois, il est possible que des ARNs anti-sens existent, mais que des modifications chimiques de leurs extrémités empêchent leur séquençage. Il est aussi possible que le guide soit plutôt une protéine. Alternativement, la machinerie d'édition par polyuridylation peut synthétiser une longue queue poly-U puis couper le transcrit jusqu'à la taille requise. Des systèmes de "mètres moléculaires" ("molecular ruler") existent dans le monde ARN. Par exemple les enzymes Drosha et Dicer impliquées dans le mécanisme d'interférence de l'ARN coupent l'ARN double brin en fragments de 11 et 25 pb [348], [349]. Un autre exemple est l'endonucléase impliquée dans l'épissage des introns de type ARNt qui l'utilise pour identifier le site de coupure [350]. Jusqu'à présent, les exemples de la polyuridylation de *nad4* (2 uridines) et *cox1* (6 uridines) ont montré chez les diplonémides une stricte conservation du nombre de 'U'. Il serait intéressant d'élargir l'analyse aux autres protéines mitochondriales, en particulier des gènes inconnus '*y*' qui portent de longues insertions d'uridines. Une variation du nombre d'uridines corrélée à la perte d'acides aminés aux jonctions corroborerait l'hypothèse de compensation de taille et du "mètre moléculaire".

Ce phénomène de compensation de taille par édition est d'ailleurs observé chez les dinoflagellés. Chez certains membres de ce groupe, le gène *cox3* est scindé en deux cistrons transcrits indépendemment, puis épissés en trans. À la jonction des deux fragments, l'épissage retient plusieurs 'A' de la queue poly-A du premier module [83]. De façon intéressante, le nombre de A retenus est tel que le cadre de lecture et la taille du produit final sont rétablis.

3. La polyuridylation des transcrits ne peut survenir que dans des régions correspondant à des zones spécifiques de la protéines

La polyuridylation aboutit majoritairement à l'ajout de codons 'UUU' codant pour la phénylalanine, un acide aminé aromatique et hydrophobe. Dans les autres cas, les uridines complètent des codons. Six gènes protéiques sont édités par uridylation (*cox1*, *nad4*, *y2*, *y3*, *y4*, *y5*) et jusqu'à 29 uridines peuvent être ajoutés entre les modules de gènes. Étant données les propriétés physico-chimiques des phénylalanines, leur ajout n'est probablement possible qu'à la jonction de certains modules.

Pour étudier l'éventuel biais d'insertion de 'U' dans les protéines, il est important de se rappeler que le génome mitochondrial encode essentiellement des protéines membranaires, donc avec de larges domaines hydrophobes. Tel est le cas des protéines de la chaîne respiratoire. Il a même été postulé que la rétention des gènes dans le génome mitochondrial était en partie dictée par leur hydrophobicité. Dans les protéines trans-membranaires, la phenylalanine est préférentiellement trouvée dans les hélices membranaires, souvent dans les motifs FxxGxxG. En interagissant entre elles, les phénylalanines stabilisent les hélices intra-protéines [351] ou permettent la dimérisation de protéines [352] grace aux propriété d'empilement des cycles aromatiques [353]. Cependant, les homopolymères de phenylalanine ont plutôt été montré comme cytotoxiques [354] car ils provoquent l'aggrégation des protéines [355]. Les dimers de phenylalanine du peptide amyloid beta sont d'ailleurs responsables de la formation de fibrilles dans la maladie d'Alzheimer [356]. On peut donc supposer que **la**

localisation d'homopolymères de phénylalanine dans la protéine est soumise à une forte pression de sélection et survient dans des régions privilégiées de la protéine. Il serait intéressant de résoudre la structure des protéines dont les ARNs sont abondamment polyuridylés, en particulier les protéines "Y", et d'élucider le rôle fonctionnel ou structural de ces homopolymères.

Dans les organites des autres eucaryotes, la polyuridylation survient dans un contexte très différent de ce qu'on observe chez *Diplonema*. La polyuridylation consiste en l'ajout ponctuel de une ou deux uridines, rarement un codon entier chez les kinétoplastides, ou chez les *slime molds* comme *Physarum*. Dans les plastides des dinoflagellés [116], [357], l'ajout d'une queue poly-U ne change pas la région codante car elle survient après le codon stop. Dans les mitochondries des dinoflagellés, 8 uridines sont ajoutés en 5' des deux gènes (*cox1* et le gène fusionnné *cob-cox3*) mais sont en amont de la phase ouverte de lecture [36].

Le cas du gène *cox3* des dinoflagellés évoqué précédemment se rapproche de ce qu'on observe chez les diplonémides avec la rétention de plusieurs adénines à la jonction des deux fragments du gène [83]. La coupure entre les deux exons se situe entre les hélices 6 et 7, comme la jonction entre les modules 2 et 3 du gène *cox3* chez *Diplonema*. La polyadénylation entre les modules rajoute des codons 'AAA' codant pour la lysine, un acide aminé composé d'une courte chaîne à 3 carbones terminée par un groupe ammonium chargé positivement. Les lysines sont fréquemment trouvées en bordure des hélices transmembranaires et se positionnent comme un périscope (*snorkel*) où la chaîne aliphatique est enfoncée dans la membrane alors que le groupe chargé pointe à la surface. La position du poly-A entre les hélices 6 et 7 de *cox3* chez les dinoflagellés apparait donc biochimiquement favorable. On voit que les contraintes physico-chimiques pour la lysine sont très différentes de celles pour la phénylalanine. Ainsi, la situation chez les dinoflagellés est mécanistiquement très similaire à ce qu'on observe chez les diplonémides mais chimiquement différente.

Il est possible que les machineries qui effectuent l'uridylation/adénylation couplée à l'épissage en trans soient similaires entre les deux groupes, ce qui serait un example supplémentaire de **convergence évolutive** entre les dinoflagellés et les Euglenozoa [358]. Comme chez les diplonémides, l'épissage en trans chez les dinoflagellés ne présente aucun des signaux d'épissage connus. Un autre point commun avec les diplonémides est l'abondance de sites d'édition d'ARN par substitution. Chez les dinoflagellés, ils ne sont cependant pas groupés et présentent tous les types de substitution possibles sauf 3 bien que A-vers-G et C-vers-U soient les plus fréquents. Ces deux espèces sont très éloignées phylogénétiquement et le trans-splicing couplé à la polyuridylation/adénylation n'étant connu que pour ces deux espèces, il est plus probable que la machinerie ait émergée indépendament plutot qu'héritée d'un ancêtre commun et perdue dans les autres taxons. Alternativement, la machinerie aurait pu être transférée horizontalement entre ces deux espèces, ce qui serait possible car ils partagent les mêmes écosystèmes (tous deux sont des composants du plancton) [165].

II.Des candidats pour le traitement post-transcriptionnels des ARNs mitochondriaux

La machinerie de traitement post-transcriptionnels des ARNs mitochondriaux des diplonémides est inédite car aucun signal en *cis* ni acteur en *trans* caractéristiques des machineries d'épissage connues n'a été trouvé. En particulier, je n'ai pas détecté d'ARN anti-sens qui permettraient de guider l'épissage en trans, ou de spécifier les sites d'édition comme c'est le cas chez le groupes frère des diplonémides, les kinétoplastides. Afin d'en identifier les composants dans le génome nucléaire, j'ai utilisé plusieurs approches bio-informatiques que j'ai basées sur les hypothèses suivantes. Au niveau **cellulaire**, nous savons que ces protéines doivent être adressées à la mitochondrie. **Biochimiquement**, plusieurs de ces protéines doivent posséder des domaines d'interaction avec l'ARN. **Enzymatiquement**, certaines réactions connues pourraient catalyser les réactions que nous observons. **Fonctionnellement**, des protéines impliquées dans des processus similaires sont caractérisées chez d'autres organismes et leurs homologues pourraient être à l'oeuvre chez *Diplonema*. **Evolutivement**, les complexes mitochondriaux impliqués dans l'édition et l'expression des gènes des kinétoplastides et ceux des diplonémides ont pu évoluer à partir de même machineries ancestrales et partagent probablement des homologues.

L'approche purement bio-informatique décrite au chapitre 5 m'a permis de faire le tri dans les milliers de protéines du génome, et d'identifier des candidats impliqués dans **(i) l'édition par polyuridylation**, **(ii) l'édition par substitution** et **(iii) l'épissage en trans**. Le tableau II, chapitre 5 présente la liste des candidats.

1. Candidats pour l'édition par polyuridylation

Le génome nucléaire de *Diplonema* contient plusieurs gènes codant pour des protéines de la famille des Terminal-Uridyl-Transferases (TUTases) dont il existe aussi des représentants chez les kinetoplastides. Les **TUTases** de la famille des poly-A polymérases (PAPs) sont des enzymes capables d'ajouter plusieurs uridines en 3' des ARNs. Chez les kinétoplastides, deux classes de TUTases sont fonctionnellement intéressantes: la TUTase de l'**éditosome** qui rajoute des uridines ponctuellement aux sites d'édition et celle du **3' processome** qui rajoute une queue poly-U en 3' des ARNs guides et un nombre limité d'uridines aux ARN ribosomiques (11 uridines pour l'ARNr 9S et entre 2 et 17 pour l'ARNr 12S) [115], [270], [359]. Il est important de souligner que la TUTase utilisée pour les ARNr et les ARN guides est différente de celle des ARNm. Chez *Diplonema*, il est possible que les TUTases, si elles catalysent effectivement la polyuridylation, aient la même spécificité de substrat. Une analyse phylogénétique des TUTases identifiées chez *Diplonema* et celles des kinetoplastides nous permettrait de mieux comprendre l'histoire évolutive de ces enzymes et de formuler des hypothèses sur leur fonction. Cependant les distances évolutives très grandes entre ces espèces peuvent gêner les reconstructions phylogénétiques en raison du phénomène d'attraction des longues branches [311]. J'en ai fait l'expérience lors de l'analyse phylogénétique des ARN ligase 2 (voir le paragraphe "candidats pour l'épissage en trans")[150]. Toutefois, les TUTases sont des protéines plus longues que les ARN ligases 2, multi-géniques, et avec plusieurs domaines bien conservés. Ces caractéristiques ajoutent du signal pour l'alignement des séquences et faciliteraient la reconstruction phylogénétique.

La TUTase du 3' processosome se distingue des autres TUTases car il existe plusieurs similitudes entre l'expression des ARNs guides et celles des gènes des diplonémides. Comme chez *Diplonema*, les mini-cercles sur lesquels se trouvent les gènes des ARNs guides sont transcrits dans les deux directions, les régions flanquantes des transcrits sont excisées et les ARNs guides sont polyuridylés en 3'. Etant données ces similarités des processus, il est possible que les homologues des autres composants du 3' Processosome jouent un rôle dans la maturation des transcrits mitochondriaux chez *Diplonema*. L' exonucléase DSS1 qui excise les régions flanquantes des ARNs guides, par exemple, a un homologue chez *D. papillatum* dont le rôle pourrait être similaire.

2. Candidats pour l'édition par substitution C-vers-U, et A-vers-G

Chez *Diplonema*, la machinerie d'édition par substitution pourrait impliquer des homologues des déaminases du métabolisme des bases azotées.

Nous avons trouvé des homologues des enzymes du **métabolisme des purines et des pyrimidines** adressées à la mitochondrie dont trois adénosines déaminases et une cytidine déaminase. Ces enzymes métaboliques ont pu évoluer chez *Diplonema* pour prendre en charge l'édition. Nous avons également identifié des homologues de **ADAR** et **APOBEC** (voir partie I.2) mais les logiciels de prédictions d'adressage subcellulaire ne prédisent pas qu'ils sont localisés dans la mitochondrie. Toutefois, ces logiciels requièrent une annotation précise de la portion N-terminale de la protéine dont je n'ai pu faire qu'une prédiction basée sur les données de séquençage ARN. Par ailleurs, la recherche d'une séquence typique d'adressage en position N-terminale ou d'une composition caractéristique de la séquence sont basés sur des espèces modèle éloignées de *Diplonema*, ce qui rend la prédiction moins fiable.

Le génome nucléaire de *Diplonema* contient également 96 **protéines pentatricopeptides (PPR)** dont une protéine du type DYW. Leur présence est surprenante car ces protéines ne sont trouvées que sporadiquement en dehors des plantes. Il est remarquable que les protistes dont la machinerie d'édition est inconnue présentent également un nombre élevé de protéines de la famille des PPR (plus de 800 chez le dinoflagellé *Alexandrium tamarenses*, 116 chez l'Amoebozoa *Physarum polycephalum* et 40 chez l'Heterolobosea *Naegleria gruberi*) [360]. L'hypothèse de l'implication des protéines PPR dans l'édition chez ces eucaryotes [111], [361] reste à démontrer expérimentalement.

Quelle est l'origine évolutive des protéines PPR chez *Diplonema* ? Elles ont plus probablement été héritées verticalement de l'ancêtre des kinétoplastides + diplonémides que transférées horizontalement. Les kinetoplastides possèdent 28 PPRs dont deux interagissent avec la TUTase impliquée dans la polyuridylation des ARNr mitochondriaux [362]. Il est donc probable que les PPR chez ces deux groupes aient une origine commune. Certains euglénides ont un chloroplaste acquis par endosymbiose secondaire. Il est peu probable qu'il soit à l'origine des PPR observés chez les diplonémides et les kinétoplastides car son acquisition s'est faite secondairement à la divergence des euglénides et de la branche des diplonémides + kinétoplastides [363], [364]. Une analyse phylogénétique des protéines PPR serait nécessaire pour tester si les PPR ont une origine plus ancienne et auraient été massivement perdues chez de nombreuses espèces.

3. Candidats pour l'épissage en trans

Le génome nucléaire de *Diplonema* contient plusieurs ARN ligases qui, enzymatiquement, pourraient catalyser l'épissage en trans des ARNs mitochondriaux. Nous avons trouvé des protéines de la famille des ARN ligases 2 dont font partie celles de l'éditosome des kinetoplastides [150], [289]. Nous avons également identifié des ARN ligases de type RtcB, une famille récemment découverte et impliquée dans l'épissage des introns d'ARNt chez les animaux, les bactéries et les archées. Ces ligases ont la particularité de catalyser la ligation entre des extrémités 5'-OH et 3'P ou 2',3'-phosphate cyclique.

La validation bio-informatique par analyse tridimensionnelle de l'une des ARN ligase 2 de *D. papillatum* a permis de conclure qu'elle ne partage ni certaines caractéristiques structurales, ni un ancêtre commun avec la ligase des kinétoplastides. Toutefois, ceci n'exclut pas sa participation dans le trans-splicing qui serait à valider expérimentalement.

Conclusion

Nous avons vu dans le chapitre d'introduction et par mon travail de thèse que les processus cellulaires chez les protistes peuvent être très éloignés des processus canoniques décrits dans les manuels scolaires. Ceci reflète le biais des efforts de recherche vers certains groupes taxonomiques (métazoaire, champignons, plantes terrestres) ; une myopie taxonomique que les nouvelles technologies de séquençage permettent en partie de corriger. Nul doute donc, que l'avenir nous réserve encore de surprenantes découvertes biochimique chez les protistes, élargissant notre vision de l'évolution des nanomachines cellulaires. Par ailleurs, ces innovations recèleront certainement un énorme potentiel d'applications biotechnologiques, à l'image de la machinerie de traitement des ARNs mitochondriaux des diplonémides capable de recoudre des morceaux épars d'ARNs avec une précision surprenante.

Références

[1] J. Archibald, *One plus one equals one: symbiosis and the evolution of complex life*. 2014.

[2] G. Burger, M. W. Gray, and B. Franz Lang, "Mitochondrial genomes: anything goes," *Trends in Genetics*, vol. 19, no. 12, pp. 709–716, Dec. 2003.

[3] M. W. Gray, D. Sankoff, and R. J. Cedergren, "On the evolutionary descent of organisms and organelles: a global phylogeny based on a highly conserved structural core in small subunit ribosomal RNA," *Nucleic Acids Res*, vol. 12, no. 14, pp. 5837–5852, 1984.

[4] R. Cedergren, M. W. Gray, Y. Abel, and D. Sankoff, "The evolutionary relationships among known life forms," *Journal of Molecular Evolution*, vol. 28, no. 1, pp. 98–112, Dec. 1988.

[5] S. G. E. Andersson, A. Zomorodipour, J. O. Andersson, T. Sicheritz-Pontén, U. C. M. Alsmark, R. M. Podowski, A. K. Näslund, A.-S. Eriksson, H. H. Winkler, and C. G. Kurland, "The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria," *Nature*, vol. 396, no. 6707, pp. 133–140, Nov. 1998.

[6] A. H. Knoll, "Paleobiological Perspectives on Early Eukaryotic Evolution," *Cold Spring Harbor Perspectives in Biology*, vol. 6, no. 1, pp. a016121–a016121, Jan. 2014.

[7] B. F. Lang, M. W. Gray, and G. Burger, "Mitochondrial Genome Evolution and the Origin of Eukaryotes," *http://dx.doi.org/10.1146/annurev.genet.33.1.351*, vol. 33, no. 1, pp. 351–397, Nov. 1999.

[8] W. F. Martin, S. Garg, and V. Zimorski, "Endosymbiotic theories for eukaryote origin," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 370, no. 1678, pp. 20140330–18, Aug. 2015.

[9] P. López-García and D. Moreira, "Open Questions on the Origin of Eukaryotes," *Trends in Ecology & Evolution*, vol. 30, no. 11, pp. 697–708, Nov. 2015.

[10] M. W. Gray, "Mitochondrial Evolution," *Cold Spring Harbor Perspectives in Biology*, vol. 4, no. 9, pp. a011403–a011403, Sep. 2012.

[11] S. M. Adl, A. G. B. Simpson, C. E. Lane, J. Lukeš, D. Bass, S. S. Bowser, M. W. Brown, F. Burki, M. Dunthorn, V. Hampl, A. Heiss, M. Hoppenrath, E. Lara, L. le Gall, D. H. Lynn, H. McManus, E. A. D. Mitchell, S. E. Mozley-Stanridge, L. W. Parfrey, J. Pawlowski, S. Rueckert, L. Shadwick, C. L. Schoch, A. Smirnov, and F. W. Spiegel, "The Revised Classification of Eukaryotes," *J. Eukaryot. Microbiol.*, vol. 59, no. 5, pp. 429–514, Sep. 2012.

[12] A. M. Shiflett and P. J. Johnson, "Mitochondrion-Related Organelles in Eukaryotic Protists," *Annu. Rev. Microbiol.*, vol. 64, no. 1, pp. 409–429, Oct. 2010.

[13] M. van der Giezen and J. Tovar, "Degenerate mitochondria," *EMBO Rep*, vol. 6, no. 6, pp. 525–530, Jun. 2005.

[14] J. H. P. Hackstein, J. Tjaden, and M. Huynen, "Mitochondria, hydrogenosomes and mitosomes: products of evolutionary tinkering!," *Curr Genet*, vol. 50, no. 4, pp. 225–245, Aug. 2006.

[15] A. Karnkowska, V. Vacek, Z. Zubáčová, S. C. Treitli, R. Petrželková, L. Eme, L. Novák, V. Žárský, L. D. Barlow, E. K. Herman, P. Soukal, M. Hroudová, P. Doležal, C. W. Stairs, A. J. Roger, M. Eliáš, J. B. Dacks, Č. Vlček, and V. Hampl, "A Eukaryote without a Mitochondrial Organelle," *Current Biology*, vol. 26, no. 10, pp. 1274–1284, May 2016.

[16] M. W. Gray, "Mitochondrial genome diversity and the evolution of mitochondrial DNA," *Can. J. Biochem.*, vol. 60, no. 3, pp. 157–171, Mar. 1982.

[17] A. J. Bendich, "The size and form of chromosomes are constant in the nucleus, but highly variable in bacteria, mitochondria and chloroplasts," *Bioessays*, vol. 29, no. 5, pp. 474–483, 2007.

[18] A. J. Bendich, "Reaching for the ring: the study of mitochondrial genome structure," *Curr Genet*, vol. 24, no. 4, pp. 279–290, 1993.

[19] B. F. Lang, E. SEIF, M. W. Gray, C. J. O'KELLY, and G. Burger, "A Comparative Genomics Approach to the Evolution of Eukaryotes and their Mitochondria," *Journal of Eukaryotic Microbiology*, vol. 46, no. 4, pp. 320–326, Jul. 1999.

[20] J. L. Boore, "Animal mitochondrial genomes.," *Nucleic Acids Res*, vol. 27, no. 8, pp. 1767–1780, Apr. 1999.

[21] R. Shao, E. F. Kirkness, and S. C. Barker, "The single mitochondrial chromosome typical of animals has evolved into 18 minichromosomes in the human body louse, *Pediculus humanus*," *Genome Res.*, vol. 19, no. 5, pp. 904–912, May 2009.

[22] R. Shao, X. Q. Zhu, S. C. Barker, and K. Herd, "Evolution of Extensively Fragmented Mitochondrial Genomes in the Lice of Humans," *Genome Biol Evol*, vol. 4, no. 11, pp. 1088–1101, Nov. 2012.

[23] T. Gibson, V. C. Blok, M. S. Phillips, G. Hong, D. Kumarasinghe, I. T. Riley, and M. Dowton, "The Mitochondrial Subgenomes of the Nematode *Globodera pallida* Are Mosaics: Evidence of Recombination in an Animal Mitochondrial Genome," *Journal of Molecular Evolution*, vol. 64, no. 4, pp. 463–471, Mar. 2007.

[24] E. Kayal and D. V. Lavrov, "The mitochondrial genome of *Hydra oligactis* (Cnidaria, Hydrozoa) sheds new light on animal mtDNA evolution and cnidarian phylogeny," *Gene*, vol. 410, no. 1, pp. 177–186, Feb. 2008.

[25] D. Williamson, "The curious history of yeast mitochondrial DNA," *Nat Rev Genet*, 2002.

[26] Y. Suyama and K. Miura, "Size and structural variations of mitochondrial DNA," *Proc Natl Acad Sci USA*, 1968.

[27] J. M. Gerhold, A. Aun, T. Sedman, P. Jõers, and J. Sedman, "Strand Invasion Structures in the Inverted Repeat of *Candida albicans* Mitochondrial DNA Reveal a Role for Homologous Recombination in Replication," *Molecular Cell*, vol. 39, no. 6, pp. 851–861, Sep. 2010.

[28] D. Fricova, M. Valach, Z. Farkas, I. Pfeiffer, J. Kucsera, L. Tomaska, and J. Nosek, "The mitochondrial genome of the pathogenic yeast *Candida subhashii*: GC-rich linear DNA with a protein covalently attached to the 5' termini," *Microbiology*, vol. 156, no. 7, pp. 2153–2163, Jul. 2010.

[29] M. Valach, Z. Farkas, D. Fricova, J. Kovac, B. Brejova, T. Vinar, I. Pfeiffer, J. Kucsera, L. Tomaska, B. F. Lang, and J. Nosek, "Evolution of linear chromosomes and multipartite genomes in yeast mitochondria," *Nucleic Acids Res*, vol. 39, no. 10, pp. 4202–4219, May 2011.

[30] D. J. Oldenburg and A. J. Bendich, "Mitochondrial DNA from the liverwort *Marchantia polymorpha*: circularly permuted linear molecules, head-to-tail concatemers, and a 5′ protein11Edited by N.-M. Chua," *Journal of Molecular Biology*, vol. 310, no. 3, pp. 549–562, Jul. 2001.

[31] D. B. Sloan, A. J. Alverson, J. P. Chuckalovcak, M. Wu, D. E. McCauley, J. D. Palmer, and D. R. Taylor, "Rapid Evolution of Enormous, Multichromosomal Genomes in Flowering Plant Mitochondria with Exceptionally High Mutation Rates," *PLoS Biol*, vol. 10, no. 1, pp. e1001241–17, Jan. 2012.

[32] J. M. Gualberto, D. Mileshina, C. Wallet, A. K. Niazi, F. Weber-Lotfi, and A. Dietrich, "The plant mitochondrial genome: Dynamics and maintenance," *Biochimie*, vol. 100, pp. 107–120, May 2014.

[33] J. E. Feagin, "The 6-kb element of *Plasmodium falciparum* encodes mitochondrial cytochrome genes," *Molecular & Biochemical Parasitology*, vol. 52, no. 1, pp. 145–148, May 1992.

[34] J. Pawlowski, S. Audic, S. Adl, D. Bass, L. Belbahri, C. Berney, S. S. BOWSER, I. Cepicka, J. Decelle, M. Dunthorn, A. M. Fiore-Donno, G. H. Gile, M. Holzmann, R. Jahn, M. Jirků, P. J. Keeling, M. Kostka, A. Kudryavtsev, E. Lara, J. Lukeš, D. G. MANN, E. A. D. Mitchell, F. Nitsche, M. Romeralo, G. W. Saunders, A. G. B. Simpson, A. V. SMIRNOV, J. L. Spouge, R. F. Stern, T. Stoeck, J. Zimmermann, D. Schindel, and C. de Vargas, "CBOL Protist Working Group: Barcoding Eukaryotic Richness beyond the Animal, Plant, and Fungal Kingdoms," *PLoS Biol*, vol. 10, no. 11, p. e1001419, 2012.

[35] M. J. Caley, R. Fisher, and K. Mengersen, "Global species richness estimates have not converged," *Trends in Ecology & Evolution*, vol. 29, no. 4, pp. 187–188, Apr. 2014.

[36] C. H. Slamovits, J. F. Saldarriaga, A. Larocque, and P. J. Keeling, "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*, vol. 372, no. 2, pp. 356–368, Sep. 2007.

[37] E. A. Nash, R. E. R. Nisbet, A. C. Barbrook, and C. J. Howe, "Dinoflagellates: a mitochondrial genome all at sea," *Trends in Genetics*, vol. 24, no. 7, pp. 328–335, Jul. 2008.

[38] R. F. Waller and C. J. Jackson, "Dinoflagellate mitochondrial genomes: stretching the rules of molecular biology," *BioEssays*, vol. 31, no. 2, pp. 237–245, Feb. 2009.

[39] D. Faktorová, E. Dobáková, P. Peña-Diaz, and J. Lukeš, "From simple to supercomplex: mitochondrial genomes of euglenozoan protists," *F1000Res*, vol. 5, pp. 392–9, Mar. 2016.

[40] D. F. Spencer and M. W. Gray, "Ribosomal RNA genes in *Euglena gracilis* mitochondrial DNA: fragmented genes in a seemingly fragmented genome," *Mol Genet Genomics*, vol. 285, no. 1, pp. 19–31, 2011.

[41] E. Dobáková, P. Flegontov, T. Skalický, and J. Lukeš, "Unexpectedly Streamlined Mitochondrial Genome of the Euglenozoan *Euglena gracilis*," *Genome Biol Evol*, vol. 7, no. 12, pp. 3358–3367, Dec. 2015.

[42] J. Lukes, D. Lys Guilbride, J. Votypka, A. Zikova, R. Benne, and P. T. Englund, "Kinetoplast DNA Network: Evolution of an Improbable Structure," *Eukaryotic Cell*, vol. 1, no. 4, pp. 495–502, Aug. 2002.

[43] W. Marande, J. Lukes, and G. Burger, "Unique Mitochondrial Genome Structure in Diplonemids, the Sister Group of Kinetoplastids," *Eukaryotic Cell*, vol. 4, no. 6, pp. 1137–1146, Jun. 2005.

[44] W. Marande and G. Burger, "Mitochondrial DNA as a Genomic Jigsaw Puzzle," *Science*, vol. 318, no. 5849, pp. 415–415, Oct. 2007.

[45] Č. Vlček, W. Marande, S. Teijeiro, J. Lukeš, and G. Burger, "Systematically fragmented genes in a multipartite mitochondrial genome," *Nucleic Acids Res*, vol. 39, no. 3, pp. gkq883–988, Oct. 2010.

[46] M. W. Gray, G. Burger, and B. F. Lang, "Mitochondrial Evolution," *Science*, vol. 283, no. 5407, pp. 1476–1481, Mar. 1999.

[47] K. Adams, "Evolution of mitochondrial gene content: gene loss and transfer to the nucleus," *Molecular Phylogenetics and Evolution*, vol. 29, no. 3, pp. 380–395, Dec. 2003.

[48] B. F. Lang, G. Burger, C. J. O'Kelly, R. Cedergren, G. B. Golding, C. Lemieux, D. Sankoff, M. Turmel, and M. W. Gray, "An ancestral mitochondrial DNA resembling a eubacterial genome in miniature," , *Published online: 29 May 1997; | doi:10.1038/387493a0*, vol. 387, no. 6632, pp. 493–497, May 1997.

[49] G. Burger, M. W. Gray, L. Forget, and B. F. Lang, "Strikingly Bacteria-Like and Gene-Rich Mitochondrial Genomes throughout Jakobid Protists," *Genome Biol Evol*, vol. 5, no. 2, pp. 418–438, Feb. 2013.

[50] J.-W. Taanman, "The mitochondrial genome: structure, transcription, translation and replication," *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, vol. 1410, no. 2, pp. 103–123, Feb. 1999.

[51] X. Perez-Martinez, A. Antaramian, M. Vazquez-Acevedo, S. Funes, E. Tolkunova, J. d'Alayer, M. G. Claros, E. Davidson, M. P. King, and D. González-Halphen, "Subunit II of cytochrome c oxidase in Chlamydomonad algae is a heterodimer encoded by two independent nuclear genes," *Journal of Biological Chemistry*, vol. 276, no. 14, pp. 11302–11309, Apr. 2001.

[52] R. F. Waller and P. J. Keeling, "Alveolate and chlorophycean mitochondrial cox2 genes split twice independently," *Gene*, vol. 383, pp. 33–37, Nov. 2006.

[53] G. Burger, Y. Zhu, T. G. Littlejohn, S. J. Greenwood, M. N. Schnare, B. F. Lang, and M. W. Gray, "Complete sequence of the mitochondrial genome of *Tetrahymena pyriformis* and comparison with *Paramecium aurelia* mitochondrial DNA," *Journal of Molecular Biology*, vol. 297, no. 2, pp. 365–380, Mar. 2000.

[54] J. P. Gogarten, A. G. Senejani, O. Zhaxybayeva, L. Olendzenski, and E. Hilario, "Inteins: Structure, Function, and Evolution," *Annu. Rev. Microbiol.*, vol. 56, no. 1, pp. 263–287, Oct. 2002.

[55] K. Hikosaka, K. Kita, and K. Tanabe, "Diversity of mitochondrial genome structure in the phylum Apicomplexa," *Molecular & Biochemical Parasitology*, vol. 188, no. 1, pp. 26–33, Mar. 2013.

[56] E. C. Swart, M. Nowacki, J. Shum, H. Stiles, B. P. Higgins, T. G. Doak, K. Schotanus, V. J. Magrini, P. Minx, E. R. Mardis, and L. F. Landweber, "The *Oxytricha trifallax* Mitochondrial Genome," *Genome Biol Evol*, vol. 4, no. 2, pp. 136–154, Feb. 2012.

[57] P. H. Boer and M. W. Gray, "Scrambled ribosomal RNA gene pieces in *Chlamydomonas reinhardtii* mitochondrial DNA," *Cell*, vol. 55, no. 3, pp. 399–411, Nov. 1988.

[58] K. M. Lonergan and M. W. Gray, "Expression of a Continuous Open Reading Frame Encoding Subunits 1 and 2 of Cytochrome c Oxidase in the Mitochondrial DNA of *Acanthamoeba castellanii*," *Journal of Molecular Biology*, vol. 257, no. 5, pp. 1019–1030, Apr. 1996.

[59] M. W. Gray, "Mosaic nature of the mitochondrial proteome: Implications for the origin and evolution of mitochondria," *Proc Natl Acad Sci USA*, vol. 112, no. 33, pp. 10133–10138, Aug. 2015.

[60] D. G. S. Smith, R. M. R. Gawryluk, D. F. Spencer, R. E. Pearlman, K. W. M. Siu, and M. W. Gray, "Exploring the mitochondrial proteome of the ciliate protozoon Tetrahymena thermophila: direct analysis by tandem mass spectrometry.," *Journal of Molecular Biology*, vol. 374, no. 3, pp. 837–863, Nov. 2007.

[61] S. E. Calvo, K. R. Clauser, and V. K. Mootha, "MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins," *Nucleic Acids Res*, vol. 44, no. 1, pp. D1251–D1257, Jan. 2016.

[62] R. M. R. Gawryluk, K. A. Chisholm, D. M. Pinto, and M. W. Gray, "Compositional complexity of the mitochondrial proteome of a unicellular eukaryote (*Acanthamoeba castellanii*, supergroup Amoebozoa) rivals that of animals, fungi, and plants," *Journal of Proteomics*, vol. 109, no. C, pp. 400–416, Sep. 2014.

[63] C. P. Lee, N. L. Taylor, and A. H. Millar, "Recent Advances in the Composition and Heterogeneity of the *Arabidopsis* Mitochondrial Proteome," *Front. Plant Sci.*, vol. 4, Jan. 2013.

[64] J. Reinders, R. P. Zahedi, N. Pfanner, C. Meisinger, and A. Sickmann, "Toward the Complete Yeast Mitochondrial Proteome: Multidimensional Separation Techniques for Mitochondrial Proteomics," *J. Proteome Res.*, vol. 5, no. 7, pp. 1543–1554, Jul. 2006.

[65] A. K. Panigrahi, Y. Ogata, A. Zíková, A. Anupama, R. A. Dalley, N. Acestor, P. J. Myler, and K. D. Stuart, "A comprehensive analysis of *Trypanosoma brucei* mitochondrial proteome," *Proteomics*, vol. 9, no. 2, pp. 434–450, Jan. 2009.

[66] T. Gabaldón and M. A. Huynen, "From Endosymbiont to Host-Controlled Organelle: The Hijacking of Mitochondrial Protein Synthesis and Metabolism," *PLoS Comput Biol*, vol. 3, no. 11, pp. e219–10, 2007.

[67] I. A. Stanley, S. M. Ribeiro, A. Giménez-Cassina, E. Norberg, and N. N. Danial, "Changing appetites: the adaptive advantages of fuel choice," *Trends in Cell Biology*, vol. 24, no. 2, pp. 118–127, Feb. 2014.

[68] R. Rizzuto, D. De Stefani, A. Raffaello, and C. Mammucari, "Mitochondria as sensors and regulators of calcium signalling," *Nature Reviews Molecular Cell Biology*, vol. 13, no. 9, pp. 566–578, Sep. 2012.

[69] D. A. Clayton, "Transcription and replication of mitochondrial DNA," *Hum. Reprod.*, vol. 15, no. 2, pp. 11–17, Jul. 2000.

[70] K. Hammani and P. Giegé, "RNA metabolism in plant mitochondria," *Trends in Plant Science*, vol. 19, no. 6, pp. 380–389, Jun. 2014.

[71] B. Schäfer, "RNA maturation in mitochondria of *S. cerevisiae* and *S. pombe*," *Gene*, vol. 354, pp. 80–85, Jul. 2005.

[72] A. P. Deshpande and S. S. Patel, "Mechanism of transcription initiation by the yeast mitochondrial RNA polymerase," *BBA - Gene Regulatory Mechanisms*, vol. 1819, no. 9, pp. 930–938, Sep. 2012.

[73] K. Liere and T. Börner, "Transcription in Plant Mitochondria," in *Plant Mitochondria*, no. 4, New York, NY: Springer New York, 2011, pp. 85–105.

[74] P. Le, P. R. Fisher, and C. Barth, "Transcription of the *Dictyostelium discoideum* mitochondrial genome occurs from a single initiation site," *RNA*, vol. 15, no. 12, pp. 2321–2330, Nov. 2009.

[75] N. Cermakian, T. M. Ikeda, P. Miramontes, B. F. Lang, M. W. Gray, and R. Cedergren, "On the Evolution of the Single-Subunit RNA Polymerases," *Journal of Molecular Evolution*, vol. 45, no. 6, pp. 671–681, Dec. 1997.

[76] D. Ojala, J. Montoya, and G. Attardi, "tRNA punctuation model of RNA processing in human mitochondria," *, Published online: 09 April 1981; | doi:10.1038/290470a0*, vol. 290, no. 5806, pp. 470–474, Apr. 1981.

[77] J. Asin-Cayuela and C. M. Gustafsson, "Mitochondrial transcription and its regulation in mammalian cells," *Trends in Biochemical Sciences*, vol. 32, no. 3, pp. 111–117, Mar. 2007.

[78] W. Rossmanith, "Of P and Z: mitochondrial tRNA processing enzymes.," *Biochim. Biophys. Acta*, vol. 1819, no. 9, pp. 1017–1026, Sep. 2012.

[79] B. Hoffmann, J. Nickel, F. Speer, and B. Schafer, "The 3′ Ends of Mature Transcripts Are Generated by a Processosome Complex in Fission Yeast Mitochondria," *Journal of Molecular Biology*, vol. 377, no. 4, pp. 1024–1037, Apr. 2008.

[80] E. M. Turk, V. Das, R. D. Seibert, and E. D. Andrulis, "The Mitochondrial RNA Landscape of *Saccharomyces cerevisiae*," *PLoS ONE*, vol. 8, no. 10, pp. e78105–21, Oct. 2013.

[81] S. Moreira, S. Breton, and G. Burger, "Unscrambling genetic information at the RNA level," *WIREs RNA*, vol. 3, no. 2, pp. 213–228, Jan. 2012.

[82] E. L. Lasda and T. Blumenthal, "Trans-splicing," *WIREs RNA*, vol. 2, no. 3, pp. 417–434, Jan. 2011.

[83] C. J. Jackson and R. F. Waller, "A Widespread and Unusual RNA Trans-Splicing Type in Dinoflagellate Mitochondria," *PLoS ONE*, vol. 8, no. 2, p. e56777, 2013.

[84] R. Benne, J. Van Den Burg, J. P. Brakenhoff, P. Sloof, J. H. Van Boom, and M. C. Tromp, "Major transcript of the frameshifted *coxII* gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA.," *Cell*, vol. 46, no. 6, pp. 819–826, Sep. 1986.

[85] S. C. Garrett and J. J. C. Rosenthal, "A Role for A-to-I RNA Editing in Temperature Adaptation," *Physiology*, vol. 27, no. 6, pp. 362–369, Dec. 2012.

[86] H. Liu, Q. Wang, Y. He, L. Chen, C. Hao, C. Jiang, Y. Li, Y. Dai, Z. Kang, and J.-R. Xu, "Genome-wide A-to-I RNA editing in fungi independent of ADAR enzymes," *Genome Res.*, vol. 26, no. 4, pp. 499–509, Apr. 2016.

[87] D. R. Smith, "The past, present and future of mitochondrial genomics: have we sequenced enough mtDNAs?," *Briefings in Functional Genomics*, vol. 15, no. 1, pp. 47–54, Jan. 2016.

[88] F. Juhling, M. Morl, R. K. Hartmann, M. Sprinzl, P. F. Stadler, and J. Putz, "tRNAdb 2009: compilation of tRNA sequences and tRNA genes," *Nucleic Acids Res*, vol. 37, no. Database, pp. D159–D162, Jan. 2009.

[89] B. El Yacoubi, M. Bailly, and V. de Crécy-Lagard, "Biosynthesis and Function of Posttranscriptional Modifications of Transfer RNAs," *Annu. Rev. Genet.*, vol. 46, no. 1, pp. 69–95, Dec. 2012.

[90] Z. Paris, I. M. C. Fleming, and J. D. Alfonzo, "Determinants of tRNA editing and modification: Avoiding conundrums, affecting function," *Seminars in Cell and Developmental Biology*, vol. 23, no. 3, pp. 269–274, May 2012.

[91] A. K. Hopper, "Transfer RNA Post-Transcriptional Processing, Turnover, and Subcellular Dynamics in the Yeast *Saccharomyces cerevisiae*," *Genetics*, vol. 194, no. 1, pp. 43–67, Apr. 2013.

[92] A. P. Gerber and W. Keller, "An Adenosine Deaminase that Generates Inosine at the Wobble Position of tRNAs," *Science*, vol. 286, no. 5442, pp. 1146–1149, Nov. 1999.

[93] W. A. Decatur and M. J. Fournier, "rRNA modifications and ribosome function," *Trends in Biochemical Sciences*, vol. 27, no. 7, pp. 344–351, Jul. 2002.

[94] A.-L. Chateigner-Boutin and I. Small, "Organellar RNA editing," *WIREs RNA*, vol. 2, no. 4, pp. 493–506, Jan. 2011.

[95] C. Schmitzlinneweber and I. Small, "Pentatricopeptide repeat proteins: a socket set for organelle gene expression," *Trends in Plant Science*, vol. 13, no. 12, pp. 663–670, Dec. 2008.

[96] S. Cheng, B. Gutmann, X. Zhong, Y. Ye, M. F. Fisher, F. Bai, I. Castleden, Y. Song, B. Song, J. Huang, X. Liu, X. Xu, B. L. Lim, C. S. Bond, S.-M. Yiu, and I. Small, "Redefining the structural motifs that determine RNA binding and RNA editing by pentatricopeptide repeat proteins in land plants," *Plant J*, vol. 85, no. 4, pp. 532–547, Feb. 2016.

[97] P. Yin, Q. Li, C. Yan, Y. Liu, J. Liu, F. Yu, Z. Wang, J. Long, J. He, H.-W. Wang, J. Wang, J.-K. Zhu, Y. Shi, and N. Yan, "Structural basis for the modular recognition of single-stranded RNA by PPR proteins," *Nature*, vol. 504, no. 7478, pp. 168–171, Nov. 2013.

[98] E. A. Nash, A. C. Barbrook, R. K. Edwards-Stuart, K. Bernhardt, C. J. Howe, and R. E. R. Nisbet, "Organization of the Mitochondrial Genome in the Dinoflagellate *Amphidinium carterae*," *Molecular Biology and Evolution*, vol. 24, no. 7, pp. 1528–1536, Mar. 2007.

[99] S. Hajduk and T. Ochsenreiter, "RNA editing in kinetoplastids," *RNA Biology*, vol. 7, no. 2, pp. 229–236, Mar. 2010.

[100] Y. W. Cheng, L. M. V. Robic, and J. M. Gott, "Non-templated addition of nucleotides to the 3′ end of nascent RNA during RNA editing in *Physarum*," *EMBO J*, vol. 20, no. 6, pp. 1405–1414, Mar. 2001.

[101] R. Bundschuh, J. Altmuller, C. Becker, P. Nurnberg, and J. M. Gott, "Complete characterization of the edited transcriptome of the mitochondrion of *Physarum polycephalum* using deep sequencing of RNA," *Nucleic Acids Res*, vol. 39, no. 14, pp. 6044–6055, Aug. 2011.

[102] L. M. Powell, S. C. Wallis, R. J. Pease, Y. H. Edwards, T. J. Knott, and J. Scott, "A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine," *Cell*, vol. 50, no. 6, pp. 831–840, Sep. 1987.

[103] B. Teng, C. F. Burant, and N. O. Davidson, "Molecular cloning of an apolipoprotein B messenger RNA editing protein," *Science*, vol. 260, no. 5115, pp. 1816–1819, Jun. 1993.

[104] B. A. Knisbacher, D. Gerber, and E. Y. Levanon, "DNA Editing by APOBECs: A Genomic Preserver and Transformer," *Trends in Genetics*, vol. 32, no. 1, pp. 16–28, Jan. 2016.

[105] S. G. Conticello, "The AID/APOBEC family of nucleic acid mutators," *Genome Biol.*, vol. 9, no. 6, p. 1, Jun. 2008.

[106] L. F. Grice, "The origin of the ADAR gene family and animal RNA editing," pp. 1–7, Feb. 2015.

[107] J. L. Casey, "Control of ADAR1 Editing of Hepatitis Delta Virus RNAs," in *Adenosine Deaminases Acting on RNA (ADARs) and A-to-I Editing*, vol. 353, no. 146, Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 123–143.

[108] A. Sanchez, S. G. Trappier, B. W. Mahy, C. J. Peters, and S. T. Nichol, "The virion glycoproteins of Ebola viruses are encoded in two reading frames and are expressed through transcriptional editing.," *Proc Natl Acad Sci USA*, vol. 93, no. 8, pp. 3602–3607, Apr. 1996.

[109] S. Hausmann, D. Garcin, C. Delenda, and D. Kolakofsky, "The versatility of paramyxovirus RNA polymerase stuttering.," *Journal of Virology*, vol. 73, no. 7, pp. 5568–5576, Jul. 1999.

[110] M. Mehedi, D. Falzarano, J. Seebach, X. Hu, M. S. Carpenter, H. J. Schnittler, and H. Feldmann, "A New Ebola Virus Nonstructural Glycoprotein Expressed through RNA Editing," *Journal of Virology*, vol. 85, no. 11, pp. 5406–5414, May 2011.

[111] M. Rudinger, L. Fritz-Laylin, M. Polsakiewicz, and V. Knoop, "Plant-type mitochondrial RNA editing in the protist *Naegleria gruberi*," *RNA*, vol. 17, no. 12, pp. 2058–2062, Nov. 2011.

[112] D. Gagliardi, P. P. Stepien, R. J. Temperley, R. N. Lightowlers, and Z. M. A. Chrzanowska-Lightowlers, "Messenger RNA stability in mitochondria: different means to an end," *Trends in Genetics*, vol. 20, no. 6, pp. 260–267, Jun. 2004.

[113] J. H. Chang and L. Tong, "Mitochondrial poly(A) polymerase and polyadenylation," *BBA - Gene Regulatory Mechanisms*, vol. 1819, no. 9, pp. 992–997, Sep. 2012.

[114] C.-Y. Kao and L. K. Read, "Opposing effects of polyadenylation on the stability of edited and unedited mitochondrial RNAs in *Trypanosoma brucei*," *Molecular and Cellular Biology*, vol. 25, no. 5, pp. 1634–1644, Mar. 2005.

[115] B. K. Adler, M. E. Harris, K. I. Bertrand, and S. L. Hajduk, "Modification of *Trypanosoma brucei* mitochondrial rRNA by posttranscriptional 3' polyuridine tail formation.," *Molecular and Cellular Biology*, vol. 11, no. 12, pp. 5878–5884, Dec. 1991.

[116] Y. Wang, "Rampant polyuridylylation of plastid gene transcripts in the dinoflagellate *Lingulodinium*," *Nucleic Acids Res*, vol. 34, no. 2, pp. 613–619, Jan. 2006.

[117] R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, and J. M. Merrick, "Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd.," *Science*, vol. 269, no. 5223, pp. 496–512, Jul. 1995.

[118] S. C. Schuster, "Next-generation sequencing transforms today's biology," *Nature Methods*, vol. 5, no. 1, pp. 16–18, Dec. 2007.

[119] E. L. van Dijk, H. Auger, Y. Jaszczyszyn, and C. Thermes, "Ten years of next-generation sequencing technology," *Trends in Genetics*, vol. 30, no. 9, pp. 418–426, Sep. 2014.

[120] J.-H. Lee, J. K. Ang, and X. Xiao, "Analysis and design of RNA sequencing experiments for identifying RNA editing and other single-nucleotide variants.," *RNA*, vol. 19, no. 6, pp. 725–732, Jun. 2013.

[121] Q. Zhang and X. Xiao, "Genome sequence-independent identification of RNA editing sites.," *Nature Methods*, vol. 12, no. 4, pp. 347–350, Apr. 2015.

[122] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with Bowtie 2," *Nature Methods*, vol. 9, no. 4, pp. 357–359, Mar. 2012.

[123] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter, "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks.," *Nature Protocols*, vol. 7, no. 3, pp. 562–578, Mar. 2012.

[124] A. Dobin, C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson, and T. R. Gingeras, "STAR: ultrafast universal RNA-seq aligner," *Bioinformatics*, vol. 29, no. 1, pp. 15–21, Dec. 2012.

[125] J. H. Bahn, J.-H. Lee, G. Li, C. Greer, G. Peng, and X. Xiao, "Accurate identification of A-to-I RNA editing in human by transcriptome sequencing.," *Genome Res.*, vol. 22, no. 1, pp. 142–150, Jan. 2012.

[126] E. Garrison and G. Marth, “Haplotype-based variant detection from short-read sequencing.” 17-Jul-2012.

[127] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly, “A framework for variation discovery and genotyping using next-generation DNA sequencing data.,” *Nature Publishing Group*, vol. 43, no. 5, pp. 491–498, May 2011.

[128] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, 1000 Genome Project Data Processing Subgroup, “The Sequence Alignment/Map format and SAMtools,” *Bioinformatics*, vol. 25, no. 16, pp. 2078–2079, Aug. 2009.

[129] H. Li, J. Wang, G. Mor, and J. Sklar, “A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells.,” *Science*, vol. 321, no. 5894, pp. 1357–1361, Sep. 2008.

[130] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg, “Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.,” *Genome Biol.*, vol. 10, no. 3, p. R25, 2009.

[131] R. Ekblom and J. B. W. Wolf, “A field guide to whole-genome sequencing, assembly and annotation,” *Evol Appl*, vol. 7, no. 9, pp. 1026–1042, Jun. 2014.

[132] H. Ellegren, “Genome sequencing and population genomics in non-model organisms,” *Trends in Ecology & Evolution*, vol. 29, no. 1, pp. 51–63, Jan. 2014.

[133] G. Parra, K. Bradnam, and I. Korf, “CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes,” *Bioinformatics*, vol. 23, no. 9, pp. 1061–1067, May 2007.

[134] V. Ter-Hovhannisyan, A. Lomsadze, Y. O. Chernoff, and M. Borodovsky, “Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training,” *Genome Res.*, vol. 18, no. 12, pp. 1979–1990, Oct. 2008.

[135] M. Stanke and S. Waack, “Gene prediction with a hidden Markov model and a new intron submodel,” *Bioinformatics*, vol. 19, no. 2, pp. ii215–ii225, Oct. 2003.

[136] I. Korf, “Gene finding in novel genomes,” *BMC Bioinformatics*, vol. 5, no. 1, p. 1, May 2004.

[137] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev, “Full-length transcriptome assembly from RNA-Seq data without a reference genome,” *Nat Biotechnol*, vol. 29, no. 7, pp. 644–652, May 2011.

[138] M. Martin, “Cutadapt removes adapter sequences from high-throughput sequencing reads,” *EMBnet.journal*, vol. 17, no. 1, p. 10, Aug. 2011.

[139] The UniProt Consortium, “Ongoing and future developments at the Universal Protein Resource,” *Nucleic Acids Res*, vol. 39, no. Database, pp. D214–D219, Dec. 2010.

[140] S. Altschul, “Gapped BLAST and PSI-BLAST: a new generation of protein database search programs,” *Nucleic Acids Res*, vol. 25, no. 17, pp. 3389–3402, Sep. 1997.

[141] G. Slater and E. Birney, “BMC Bioinformatics,” *BMC Bioinformatics*, vol. 6, no. 1, pp. 31–11, 2005.

[142] S. Hunter, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. A. Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats, “InterPro: the integrative protein signature database,” *Nucleic Acids Res*, vol. 37, no. Database, pp. D211–D215, Jan. 2009.

[143] M. Punta, P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, N. Pang, K. Forslund, G. Ceric, J. Clements, A. Heger, L. Holm, E. L. L. Sonnhammer, S. R. Eddy, A. Bateman, and R. D. Finn, “The Pfam protein families database,” *Nucleic Acids Res*, vol. 40, no. 1, pp. D290–D301, Dec. 2011.

[144] C. Bru, “The ProDom database of protein domain families: more emphasis on 3D,” *Nucleic Acids Res*, vol. 33, no. Database issue, pp. D212–D215, Dec. 2004.

[145] C. J. A. Sigrist, “PROSITE: A documented database using patterns and profiles as motif descriptors,” *Briefings in Bioinformatics*, vol. 3, no. 3, pp. 265–274, Jan. 2002.

[146] H. Mi, N. Guo, A. Kejariwal, and P. D. Thomas, “PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways,” *Nucleic Acids Res*, vol. 35, no. Database, pp. D247–D252, Jan. 2007.

[147] E. P. Nawrocki and S. R. Eddy, “Infernal 1.1: 100-fold faster RNA homology searches,” *Bioinformatics*, vol. 29, no. 22, pp. 2933–2935, Oct. 2013.

[148] P. Schattner, A. N. Brooks, and T. M. Lowe, “The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs,” *Nucleic Acids Res*, vol. 33, no. Web Server, pp. W686–W689, Jul. 2005.

[149] C. Holt and M. Yandell, “MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects.,” *BMC Bioinformatics*, vol. 12, no. 1, pp. 491–14, 2011.

[150] S. Moreira, E. Noutahi, G. Lamoureux, and G. Burger, “Three-dimensional structure model and predicted ATP interaction rewiring of a deviant RNA ligase 2.,” *BMC Structural Biology*, vol. 15, no. 1, pp. 20–15, 2015.

[151] R. P. Huntley, M. A. Harris, Y. Alam-Faruque, J. A. Blake, S. Carbon, H. Dietze, E. C. Dimmer, R. E. Foulger, D. P. Hill, V. K. Khodiyar, A. Lock, J. Lomax, R. C. Lovering, P. Mutowo-Meullenet, T. Sawford, K. Van Auken, V. Wood, and C. J. Mungall, "A method for increasing expressivity of Gene Ontology annotations using a compositional approach," *BMC Bioinformatics*, vol. 15, no. 1, pp. 155–11, 2014.

[152] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: back to metabolism in KEGG.," *Nucleic Acids Res*, vol. 42, no. Database issue, pp. D199–205, Jan. 2014.

[153] R. Caspi, R. Billington, L. Ferrer, H. Foerster, C. A. Fulcher, I. M. Keseler, A. Kothari, M. Krummenacker, M. Latendresse, L. A. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, and P. D. Karp, "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases," *Nucleic Acids Res*, vol. 44, no. 1, pp. D471–D480, Jan. 2016.

[154] A. Stoltzfus, "Constructive neutral evolution: exploring evolutionary theorys curious disconnect," *Biology Direct*, vol. 7, no. 1, pp. 1–1, Oct. 2012.

[155] T. Cavalier-Smith, "Eukaryote kingdoms: seven or nine?," *BioSystems*, vol. 14, no. 3, pp. 461–481, 1981.

[156] A. G. B. Simpson, "The identity and composition of the Euglenozoa," *Archiv fr Protistenkunde : Protozoen, Algen, Pilze*, vol. 148, no. 3, pp. 318–328, Oct. 1997.

[157] D. A. Maslov, S. Yasuhira, and L. Simpson, "Phylogenetic Affinities of *Diplonema* within the Euglenozoa as Inferred from the SSU rRNA Gene and Partial COI Protein Sequences," *Protist*, vol. 150, no. 1, pp. 33–42, Mar. 1999.

[158] D. Dooijes, I. Chaves, R. Kieft, A. Dirks-Mulder, W. Martin, and P. Borst, "Base J originally found in kinetoplastida is also a minor constituent of nuclear DNA of Euglena gracilis.," *Nucleic Acids Res*, vol. 28, no. 16, pp. 3017–3021, Aug. 2000.

[159] N. R. Sturm, D. A. Maslov, E. C. Grisard, and D. A. Campbell, "*Diplonema* spp. possess spliced leader RNA genes similar to the Kinetoplastida.," *J Eukaryotic Microbiology*, vol. 48, no. 3, pp. 325–331, May 2001.

[160] P. Borst and R. Sabatini, "Base J: discovery, biosynthesis, and possible functions.," *Annu. Rev. Microbiol.*, vol. 62, no. 1, pp. 235–251, 2008.

[161] P. J. Keeling, "The endosymbiotic origin, diversification and fate of plastids," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 365, no. 1541, pp. 729–748, Feb. 2010.

[162] M. L. Kent, R. A. Elston, T. A. Nerad, and T. K. Sawyer, "An *Isonema*-like flagellate (Protozoa: Mastigophora) infection in larval geoduck clams, Panope abrupta.," *J. Invertebr. Pathol.*, vol. 50, no. 3, pp. 221–229, Nov. 1987.

[163] R. E. Triemer and D. W. Ott, "Ultrastructure of *Diplonema ambulator* larsen & patterson (euglenozoa) and its relationship to *Isonema*.," *European Journal of Protistology*, vol. 25, no. 4, pp. 316–320, Jun. 1990.

[164] E. Lara, D. Moreira, A. Vereshchaka, and P. López-García, "Pan-oceanic distribution of new highly diverse clades of deep-sea diplonemids," *Environmental Microbiology*, vol. 11, no. 1, pp. 47–55, Jan. 2009.

[165] C. de Vargas, S. Audic, N. Henry, J. Decelle, F. Mahé, R. Logares, E. Lara, C. Berney, N. Le Bescot, I. Probert, M. Carmichael, J. Poulain, S. Romac, S. Colin, J.-M. Aury, L. Bittner, S. Chaffron, M. Dunthorn, S. Engelen, O. Flegontova, L. Guidi, A. Horák, O. Jaillon, G. Lima-Mendez, J. Lukeš, S. Malviya, R. Morard, M. Mulot, E. Scalco, R. Siano, F. Vincent, A. Zingone, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, Tara Oceans Coordinators, S. G. Acinas, P. Bork, C. Bowler, G. Gorsky, N. Grimsley, P. Hingamp, D. Iudicone, F. Not, H. Ogata, S. Pesant, J. Raes, M. E. Sieracki, S. Speich, L. Stemmann, S. Sunagawa, J. Weissenbach, P. Wincker, and E. Karsenti, "Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean.," *Science*, vol. 348, no. 6237, pp. 1261605–1261605, May 2015.

[166] E. Schnepf, "Light and Electron Microscopical Observations in *Rhynchopus coscinodiscivorus* spec. nov., a Colorless, Phagotrophic Euglenozoon with Concealed Flagella)," *Archiv für Protistenkunde*, vol. 144, no. 1, pp. 63–74, Mar. 1994.

[167] S. von der Heyden, E. E. Chao, K. Vickerman, and T. Cavalier-Smith, "Ribosomal RNA phylogeny of bodonid and diplonemid flagellates and the evolution of euglenozoa.," *J Eukaryotic Microbiology*, vol. 51, no. 4, pp. 402–416, Jul. 2004.

[168] G. N. Kiethega, M. Turcotte, and G. Burger, "Evolutionarily Conserved *cox1* Trans-Splicing Without cis-Motifs," *Molecular Biology and Evolution*, vol. 28, no. 9, pp. 2425–2428, Aug. 2011.

[169] M. W. Gray, B. F. Lang, and G. Burger, "Mitochondria of Protists," *Annu. Rev. Genet.*, vol. 38, no. 1, pp. 477–524, Dec. 2004.

[170] J. E. Feagin, M. I. Harrell, J. C. Lee, K. J. Coe, B. H. Sands, J. J. Cannone, G. Tami, M. N. Schnare, and R. R. Gutell, "The Fragmented Mitochondrial Ribosomal RNAs of *Plasmodium falciparum*," *PLoS ONE*, vol. 7, no. 6, pp. e38320–20, Jun. 2012.

[171] C. J. Jackson, J. E. Norman, M. N. Schnare, M. W. Gray, P. J. Keeling, and R. F. Waller, "Broad genomic and transcriptional analysis reveals a highly derived genome in dinoflagellate mitochondria," *BMC Biol*, vol. 5, no. 1, pp. 41–17, 2007.

[172] C. J. Jackson, S. G. Gornik, and R. F. Waller, "The Mitochondrial Genome and Transcriptome of the Basal Dinoflagellate *Hematodinium* sp.: Character Evolution within the Highly Derived Mitochondrial Genomes of Dinoflagellates," *Genome Biol Evol*, vol. 4, no. 1, pp. 59–72, Jan. 2012.

[173] D. E. Gillespie, N. A. Salazar, D. H. Rehkopf, and J. E. Feagin, "The fragmented mitochondrial ribosomal RNAs of *Plasmodium falciparum* have short A tails.," *Nucleic Acids Res*, vol. 27, no. 11, pp. 2416–2422, Jun. 1999.

[174] G. N. Kiethega, Y. Yan, M. Turcotte, and G. Burger, "RNA-level unscrambling of fragmented genes in *Diplonema* mitochondria," *RNA Biology*, vol. 10, no. 2, pp. 301–313, Oct. 2014.

[175] B. F. Lang and G. Burger, "Purification of mitochondrial and plastid DNA," *Nature Protocols*, vol. 2, no. 3, pp. 652–660, Mar. 2007.

[176] Y.-Q. Shen, E. O'Brien, L. Koski, B. F. Lang, and G. Burger, "EST Databases and Web Tools for EST Projects," in *Bioinformatics*, vol. 533, no. 11, J. M. Keith, Ed. Totowa, NJ: Humana Press, 2009, pp. 241–256.

[177] I. Milne, G. Stephen, M. Bayer, P. J. A. Cock, L. Pritchard, L. Cardle, P. D. Shaw, and D. Marshall, "Using Tablet for visual exploration of second-generation sequencing data," *Briefings in Bioinformatics*, vol. 14, no. 2, pp. 193–202, Mar. 2013.

[178] J. J. Cannone, S. Subramanian, M. N. Schnare, J. R. Collett, L. M. D'Souza, Y. Du, B. Feng, N. Lin, L. V. Madabusi, K. M. Müller, N. Pande, Z. Shang, N. Yu, and R. R. Gutell, "The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs.," *BMC Bioinformatics*, vol. 3, no. 1, p. 2, 2002.

[179] R. Lorenz, S. H. Bernhart, C. Höner Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, "ViennaRNA Package 2.0.," *Algorithms Mol Biol*, vol. 6, no. 1, p. 26, 2011.

[180] N. Ban, P. Nissen, J. Hansen, P. B. Moore, and T. A. Steitz, "The complete atomic structure of the large ribosomal subunit at 2.4 A resolution.," *Science*, vol. 289, no. 5481, pp. 905–920, Aug. 2000.

[181] I. C. Eperon, J. W. Janssen, J. H. Hoeijmakers, and P. Borst, "The major transcripts of the kinetoplast DNA of *Trypanosoma brucei* are very small ribosomal RNAs.," *Nucleic Acids Res*, vol. 11, no. 1, pp. 105–125, Jan. 1983.

[182] J. A. Mears, J. J. Cannone, S. M. Stagg, R. R. Gutell, R. K. Agrawal, and S. C. Harvey, "Modeling a minimal ribosome based on comparative sequence analysis.," *Journal of Molecular Biology*, vol. 321, no. 2, pp. 215–234, Aug. 2002.

[183] V. F. de la Cruz, A. M. Simpson, J. A. Lake, and L. Simpson, "Primary sequence and partial secondary structure of the 12S kinetoplast (mitochondrial) ribosomal RNA from *Leishmania tarentolae*: conservation of peptidyl-transferase structural elements.," *Nucleic Acids Res*, vol. 13, no. 7, pp. 2337–2356, Apr. 1985.

[184] P. Sloof, J. Van Den Burg, A. Voogd, R. Benne, M. Agostinelli, P. Borst, R. Gutell, and H. Noller, "Further characterization of the extremely small mitochondrial ribosomal RNAs from trypanosomes: a detailed comparison of the 9S and 12S RNAs from *Crithidia fasciculata* and *Trypanosoma brucei* with rRNAs from other organisms.," *Nucleic Acids Res*, vol. 13, no. 11, pp. 4171–4190, Jun. 1985.

[185] M. M. Yusupov, G. Z. Yusupova, A. Baucom, K. Lieberman, T. N. Earnest, J. H. Cate, and H. F. Noller, "Crystal structure of the ribosome at 5.5 A resolution.," *Science*, vol. 292, no. 5518, pp. 883–896, May 2001.

[186] M. R. Sharma, T. M. Booth, L. Simpson, D. A. Maslov, and R. K. Agrawal, "Structure of a mitochondrial ribosome with minimal RNA.," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 106, no. 24, pp. 9637–9642, Jun. 2009.

[187] A. B. Vaidya, R. Akella, and K. Suplick, "Sequences similar to genes for two mitochondrial proteins and portions of ribosomal RNA in tandemly arrayed 6-kilobase-pair DNA of a malarial parasite.," *Molecular & Biochemical Parasitology*, vol. 35, no. 2, pp. 97–107, Jun. 1989.

[188] J. E. Feagin, E. Werner, M. J. Gardner, D. H. Williamson, and R. J. M. Wilson, "Homologies between the contiguous and fragmented rRNAs of the two *Plasmodium falciparum* extrachromosomal DNAs are limited to core sequences," *Nucleic Acids Res*, vol. 20, no. 4, pp. 879–887, Feb. 1992.

[189] E. M. Denovan-Wright and R. W. Lee, "Comparative structure and genomic organization of the discontinuous mitochondrial ribosomal RNA genes of *Chlamydomonas eugametos* and *Chlamydomonas reinhardtii*.," *Journal of Molecular Biology*, vol. 241, no. 2, pp. 298–311, Aug. 1994.

[190] A. M. Nedelcu, R. W. Lee, C. Lemieux, M. W. Gray, and G. Burger, "The complete mitochondrial DNA sequence of *Scenedesmus obliquus* reflects an intermediate stage in the evolution of the green algal mitochondrial genome.," *Genome Res.*, vol. 10, no. 6, pp. 819–831, Jun. 2000.

[191] J. Fan and R. W. Lee, "Mitochondrial genome of the colorless green alga *Polytomella parva*: two linear DNA molecules with homologous inverted repeat Termini.," *Molecular Biology and Evolution*, vol. 19, no. 7, pp. 999–1007, Jul. 2002.

[192] D. F. Spencer, J. C. Collings, M. N. Schnare, and M. W. Gray, "Multiple spacer sequences in the nuclear large subunit ribosomal RNA gene of *Crithidia fasciculata*.," *EMBO J*, vol. 6, no. 4, pp. 1063–1071, Apr. 1987.

[193] M. N. Schnare and M. W. Gray, "Complete modification maps for the cytosolic small and large subunit rRNAs of *Euglena gracilis*: functional and evolutionary implications of contrasting patterns between the two rRNA components.," *Journal of Molecular Biology*, vol. 413, no. 1, pp. 66–83, Oct. 2011.

[194] M. Niemann, H. Kaibel, E. Schluter, K. Weitzel, M. Brecht, and H. U. Goringer, "Kinetoplastid RNA editing involves a 3' nucleotidyl phosphatase activity," *Nucleic Acids Res*, vol. 37, no. 6, pp. 1897–1906, Jan. 2009.

[195] H. Hashimi, S. L. Zimmer, M. L. Ammerman, L. K. Read, and J. Lukeš, "Dual core processing: MRB1 is an emerging kinetoplast RNA editing complex.," *Trends in Parasitology*, vol. 29, no. 2, pp. 91–99, Feb. 2013.

[196] M. Wassenegger and G. Krczal, "Nomenclature and functions of RNA-directed RNA polymerases.," *Trends in Plant Science*, vol. 11, no. 3, pp. 142–151, Mar. 2006.

[197] C. Cogoni and G. Macino, "Gene silencing in *Neurospora crassa* requires a protein homologous to RNA-dependent RNA polymerase.," *Nature*, vol. 399, no. 6732, pp. 166–169, May 1999.

[198] B. Ding, "Viroids: self-replicating, mobile, and fast-evolving noncoding regulatory RNAs.," *WIREs RNA*, vol. 1, no. 3, pp. 362–375, Nov. 2010.

[199] J. J. Polashock and B. I. Hillman, "A small mitochondrial double-stranded (ds) RNA element associated with a hypovirulent strain of the chestnut blight fungus and ancestrally related to yeast cytoplasmic T and W dsRNAs.," *Proc Natl Acad Sci USA*, vol. 91, no. 18, pp. 8680–8684, Aug. 1994.

[200] P. M. Finnegan and G. G. Brown, "Autonomously replicating RNA in mitochondria of maize plants with S-type cytoplasm.," *Proc Natl Acad Sci USA*, vol. 83, no. 14, pp. 5175–5179, Jul. 1986.

[201] J. R. Bracht, W. Fang, A. D. Goldman, E. Dolzhenko, E. M. Stein, and L. F. Landweber, "Genomes on the Edge: Programmed Genome Instability in Ciliates," *Cell*, vol. 152, no. 3, pp. 406–416, Jan. 2013.

[202] O. Rechavi, G. Minevich, and O. Hobert, "Transgenerational inheritance of an acquired small RNA-based antiviral response in *C. elegans*.," *Cell*, vol. 147, no. 6, pp. 1248–1256, Dec. 2011.

[203] R. Mahendran, M. S. Spottswood, A. Ghate, M. L. Ling, K. Jeng, and D. L. Miller, "Editing of the mitochondrial small subunit rRNA in *Physarum polycephalum*.," *EMBO J*, vol. 13, no. 1, pp. 232–240, Jan. 1994.

[204] C. Barth, U. Greferath, M. Kotsifas, and P. R. Fisher, "Polycistronic transcription and editing of the mitochondrial small subunit (SSU) ribosomal RNA in *Dictyostelium discoideum*.," *Curr Genet*, vol. 36, no. 1, pp. 55–61, Aug. 1999.

[205] W. A. Decatur and M. J. Fournier, "RNA-guided Nucleotide Modification of Ribosomal and Other RNAs," *Journal of Biological Chemistry*, vol. 278, no. 2, pp. 695–698, Jan. 2003.

[206] A. S. Petrov, C. R. Bernier, E. Hershkovits, Y. Xue, C. C. Waterbury, C. Hsiao, V. G. Stepanov, E. A. Gaucher, M. A. Grover, S. C. Harvey, N. V. Hud, R. M. Wartell, G. E. Fox, and L. D. Williams, "Secondary structure and domain architecture of the 23S and 5S rRNAs," *Nucleic Acids Res*, vol. 41, no. 15, pp. 7522–7535, Aug. 2013.

[207] S. Maas, Y. Kawahara, K. M. Tamburro, and K. Nishikura, "A-to-I RNA editing and human disease.," *RNA Biology*, vol. 3, no. 1, pp. 1–9, Jan. 2006.

[208] S. Dunin-Horkawicz, A. Czerwoniec, M. J. Gajda, M. Feder, H. Grosjean, and J. M. Bujnicki, "MODOMICS: a database of RNA modification pathways.," *Nucleic Acids Res*, vol. 34, no. Database issue, pp. D145–9, Jan. 2006.

[209] J. Rodriguez, J. S. Menet, and M. Rosbash, "Nascent-seq indicates widespread cotranscriptional RNA editing in Drosophila.," *Molecular Cell*, vol. 47, no. 1, pp. 27–37, Jul. 2012.

[210] J. M. Gott, "Expanding genome capacity via RNA editing," *Comptes Rendus Biologies*, vol. 326, no. 10, pp. 901–908, Oct. 2003.

[211] V. Knoop, "When you can't trust the DNA: RNA editing changes transcript sequences," *Cell. Mol. Life Sci.*, vol. 68, no. 4, pp. 567–586, Oct. 2010.

[212] C. Basilio, A. J. Wahba, P. Lengyel, J. F. Speyer, and S. Ochoa, "Synthetic polynucleotides and the amino acid code. V.," *Proc Natl Acad Sci USA*, vol. 48, no. 4, pp. 613–616, Apr. 1962.

[213] J. E. Jackman and J. D. Alfonzo, "Transfer RNA modifications: nature's combinatorial chemistry playground.," *WIREs RNA*, vol. 4, no. 1, pp. 35–48, Jan. 2013.

[214] H. Betat, Y. Long, J. Jackman, and M. Mörl, "From End to End: tRNA Editing at 5'- and 3'-Terminal Positions," *IJMS*, vol. 15, no. 12, pp. 23975–23998, Dec. 2014.

[215] F. Grewe, S. Herres, P. Viehover, M. Polsakiewicz, B. Weisshaar, and V. Knoop, "A unique transcriptome: 1782 positions of RNA editing alter 1406 codon identities in mitochondrial mRNAs of the lycophyte *Isoetes engelmannii*," *Nucleic Acids Res*, vol. 39, no. 7, pp. 2890–2902, Apr. 2011.

[216] J. Hecht, F. Grewe, and V. Knoop, "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 Biol Evol*, vol. 3, no. 0, pp. 344–358, Jan. 2011.

[217] C. J. Fu, S. Sheikh, W. Miao, S. G. E. Andersson, and S. L. Baldauf, "Missing Genes, Multiple ORFs, and C-to-U Type RNA Editing in *Acrasis kona* (Heterolobosea, Excavata) Mitochondrial DNA," *Genome Biol Evol*, vol. 6, no. 9, pp. 2240–2257, Sep. 2014.

[218] G. Burger, Y. Yan, P. Javadi, and B. F. Lang, "Group I-intron trans-splicing and mRNA editing in the mitochondria of placozoan animals," *Trends in Genetics*, vol. 25, no. 9, pp. 381–386, Sep. 2009.

[219] A. Janke and S. Pääbo, "Editing of a tRNA anticodon in marsupial mitochondria changes its codon recognition.," *Nucleic Acids Res*, vol. 21, no. 7, pp. 1523–1525, Apr. 1993.

[220] L. Randau, B. J. Stanley, A. Kohlway, S. Mechta, Y. Xiong, and D. Söll, "A cytidine deaminase edits C to U in transfer RNAs in Archaea.," *Science*, vol. 324, no. 5927, pp. 657–659, May 2009.

[221] B. R. Rosenberg, C. E. Hamilton, M. M. Mwangi, S. Dewell, and F. N. Papavasiliou, "Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3' UTRs.," *Nat. Struct. Mol. Biol.*, vol. 18, no. 2, pp. 230–236, Feb. 2011.

[222] M. Takenaka, A. Zehrmann, D. Verbitskiy, B. Härtel, and A. Brennicke, "RNA Editing in Plants and Its Evolution," *Annu. Rev. Genet.*, vol. 47, no. 1, pp. 335–352, Nov. 2013.

[223] B.-E. Wulff and K. Nishikura, "Substitutional A-to-I RNA editing.," *WIREs RNA*, vol. 1, no. 1, pp. 90–101, Jul. 2010.

[224] C. L. Kleinman and J. Majewski, "Comment on 'Widespread RNA and DNA Sequence Differences in the Human Transcriptome'," *Science*, vol. 335, no. 6074, pp. 1302–1302, Mar. 2012.

[225] L. Simpson, O. H. Thiemann, N. J. Savill, J. D. Alfonzo, and D. A. Maslov, "Evolution of RNA editing in trypanosome mitochondria.," *Proc Natl Acad Sci USA*, vol. 97, no. 13, pp. 6986–6993, Jun. 2000.

[226] J. Lukeš, O. Flegontova, and A. Horák, "Diplonemids," *Current Biology*, vol. 25, no. 16, pp. R702–R704, Aug. 2015.

[227] M. Valach, S. Moreira, G. N. Kiethega, and G. Burger, "Trans-splicing and RNA editing of LSU rRNA in *Diplonema* mitochondria," *Nucleic Acids Res*, vol. 42, no. 4, pp. 2660–2672, Feb. 2014.

[228] N. Rodríguez-Ezpeleta, S. Teijeiro, L. Forget, G. Burger, and B. F. Lang, "Construction of cDNA Libraries: Focus on Protists and Fungi," in *Bioinformatics*, vol. 533, no. 3, J. M. Keith, Ed. Totowa, NJ: Humana Press, 2009, pp. 33–47.

[229] P. B. Cattenoz, R. J. Taft, E. Westhof, and J. S. Mattick, "Transcriptome-wide identification of A > I RNA editing sites by inosine specific cleavage," *RNA*, vol. 19, no. 2, pp. 257–270, Jan. 2013.

[230] D. P. Morse and B. L. Bass, "Detection of inosine in messenger RNA by inosine-specific cleavage.," *Biochemistry*, vol. 36, no. 28, pp. 8429–8434, Jul. 1997.

[231] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H. H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, and J. C. Venter, "A whole-genome assembly of *Drosophila*.," *Science*, vol. 287, no. 5461, pp. 2196–2204, Mar. 2000.

[232] W. Li, L. Jaroszewski, and A. Godzik, "Tolerating some redundancy significantly speeds up clustering of large protein databases.," *Bioinformatics*, vol. 18, no. 1, pp. 77–82, Jan. 2002.

[233] H. Thorvaldsdottir, J. T. Robinson, and J. P. Mesirov, "Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration," *Briefings in Bioinformatics*, vol. 14, no. 2, pp. 178–192, Mar. 2013.

[234] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Res*, vol. 32, no. 5, pp. 1792–1797, Mar. 2004.

[235] W. S. J. Valdar, "Scoring residue conservation.," *Proteins*, vol. 48, no. 2, pp. 227–241, Aug. 2002.

[236] Y. Xie, G. Wu, J. Tang, R. Luo, J. Patterson, S. Liu, W. Huang, G. He, S. Gu, S. Li, X. Zhou, T. W. Lam, Y. Li, X. Xu, G. K. S. Wong, and J. Wang, "SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads," *Bioinformatics*, vol. 30, no. 12, pp. 1660–1666, Jun. 2014.

[237] R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D. W. Cheung, S.-M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T.-W. Lam, and J. Wang, "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler.," *Gigascience*, vol. 1, no. 1, p. 18, 2012.

[238] G. A. Van der Auwera, M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy-Moonshine, T. Jordan, K. Shakir, D. Roazen, J. Thibault, E. Banks, K. V. Garimella, D. Altshuler, S. Gabriel, and M. A. DePristo, "From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline.," *Curr Protoc Bioinformatics*, vol. 43, pp. 11.10.1–33, 2013.

[239] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble, "MEME SUITE: tools for motif discovery and searching," *Nucleic Acids Res*, vol. 37, no. Web Server, pp. W202–W208, Jun. 2009.

[240] D. H. Mathews, "Using the RNAstructure Software Package to Predict Conserved RNA Structures.," *Curr Protoc Bioinformatics*, vol. 46, pp. 12.4.1–22, 2014.

[241] S. Bellaousov, J. S. Reuter, M. G. Seetin, and D. H. Mathews, "RNAstructure: web servers for RNA secondary structure prediction and analysis," *Nucleic Acids Res*, vol. 41, no. 1, pp. W471–W474, Jun. 2013.

[242] S. Will, T. Joshi, I. L. Hofacker, P. F. Stadler, and R. Backofen, "LocARNA-P: Accurate boundary prediction and improved detection of structural RNAs," *RNA*, vol. 18, no. 5, pp. 900–914, Apr. 2012.

[243] C. Smith, S. Heyne, A. S. Richter, S. Will, and R. Backofen, "Freiburg RNA Tools: a web server integrating INTARNA, EXPARNA and LOCARNA," *Nucleic Acids Res*, vol. 38, no. Web Server, pp. W373–W377, Jun. 2010.

[244] E. L. Sonnhammer and R. Durbin, "A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis.," *Gene*, vol. 167, no. 1, pp. GC1–10, Dec. 1995.

[245] L. Käll, A. Krogh, and E. L. L. Sonnhammer, "A combined transmembrane topology and signal peptide prediction method.," *Journal of Molecular Biology*, vol. 338, no. 5, pp. 1027–1036, May 2004.

[246] P. Artimo, M. Jonnalagedda, K. Arnold, D. Baratin, G. Csárdi, E. de Castro, S. Duvaud, V. Flegel, A. Fortier, E. Gasteiger, A. Grosdidier, C. Hernandez, V. Ioannidis, D. Kuznetsov, R. Liechti, S. Moretti, K. Mostaguir, N. Redaschi, G. Rossier, I. Xenarios, and H. Stockinger, "ExPASy: SIB bioinformatics resource portal.," *Nucleic Acids Res*, vol. 40, no. Web Server issue, pp. W597–603, Jul. 2012.

[247] C. Vlcek, W. Marande, S. Teijeiro, J. Lukes, and G. Burger, "Systematically fragmented genes in a multipartite mitochondrial genome," *Nucleic Acids Res*, vol. 39, no. 3, pp. 979–988, Feb. 2011.

[248] W. Pett, J. F. Ryan, K. Pang, J. C. Mullikin, M. Q. Martindale, A. D. Baxevanis, and D. V. Lavrov, "Extreme mitochondrial evolution in the ctenophore *Mnemiopsis leidyi*: Insight from mtDNA and the nuclear genome.," *Mitochondrial DNA*, vol. 22, no. 4, pp. 130–142, Aug. 2011.

[249] J. E. Feagin, B. L. Mericle, E. Werner, and M. Morris, "Identification of additional rRNA fragments encoded by the *Plasmodium falciparum* 6 kb element.," *Nucleic Acids Res*, vol. 25, no. 2, pp. 438–446, Jan. 1997.

[250] J. P. Mower and J. D. Palmer, "Patterns of partial RNA editing in mitochondrial genes of *Beta vulgaris*," *Mol Genet Genomics*, vol. 276, no. 3, pp. 285–293, Jul. 2006.

[251] E. Picardi, D. S. Horner, M. Chiara, R. Schiavon, G. Valle, and G. Pesole, "Large-scale detection and analysis of RNA editing in grape mtDNA by RNA deep-sequencing," *Nucleic Acids Res*, vol. 38, no. 14, pp. 4755–4767, Aug. 2010.

[252] M. Inada, T. Sasaki, M. Yukawa, T. Tsudzuki, and M. Sugiura, "A systematic search for RNA editing sites in pea chloroplasts: an editing event causes diversification from the evolutionarily conserved amino acid sequence.," *Plant Cell Physiol.*, vol. 45, no. 11, pp. 1615–1622, Nov. 2004.

[253] D. Verbitskiy, M. Takenaka, J. Neuwirt, J. A. van der Merwe, and A. Brennicke, "Partially edited RNAs are intermediates of RNA editing in plant mitochondria.," *Plant J*, vol. 47, no. 3, pp. 408–416, Aug. 2006.

[254] T. Cardi, P. Giegé, S. Kahlau, and N. Scotti, "Expression Profiling of Organellar Genes," in *Genomics of Chloroplasts and Mitochondria*, vol. 35, no. 14, Dordrecht: Springer Netherlands, 2012, pp. 323–355.

[255] I. Aphasizheva and R. Aphasizhev, "RET1-Catalyzed Uridylylation Shapes the Mitochondrial Transcriptome in *Trypanosoma brucei*," *Molecular and Cellular Biology*, vol. 30, no. 6, pp. 1555–1567, Mar. 2010.

[256] S. Mungpakdee, C. Shinzato, T. Takeuchi, T. Kawashima, R. Koyanagi, K. Hisata, M. Tanaka, H. Goto, M. Fujie, S. Lin, N. Satoh, and E. Shoguchi, "Massive Gene Transfer and Extensive RNA Editing of a Symbiotic Dinoflagellate Plastid Genome," *Genome Biol Evol*, vol. 6, no. 6, pp. 1408–1422, Jun. 2014.

[257] J. Roy, D. Faktorová, O. Benada, J. Lukeš, and G. Burger, "Description of Rhynchopus euleeides n. sp. (Diplonemea), a free-living marine euglenozoan.," *J Eukaryotic Microbiology*, vol. 54, no. 2, pp. 137–145, Mar. 2007.

[258] L. Simpson, "The mitochondrial genome of kinetoplastid protozoa: genomic organization, transcription, replication, and evolution.," *Annu. Rev. Microbiol.*, vol. 41, no. 1, pp. 363–382, 1987.

[259] D. V. Lavrov, M. Adamski, P. Chevaldonné, and M. Adamska, "Extensive Mitochondrial mRNA Editing and Unusual Mitochondrial Genome Organization in Calcaronean Sponges.," *Curr. Biol.*, vol. 26, no. 1, pp. 86–92, Jan. 2016.

[260] I. Alseth, B. Dalhus, and M. Bjørås, "Inosine in DNA and RNA.," *Curr. Opin. Genet. Dev.*, vol. 26, pp. 116–123, Jun. 2014.

[261] M. W. Gray, "O2'-Methylinosine, a constituent of the ribosomal RNA of *Crithidia fasciculata*.," *Nucleic Acids Res*, vol. 3, no. 4, pp. 977–988, Apr. 1976.

[262] F. V. Murphy and V. Ramakrishnan, "Structure of a purine-purine wobble base pair in the decoding center of the ribosome.," *Nat. Struct. Mol. Biol.*, vol. 11, no. 12, pp. 1251–1252, Dec. 2004.

[263] D. M. Driscoll and T. L. Innerarity, "RNA editing by cytidine deamination in mammals ," in *RNA Editing*, vol. 34, Elsevier, 2007, pp. 61–76.

[264] V. Blanc, E. Park, S. Schaefer, M. Miller, Y. Lin, S. Kennedy, A. M. Billing, H. Ben Hamidane, J. Graumann, A. Mortazavi, J. H. Nadeau, and N. O. Davidson, "Genome-wide identification and functional analysis of Apobec-1-mediated C-to-U RNA editing in mouse small intestine and liver.," *Genome Biol.*, vol. 15, no. 6, p. R79, 2014.

[265] R. F. Hough and B. L. Bass, "Adenosine deaminases that act on RNA ," in *RNA Editing*, Elsevier, 2007, pp. 77–108.

[266] R. Aphasizhev and I. Aphasizheva, "Uridine insertion/deletion editing in trypanosomes: a playground for RNA-guided information transfer.," *WIREs RNA*, vol. 2, no. 5, pp. 669–685, Sep. 2011.

[267] K. Nishikura, "Functions and Regulation of RNA Editing by ADAR Deaminases," *Annu. Rev. Biochem.*, vol. 79, no. 1, pp. 321–349, Jun. 2010.

[268] A. Stoltzfus, "On the Possibility of Constructive Neutral Evolution," pp. 1–13, Jul. 1999.

[269] F. Jacob, "Evolution and tinkering.," *Science*, vol. 196, no. 4295, pp. 1161–1166, Jun. 1977.

[270] H. U. Göringer, "'Gestalt,' Composition and Function of the *Trypanosoma brucei* Editosome," *Annu. Rev. Microbiol.*, vol. 66, no. 1, pp. 65–82, Oct. 2012.

[271] P. A, W. T, P. EV, and P. JD, "RNA purification by preparative polyacrylamide gel electrophoresis.," *Methods Enzymol*, vol. 530, pp. 315–330, 2013.

[272] W. SE and L. J, "Reverse transcriptase dideoxy sequencing of RNA.," *Methods Enzymol*, vol. 530, pp. 347–359, 2013.

[273] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, "NIH Image to ImageJ: 25 years of image analysis," *Nat methods*, 2012.

[274] J. Archer, G. Whiteley, N. R. Casewell, R. A. Harrison, and S. C. Wagstaff, "VTBuilder: a tool for the assembly of multi isoform transcriptomes," *BMC Bioinformatics*, vol. 15, no. 1, p. 389, Dec. 2014.

[275] G. S. Pearson, *The UNSCOM Saga*. London: Palgrave Macmillan UK, 1999.

[276] S. Janssen and R. Giegerich, "The RNA shapes studio," *Bioinformatics*, vol. 31, no. 3, pp. btu649–425, Oct. 2014.

[277] C. Trapnell, L. Pachter, and S. L. Salzberg, "TopHat: discovering splice junctions with RNA-Seq.," *Bioinformatics*, vol. 25, no. 9, pp. 1105–1111, May 2009.

[278] S. W. Smith, R. Overbeek, C. R. Woese, W. Gilbert, and P. M. Gillevet, "The genetic data environment an expandable GUI for multiple sequence analysis.," *Comput. Appl. Biosci.*, vol. 10, no. 6, pp. 671–675, Dec. 1994.

[279] K. D. Stuart, A. Schnaufer, N. L. Ernst, and A. K. Panigrahi, "Complex management: RNA editing in trypanosomes," *Trends in Biochemical Sciences*, vol. 30, no. 2, pp. 97–105, Feb. 2005.

[280] M. W. Gray, J. Lukes, J. M. Archibald, P. J. Keeling, and W. F. Doolittle, "Irremediable Complexity?," *Science*, vol. 330, no. 6006, pp. 920–921, Nov. 2010.

[281] M. Lynch, B. Koskella, and S. Schaack, "Mutation pressure and the evolution of organelle genomic architecture.," *Science*, vol. 311, no. 5768, pp. 1727–1730, Mar. 2006.

[282] J. M. Pascal, "DNA and RNA ligases: structural variations and shared mechanisms," *Current Opinion in Structural Biology*, vol. 18, no. 1, pp. 96–105, Feb. 2008.

[283] H. S. Subramanya, A. J. Doherty, S. R. Ashford, and D. B. Wigley, "Crystal structure of an ATP-dependent DNA ligase from bacteriophage T7.," *Cell*, vol. 85, no. 4, pp. 607–615, May 1996.

[284] C. L. Greer, C. L. Peebles, P. Gegenheimer, and J. Abelson, "Mechanism of action of a yeast RNA ligase in tRNA splicing.," *Cell*, vol. 32, no. 2, pp. 537–546, Feb. 1983.

[285] E. A. Arn and J. N. Abelson, "The 2'-5' RNA Ligase of *Escherichia coli*: Purification, cloning, and genomic disruption," *Journal of Biological Chemistry*, vol. 271, no. 49, pp. 31145–31153, Dec. 1996.

[286] R. Mazumder, L. M. Iyer, S. Vasudevan, and L. Aravind, "Detection of novel members, structure-function analysis and evolutionary classification of the 2H phosphoesterase superfamily.," *Nucleic Acids Res*, vol. 30, no. 23, pp. 5229–5243, Dec. 2002.

[287] J. Popow, M. Englert, S. Weitzer, A. Schleiffer, B. Mierzwa, K. Mechtler, S. Trowitzsch, C. L. Will, R. Luhrmann, D. Soll, and J. Martinez, "HSPC117 Is the Essential Subunit of a Human tRNA Splicing Ligase Complex," *Science*, vol. 331, no. 6018, pp. 760–764, Feb. 2011.

[288] N. Tanaka, B. Meineke, and S. Shuman, "RtcB, a Novel RNA Ligase, Can Catalyze tRNA Splicing and HAC1 mRNA Splicing in Vivo," *Journal of Biological Chemistry*, vol. 286, no. 35, pp. 30253–30257, Aug. 2011.

[289] J. Popow, A. Schleiffer, and J. Martinez, "Diversity and roles of (t)RNA ligases," *Cell. Mol. Life Sci.*, vol. 69, no. 16, pp. 2657–2670, Mar. 2012.

[290] M. Amitsur, R. Levitz, and G. Kaufmann, "Bacteriophage T4 anticodon nuclease, polynucleotide kinase and RNA ligase reprocess the host lysine tRNA.," *EMBO J*, vol. 6, no. 8, pp. 2499–2503, Aug. 1987.

[291] C. K. Ho and S. Shuman, "Bacteriophage T4 RNA ligase 2 (gp24.1) exemplifies a family of RNA ligases found in all phylogenetic domains.," *Proc Natl Acad Sci USA*, vol. 99, no. 20, pp. 12709–12714, Oct. 2002.

[292] N. Bakalara, A. M. Simpson, and L. Simpson, "The *Leishmania* kinetoplast-mitochondrion contains terminal uridylyltransferase and RNA ligase activities.," *Journal of Biological Chemistry*, vol. 264, no. 31, pp. 18679–18686, Nov. 1989.

[293] K. Stuart, R. Brun, S. Croft, A. Fairlamb, R. E. Gürtler, J. McKerrow, S. Reed, and R. Tarleton, "Kinetoplastids: related protozoan pathogens, different diseases," *J. Clin. Invest.*, vol. 118, no. 4, pp. 1301–1310, Apr. 2008.

[294] L. Simpson and A. Da Silva, "Isolation and characterization of kinetoplast DNA from *Leishmania tarentolae*.," *Journal of Molecular Biology*, vol. 56, no. 3, pp. 443–473, Mar. 1971.

[295] N. R. Sturm and L. Simpson, "Kinetoplast DNA minicircles encode guide RNAs for editing of cytochrome oxidase subunit III mRNA.," *Cell*, vol. 61, no. 5, pp. 879–884, Jun. 1990.

[296] R. Aphasizhev and I. Aphasizheva, "Mitochondrial RNA editing in trypanosomes: Small RNAs in control," *Biochimie*, vol. 100, no. c, pp. 125–131, May 2014.

[297] J. W. Cranston, R. Silber, V. G. Malathi, and J. Hurwitz, "Studies on ribonucleic acid ligase. Characterization of an adenosine triphosphate-inorganic pyrophosphate exchange reaction and demonstration of an enzyme-adenylate complex with T4 bacteriophage-induced enzyme.," *Journal of Biological Chemistry*, vol. 249, no. 23, pp. 7447–7456, Dec. 1974.

[298] S. Yin, C. K. Ho, and S. Shuman, "Structure-Function Analysis of T4 RNA Ligase 2," *Journal of Biological Chemistry*, vol. 278, no. 20, pp. 17601–17608, May 2003.

[299] C. K. Ho, L. K. Wang, C. D. Lima, and S. Shuman, "Structure and Mechanism of RNA Ligase," *Structure*, vol. 12, no. 2, pp. 327–339, Feb. 2004.

[300] J. Nandakumar, S. Shuman, and C. D. Lima, "RNA Ligase Structures Reveal the Basis for RNA Specificity and Conformational Changes that Drive Ligation Forward," *Cell*, vol. 127, no. 1, pp. 71–84, Oct. 2006.

[301] J. Deng, A. Schnaufer, R. Salavati, K. D. Stuart, and W. G. J. Hol, "High Resolution Crystal Structure of a Key Editosome Enzyme from *Trypanosoma brucei*: RNA Editing Ligase 1," *Journal of Molecular Biology*, vol. 343, no. 3, pp. 601–613, Oct. 2004.

[302] S. Altschul, "Basic Local Alignment Search Tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, Oct. 1990.

[303] L. Holm and P. Rosenstrom, "Dali server: conservation mapping in 3D," *Nucleic Acids Res*, vol. 38, no. Web Server, pp. W545–W549, Jun. 2010.

[304] Y. Zhang, "I-TASSER server for protein 3D structure prediction.," *BMC Bioinformatics*, vol. 9, no. 1, p. 40, 2008.

[305] J. U. Bowie, R. Lüthy, and D. Eisenberg, "A method to identify protein sequences that fold into a known three-dimensional structure.," *Science*, vol. 253, no. 5016, pp. 164–170, Jul. 1991.

[306] L. J. McGuffin, M. T. Buenavista, and D. B. Roche, "The ModFOLD4 server for the quality assessment of 3D protein models," *Nucleic Acids Res*, vol. 41, no. 1, pp. W368–W372, Jun. 2013.

[307] R. E. Amaro, R. V. Swift, and J. A. McCammon, "Functional and Structural Insights Revealed by Molecular Dynamics Simulations of an Essential RNA Editing Ligase in *Trypanosoma brucei*," *PLoS Neglected Tropical Diseases*, vol. 1, no. 2, pp. e68–10, Nov. 2007.

[308] M. A. Brooks, L. Meslet-Cladiére, M. Graille, J. Kuhn, K. Blondeau, H. Myllykallio, and H. van Tilbeurgh, "The structure of an archaeal homodimeric ligase which has RNA circularization activity," *Protein Sci.*, vol. 17, no. 8, pp. 1336–1345, Aug. 2008.

[309] F. B. Sheinerman, R. Norel, and B. Honig, "Electrostatic aspects of protein-protein interactions.," *Current Opinion in Structural Biology*, vol. 10, no. 2, pp. 153–159, Apr. 2000.

[310] S. Viollet, R. T. Fuchs, D. B. Munafo, F. Zhuang, and G. B. Robb, "T4 RNA Ligase 2 truncated active site mutants: improved tools for RNA analysis," *BMC Biotechnology*, vol. 11, no. 1, p. 72, Jul. 2011.

[311] N. Lartillot, H. Brinkmann, and H. Philippe, "Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model," *BMC Evol Biol*, vol. 7, no. 1, pp. S4–14, 2007.

[312] J. Felsenstein, "Cases in which Parsimony or Compatibility Methods will be Positively Misleading," *Systematic Biology*, vol. 27, no. 4, pp. 401–410, Dec. 1978.

[313] Y. J. Park, T. Budiarto, M. Wu, E. Pardon, J. Steyaert, and W. G. J. Hol, "The structure of the C-terminal domain of the largest editosome interaction protein and its role in promoting RNA binding by RNA-editing ligase L2," *Nucleic Acids Res*, vol. 40, no. 14, pp. 6966–6977, Aug. 2012.

[314] A. G. B. Simpson, E. E. Gill, H. A. Callahan, R. W. Litaker, and A. J. Roger, "Early Evolution within Kinetoplastids (Euglenozoa), and the Late Emergence of Trypanosomatids," *Protist*, vol. 155, no. 4, pp. 407–422, Dec. 2004.

[315] J. Lukeš, J. M. Archibald, P. J. Keeling, W. F. Doolittle, and M. W. Gray, "How a neutral evolutionary ratchet can build cellular complexity," *IUBMB Life*, vol. 63, no. 7, pp. 528–537, Jun. 2011.

[316] R. V. Swift, J. Durrant, R. E. Amaro, and J. A. McCammon, "Toward Understanding the Conformational Dynamics of RNA Ligation," *Biochemistry*, vol. 48, no. 4, pp. 709–719, Feb. 2009.

[317] A. E. Todd, C. A. Orengo, and J. M. Thornton, "Plasticity of enzyme active sites.," *Trends in Biochemical Sciences*, vol. 27, no. 8, pp. 419–426, Aug. 2002.

[318] J. Dudek, P. Rehling, and M. van der Laan, "Mitochondrial protein import: Common principles and physiological networks," *BBA - Molecular Cell Research*, vol. 1833, no. 2, pp. 274–285, Feb. 2013.

[319] B. Chevreux, T. Wetter, and S. S, "Genome sequence assembly using trace signals and additional sequence information," presented at the German Conference on Bioinformatics GCB GCB, 1999, pp. 45–56.

[320] R. D. Finn, J. Clements, and S. R. Eddy, "HMMER web server: interactive sequence similarity searching," *Nucleic Acids Res*, vol. 39, no. suppl, pp. W29–W37, Jun. 2011.

[321] S. R. Eddy, "Accelerated Profile HMM Searches," *PLoS Comput Biol*, vol. 7, no. 10, pp. e1002195–16, Oct. 2011.

[322] M. Aslett, C. Aurrecoechea, M. Berriman, J. Brestelli, B. P. Brunk, M. Carrington, D. P. Depledge, S. Fischer, B. Gajria, X. Gao, M. J. Gardner, A. Gingle, G. Grant, O. S. Harb, M. Heiges, C. Hertz-Fowler, R. Houston, F. Innamorato, J. Iodice, J. C. Kissinger, E. Kraemer, W. Li, F. J. Logan, J. A. Miller, S. Mitra, P. J. Myler, V. Nayak, C. Pennington, I. Phan, D. F. Pinney, G. Ramasamy, M. B. Rogers, D. S. Roos, C. Ross, D. Sivam, D. F. Smith, G. Srinivasamoorthy, C. J. Stoeckert, S. Subramanian, R. Thibodeau, A. Tivey, C. Treatman, G. Velarde, and H. Wang, "TriTrypDB: a functional genomic resource for the Trypanosomatidae," *Nucleic Acids Res*, vol. 38, no. Database, pp. D457–D462, Dec. 2009.

[323] A. Roy, A. Kucukural, and Y. Zhang, "I-TASSER: a unified platform for automated protein structure and function prediction," *Nat Protoc*, vol. 5, no. 4, pp. 725–738, Mar. 2010.

[324] D. Xu and Y. Zhang, "Improving the physical realism and structural accuracy of protein models by a two-step atomic-level energy minimization.," *Biophys. J.*, vol. 101, no. 10, pp. 2525–2534, Nov. 2011.

[325] G. Vriend, "WHAT IF: A molecular modeling and drug design program," *Journal of Molecular Graphics*, vol. 8, no. 1, pp. 52–56, Mar. 1990.

[326] S. Jo, T. Kim, V. G. Iyer, and W. Im, "CHARMM-GUI: A web-based graphical user interface for CHARMM," *J. Comput. Chem.*, vol. 29, no. 11, pp. 1859–1865, Mar. 2008.

[327] S. Pronk, S. Pall, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl, "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit," *Bioinformatics*, vol. 29, no. 7, pp. 845–854, Mar. 2013.

[328] A. D. MacKerell, D. Bashford, M. Bellott, R. L. Dunbrack, J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, B. Roux, M. Schlenkrich, J. C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, and M. Karplus, "All-atom empirical potential for molecular modeling and dynamics studies of proteins.," *J Phys Chem B*, vol. 102, no. 18, pp. 3586–3616, Apr. 1998.

[329] P. Bjelkmar, P. Larsson, M. A. Cuendet, B. Hess, and E. Lindahl, "Implementation of the CHARMM Force Field in GROMACS: Analysis of Protein Stability Effects from Correction Maps, Virtual Interaction Sites, and Water Models," *J. Chem. Theory Comput.*, vol. 6, no. 2, pp. 459–466, Feb. 2010.

[330] G. Bussi, D. Donadio, and M. Parrinello, "Canonical sampling through velocity rescaling," *J. Chem. Phys.*, vol. 126, no. 1, pp. 014101–8, 2007.

[331] M. Parrinello, "Polymorphic transitions in single crystals: A new molecular dynamics method," *J. Appl. Phys.*, vol. 52, no. 12, pp. 7182–10, 1981.

[332] S. Nosé and M. L. Klein, "Constant pressure molecular dynamics for molecular systems," *Mol. Phys.*, vol. 50, pp. 1055–1076, 2006.

[333] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features.," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, Dec. 1983.

[334] L. A. Mirny and E. I. Shakhnovich, "Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function.," *Journal of Molecular Biology*, vol. 291, no. 1, pp. 177–196, Aug. 1999.

[335] R. C. Team, "R: A language and environment for statistical computing," 2013.

[336] W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics.," *Journal of Molecular Graphics*, vol. 14, no. 1, pp. 33–8– 27–8, Feb. 1996.

[337] L. Schrödinger, "The PyMOL molecular graphics system, Version 1.3 R1," 2010.

[338] D. Frishman and P. Argos, "Knowledge-based protein secondary structure assignment," *Proteins*, vol. 23, no. 4, pp. 566–579, Dec. 1995.

[339] F. Glaser, T. Pupko, I. Paz, R. E. Bell, D. Bechor-Shental, E. Martz, and N. Ben-Tal, "ConSurf: Identification of Functional Regions in Proteins by Surface-Mapping of Phylogenetic Information," *Bioinformatics*, vol. 19, no. 1, pp. 163–164, Jan. 2003.

[340] I. Mayrose, "Comparison of Site-Specific Rate-Inference Methods for Protein Sequences: Empirical Bayesian Methods Are Superior," *Molecular Biology and Evolution*, vol. 21, no. 9, pp. 1781–1791, May 2004.

[341] M. Holst and F. Saied, "Multigrid solution of the Poisson—Boltzmann equation," *J. Comput. Chem.*, vol. 14, no. 1, pp. 105–113, Jan. 1993.

[342] M. J. Holst and F. Saied, "Numerical solution of the nonlinear Poisson–Boltzmann equation: Developing more robust and efficient methods," *J. Comput. Chem.*, vol. 16, no. 3, pp. 337–364, Mar. 1995.

[343] T. J. Dolinsky, P. Czodrowski, H. Li, J. E. Nielsen, J. H. Jensen, G. Klebe, and N. A. Baker, "PDB2PQR: expanding and upgrading automated preparation of biomolecular structures for molecular simulations," *Nucleic Acids Res*, vol. 35, no. Web Server, pp. W522–W525, May 2007.

[344] E. Jaspard and G. Hunault, "Comparison of Amino Acids Physico-Chemical Properties and Usage of Late Embryogenesis Abundant Proteins, Hydrophilins and WHy Domain," *PLoS ONE*, vol. 9, no. 10, p. e109570, Oct. 2014.

[345] S. Moreira, M. Valach, M. Aoulad-Aissa, C. Otto, and G. Burger, "Novel modes of RNA editing in mitochondria," *Nucleic Acids Res*, vol. 44, no. 10, pp. 4907–4919, Jun. 2016.

[346] A. Danchin, *Barque de Delphes (La): Ce que révèle le texte des génomes*. 1998.

[347] G. Burger, S. Moreira, and M. Valach, "Genes in Hiding.," *Trends in Genetics*, vol. 32, no. 9, pp. 553–565, Sep. 2016.

[348] I. J. MacRae, K. Zhou, F. Li, A. Repic, A. N. Brooks, W. Z. Cande, P. D. Adams, and J. A. Doudna, "Structural Basis for Double-Stranded RNA Processing by Dicer," *Science*, vol. 311, no. 5758, pp. 195–198, Jan. 2006.

[349] T. A. Nguyen, M. H. Jo, Y.-G. Choi, J. Park, S. C. Kwon, S. Hohng, V. N. Kim, and J.-S. Woo, "Functional Anatomy of the Human Microprocessor," *Cell*, vol. 161, no. 6, pp. 1374–1387, Jun. 2015.

[350] V. M. Reyes and J. Abelson, "Substrate recognition and splice site determination in yeast tRNA splicing," *Cell*, vol. 55, no. 4, pp. 719–730, Nov. 1988.

[351] S. Unterreitmeier, A. Fuchs, T. Schäffler, R. G. Heym, D. Frishman, and D. Langosch, "Phenylalanine Promotes Interaction of Transmembrane Domains via GxxxG Motifs," *Journal of Molecular Biology*, vol. 374, no. 3, pp. 705–718, Nov. 2007.

[352] M.-J. Kwon, Y. Choi, J.-H. Yun, W. Lee, I.-O. Han, and E.-S. Oh, "A Unique Phenylalanine in the Transmembrane Domain Strengthens Homodimerization of the Syndecan-2 Transmembrane Domain and Functionally Regulates Syndecan-2," *Journal of Biological Chemistry*, vol. 290, no. 9, pp. 5772–5782, Feb. 2015.

[353] G. B. McGaughey, M. Gagné, and A. K. Rappé, "pi-Stacking interactions. Alive and well in proteins.," *Journal of Biological Chemistry*, vol. 273, no. 25, pp. 15458–15463, Jun. 1998.

[354] Y. Oma, Y. Kino, N. Sasagawa, and S. Ishiura, "Comparative analysis of the cytotoxicity of homopolymeric amino acids," *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, vol. 1748, no. 2, pp. 174–179, May 2005.

[355] E. Gazit, "A possible role for pi-stacking in the self-assembly of amyloid fibrils.," *FASEB J.*, vol. 16, no. 1, pp. 77–83, Jan. 2002.

[356] M. Reches and E. Gazit, "Casting Metal Nanowires Within Discrete Self-Assembled Peptide Nanotubes," *Science*, vol. 300, no. 5619, pp. 625–627, Apr. 2003.

[357] E. Richardson, R. G. Dorrell, and C. J. Howe, "Genome-Wide Transcript Profiling Reveals the Coevolution of Plastid Gene Sequences and Transcript Processing Pathways in the Fucoxanthin Dinoflagellate *Karlodinium veneficum*," *Molecular Biology and Evolution*, vol. 31, no. 9, pp. 2376–2386, Aug. 2014.

[358] J. Lukeš, B. S. Leander, and P. J. Keeling, "Cascades of convergent evolution: the corresponding evolutionary histories of euglenozoans and dinoflagellates.," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 106, no. 1, pp. 9963–9970, Jun. 2009.

[359] T. Suematsu, L. Zhang, I. Aphasizheva, S. Monti, L. Huang, Q. Wang, C. E. Costello, and R. Aphasizhev, "Antisense Transcripts Delimit Exonucleolytic Activity of the Mitochondrial 3' Processome to Generate Guide RNAs," *Molecular Cell*, vol. 61, no. 3, pp. 364–378, Feb. 2016.

[360] M. Schallenberg-Rüdinger, H. Lenz, M. Polsakiewicz, J. M. Gott, and V. Knoop, "A survey of PPR proteins identifies DYW domains like those of land plant RNA editing factors in diverse eukaryotes," *RNA Biology*, vol. 10, no. 9, pp. 1549–1556, Jul. 2013.

[361] P. Schaap, I. Barrantes, P. Minx, N. Sasaki, R. W. Anderson, M. Bénard, K. K. Biggar, N. E. Buchler, R. Bundschuh, X. Chen, C. Fronick, L. Fulton, G. Golderer, N. Jahn, V. Knoop, L. F. Landweber, C. Maric, D. Miller, A. A. Noegel, R. Peace, G. Pierron, T. Sasaki, M. Schallenberg-Rüdinger, M. Schleicher, R. Singh, T. Spaller, K. B. Storey, T. Suzuki, C. Tomlinson, J. J. Tyson, W. C. Warren, E. R. Werner, G. Werner-Felmayer, R. K. Wilson, T. Winckler, J. M. Gott, G. Glöckner, and W. Marwan, "The *Physarum polycephalum* Genome Reveals Extensive Use of Prokaryotic Two-Component and Metazoan-Type Tyrosine Kinase Signaling," *Genome Biol Evol*, vol. 8, no. 1, pp. 109–125, Jan. 2016.

[362] R. Aphasizhev and I. Aphasizheva, "Emerging roles of PPR proteins in trypanosomes," *RNA Biology*, vol. 10, no. 9, pp. 1495–1500, Oct. 2014.

[363] B. S. Leander, "Did trypanosomatid parasites have photosynthetic ancestors?," *Trends in Microbiology*, vol. 12, no. 6, pp. 251–258, Jun. 2004.

[364] M. B. Rogers, P. R. Gilson, V. Su, G. I. McFadden, and P. J. Keeling, "The complete chloroplast genome of the chlorarachniophyte Bigelowiella natans: evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts.," *Molecular Biology and Evolution*, vol. 24, no. 1, pp. 54–62, Jan. 2007.

Appendice

G. Burger, S. Moreira, and M. Valach. 2016
"Genes in Hiding." Trends in Genetics 32(9):553–565

Trends in Genetics

CellPress

Review

Genes in Hiding

Gertraud Burger,[1,*] Sandrine Moreira,[1] and Matus Valach[1]

Unrecognizable genes are an unsettling problem in genomics. Here, we survey the various types of cryptic genes and the corresponding deciphering strategies employed by cells. Encryption that renders genes substantially different from homologs in other species includes sequence substitution, insertion, deletion, fragmentation plus scrambling, and invasion by mobile genetic elements. Cells decode cryptic genes at the DNA, RNA or protein level. We will focus on a recently discovered case of unparalleled encryption involving massive gene fragmentation and nucleotide deletions and substitutions, occurring in the mitochondrial genome of a poorly understood protist group, the diplonemids. This example illustrates that comprehensive gene detection requires not only auxiliary sequence information – transcriptome and proteome data – but also knowledge about a cell's deciphering arsenal.

Trends

Omics technologies facilitate research into organisms beyond model systems, tapping into an underexploited wealth of information.

Combination of genomics, transcriptomics, and proteomics is a potent means for uncovering hidden genes and genetic elements, assigning function to genomic regions thought to be 'junk DNA'.

Unconventional gene structures promise to reveal innovative strategies and novel molecular mechanisms in gene expression, and thus to expand our toolbox for synthetic biology, genetic engineering, and molecular therapy.

The Dilemma of the Genomics Era

Current genomics technologies allow rapid sequencing of entire genomes, but detecting genes in the 'haystack' of DNA sequences remains a challenging task, because gene finding primarily relies on sequence similarity to known genes from a limited set of (model) species. An implication is that certain genes are overlooked, especially when they contain numerous **introns**, short exons, and atypical splice-junctions. The puzzle becomes even more complex when genes are disrupted by foreign DNA inserts, carry massive mutations, or are broken up in pieces. While genomicists struggle with incognito genes, these genes are accurately decoded in the cell. Understanding cellular deciphering strategies is an important asset in assigning function to yet 'vacant' genome regions.

In this review, we summarize scenarios that render genes cryptic and briefly examine cellular decoding schemes that ensure functional gene products. We then focus on a recently discovered system whose genetic information is hidden in an unparalleled way: by massive fragmentation and scrambling, as well as nucleotide substitutions and deletions. The molecular processes involved in deciphering these extremely 'mutilated' genes are intriguingly inventive, raising questions about how such a complex system may have emerged and why it has persisted during evolution. This review aims to make genomicists aware of unconventional gene structures encountered in nature, underscoring the importance of transcriptome and proteome data for a full exploration of genome information.

Types of Gene Encryption and Corresponding Decoding Strategies

A gene is called 'cryptic' when its sequence deviates substantially from that of its product. Deviations can take various forms (Table 1, Figure 1). Nucleotide substitutions have the potential to obscure a gene if replacements are abundant. The largest number (approx. 10%) and combinatorial diversity of substitutions are reported in dinoflagellate mitochondrial and chloroplast genes [1–3]. Similar high rates, but mostly cytidine to uridine replacements (C-to-U), occur in plant organelles [4]. In both cases, genes appear highly unusual, but are still recognizable. Nucleotide substitutions in these genes are corrected after transcription, a process known as RNA editing. Substitution RNA editing of organellar pre-mRNAs and pre-tRNAs is most studied in plants, certain amoebas, and fungi, showing that the molecular processes and enzymes

[1] Department of Biochemistry and Robert-Cedergren Centre for Bioinformatics and Genomics, Université de Montréal, Montreal, Canada

*Correspondence: gertraud.burger@umontreal.ca (G. Burger).

http://dx.doi.org/10.1016/j.tig.2016.06.005

employed by these groups are radically different from each other [5]. Note that, likewise, many animal nuclear transcripts undergo nucleotide substitutions [predominantly adenosine to **inosine** (A-to-I), but also C-to-U], but positions fall outside coding regions [6,7] or, alternatively, generate tissue-dependent protein isoforms [8], a subject area covered extensively in other reviews [9]. Finally, rectification of substitutions also occurs during translation. For instance, illegitimate in-frame stop codons in viral genes are reinterpreted as amino-acid codons through suppressor tRNAs [10].

Another encryption type is nucleotide insertion and deletion. **Indels** can severely conceal genes that encode proteins, because conceptual translation will shift the reading frame at indel sites, changing the downstream protein sequence, and usually introducing premature stop codons. Encryption by indels was first discovered in trypanosome mitochondria, where thymidines (Ts) are either missing or superfluous at certain gene positions. These 'errors' are fixed by RNA editing through an enzymatic machinery (editosome) that adds uridines (Us) in some places and removes Us in other places from the precursor transcript [11]. A much different system of indel elimination operates in mitochondria of the slime mold *Physarum*, in which repair coincides with RNA synthesis without producing pre-edited RNA [12]. Again, other indels are corrected at the translational level by programmed frameshifting, as frequently seen in viral genes. Frameshifting requires 'slippery' codons, RNA pseudoknot structures, frameshift-promoting tRNAs, and probably trans-acting factors [13–15].

Genes also become unrecognizable by insertion of mobile genetic elements that are either 'selfish' mobile endonucleases/transposases or elements of viral or bacteriophage origin. The most studied insert elimination is (posttranscriptional) intron splicing, operating through various mechanisms specific to the particular intron type (spliceosomal, Group I, Group II, and tRNA-introns). Removal at the protein level exists as well and applies to insertion elements referred to as inteins, which are transcribed and translated together with their host gene, and subsequently eliminated by protein splicing [16]. Typically, inteins do not obscure gene detection, but rather impede functional assignment, as they are usually inserted in highly conserved protein domains. In contrast, 'hops' (for hopping) and 'byps' (for bypassing) are a class of insertion elements that do conceal genes. The hop element resides in gene 6 of bacteriophage T4 and was described as a 'persistently untranslated sequence' [17], while byps are inserted *en masse* in mitochondrial protein-coding genes of the yeast *Magnusiomyces capitatus* [18]. Unlike introns, hops and byps are retained in mRNA, yet 'ignored' during translation, by ribosomes that have 'learnt' to bypass these elements [19].

Fragmented and scrambled genes are most difficult to detect in genome sequences. Genes in pieces prevail in the germline nucleus of certain ciliates, with fragments assembled at the DNA level in the working copy of the genome, the somatic nucleus. For example, the germline nucleus of *Oxytricha* contains thousands of scrambled genes; reordering and assembly of gene segments is guided by epigenetically inherited antisense RNAs [20]. Fragment joining at the RNA level is more widespread. Typically, gene breakpoints are within introns. Gene fragments are transcribed separately and fragment transcripts associate via RNA base-pairing that allows splicing by the cognate machinery in trans (reviewed in [21]). An unorthodox case of **trans-splicing** was reported recently in certain dinoflagellates, where one of the three mitochondrial protein-coding genes (*cox3*) was split into two pieces that were separately transcribed and then combined to a full-length mRNA by an unknown, intron-independent mechanism [22]. Another RNA-level strategy occurs in retroviruses, whose split genes are transcribed directly into contiguous mRNA via template switching of the RNA polymerase [23]. Finally, there are cases of fragmented genes that produce fragmented products, which, in turn, associate noncovalently. Examples include mitochondrial rRNA genes of diverse taxa (e.g., [24–26]), but also protein-coding genes [27,28], with pieces engaging in RNA–RNA, RNA–protein, or protein–protein interactions [29].

Glossary

Adenosine deaminase acting on RNA/tRNA (ADAR/ADAT): adenosine deaminase acting on (double-stranded) RNA/tRNAs, respectively.

Cassette: a short sequence region that is unique to a given chromosome in diplonemid mtDNA. The remaining sequence of the chromosome is shared with the other chromosomes of the multipartite genome.

Cis element: a sequence or secondary structure motif (e.g., of DNA or RNA) that acts on another region of the molecule by playing a regulatory or auxiliary role in a given molecular process (transcription, splicing, etc.). The counterpart is a trans factor (see later).

Constructive neutral evolution: a ratchet-like process that describes the evolution of complex systems by nonadaptive forces.

Diplonemids: a monophyletic group of unicellular, flagellated eukaryotes that are generally free-living. Diplonemids, their sister-group kinetoplastids, and euglenids form the Euglenozoa.

Deep sea pelagic diplonemids clade I (DSPDI) and DSPDII: recently discovered large clades of deep sea pelagic diplonemids, so far including exclusively noncultured species.

Genetic drift: a neutral force in species evolution proceeding by random segregation of genetic variants within a population.

Indels: insertions and deletions of nucleotides in a sequence.

Inosine: a nucleoside, commonly found in tRNAs, which is composed of hypoxanthine (i.e., hydrolytically deaminated adenine) and ribose.

Introns: Group I and Group II introns are ribozymes, characterized by distinctive 3D structures. Splicing involves transesterification. Splice-site consensus sequences are present but weak in Group I introns. Spliceosomal introns have a highly conserved splice-site consensus and are removed from pre-mRNA by several hydrolytic and trans-esterification steps that are performed by a large RNA–protein complex, the spliceosome. 'Archaeal' or 'tRNA' introns are characterized by a distinct secondary structure context. They are eliminated from the

In summary, cells decipher hidden genes by employing numerous and highly diverse molecular mechanisms and at any imaginable level of gene expression (Table 1, Figure 1). Posttranscriptional gene decryption is observed most frequently, which also applies to the recently discovered camouflaged genes in **diplonemid** mitochondria that we describe in the following section.

The Diplonemid Taxon

Diplonemids are unicellular, free-living flagellates found mostly in marine environments (Figure 2A, Key Figure). Traditionally, this group consists of two genera, *Diplonema* and *Rhynchopus* [30,31]. Diplonemids, together with their kinetoplastid sister group (including the well-studied human pathogens, trypanosomes and leishmania) and euglenids, form the Euglenozoa, which share a common ancestry with heteroloboseans (Figure 2B).

With only two genera recognized, diplonemids were considered an insignificant protist taxon. However, more recent environmental explorations revealed that diplonemids contain two additional, huge clades, **deep sea pelagic diplonemids clade I (DSPDI) and DSPDII**, composed exclusively of uncultured taxa [32]. In addition, a worldwide survey of marine plankton published last year ranked diplonemids among the most abundant and genetically most diverse eukaryotes in the oceans [33,34].

Unparalleled Mitochondrial Genome Architecture and Gene Structure in Diplonemids

Mitochondrial DNA (mtDNA) of the type species *Diplonema papillatum* encodes a conventional set of 12 recognized genes, including two ribosomal RNAs (mt-rRNAs) and protein components of the respiratory chain and oxidative phosphorylation. The genome also contains six unidentified reading frames, referred to as *y* genes (Figure 3A, left panel). Transfer RNAs appear to be imported from the cytosol as in kinetoplastids, *Euglena* [35], and several other eukaryotes. Despite their orthodox mitochondrial gene complement, the genome architecture and gene structure in diplonemid mitochondria is most eccentric and complex [36]. Unlike the majority of eukaryotes, which have a single mitochondrial chromosome, mtDNA of *Diplonema* consists of at least 80 distinct molecules (Figure 2 C) that are apparently able to replicate; that is, they are not formed as subcircles from a large master circle as in plant mitochondria [37].

Chromosomes are either 6-kb- or 7-kb-long circles (class A or class B, respectively). Approximately 9/10 of their length is identical in sequence between members of a given molecule class, and they mostly contain direct repeats. Only a short stretch of approximately 160–640 nt, termed a '**cassette**', is unique to each chromosome. The center piece of cassettes is a small coding region (42–534 nt) embedded between unique flanking sequences (50 nt on average) on both sides [38].

Cassettes are too short to hold most complete mitochondrial genes. Only tRNA genes would fit, but they are absent from *Diplonema* mtDNA. Instead, cassettes carry gene fragments referred to as gene **modules**. Genes are systematically broken down into multiple modules (up to 11), each of which resides on a different chromosome, with one exception: the tentative *rns* gene, specifying the mitochondrial small subunit ribosomal RNA (mt-SSU rRNA), is squeezed wholly into a single cassette, as *rns* is particularly short in this species.

All diplonemids investigated so far share this extravagant mitochondrial genome architecture and gene structure, with little variation [39]. Yet, as data are only available for the genera *Diplonema* and *Rhynchopus*, it remains to be investigated whether the two recently recognized clades, DSPDI and DSPDII [32,34], conform to the same scheme.

immature transcript by an endonuclease and RNA ligase.
Module: sequence region in mitochondrial chromosomes of diplonemids that codes for a part of a gene.
Protists: all eukaryotic groups except animals, fungi, and plants. These three latter groups emerged from protist lineages. Protists include single-celled and multicellular groups, and represent the bulk of eukaryotic diversity.
Ribozyme: catalytically active RNA, in contrast to (proteinaceous) 'enzyme'.
RNA chaperone: a molecule, typically protein that stabilizes, or helps proper folding, of RNAs.
Terminal uridylyl transferase (TUTase): an enzyme appending Us at the 3′ end of an RNA.
Trans factor: a molecule (e.g., RNA or protein) that acts on a separate molecule, by playing a regulatory or auxiliary role in a molecular process such as transcription, splicing, etc. The counterpart of trans factor is cis element.
Trans-splicing: covalent joining of RNA or protein pieces into a single molecule. The counterpart to trans-splicing is cis-splicing, which involves a single molecule from which an insert is removed by cleavage and resealing. Known splicing machineries perform both cis- and trans-splicing.
Untranslated regions (UTRs): untranslated regions at the 5′ and 3′ end of protein-coding genes.

Table 1. Types of Gene Encryption and Decryption Strategies[a]

Gene encryption type	Decryption at level of				
	DNA	RNA		Protein	
		(Cotranscriptional)[b]	(Posttranscriptional)	(Cotranslational)[b]	(Posttranslational)
Nucleotide substitution	Conceivable: nucleotide 'reversion'	Conceivable: incorporation of non-matching nt	Substitution-RNA editing [5]**	Mis-/non-sense suppressor tRNA [10]*	Conceivable: change of amino acid physicochemical properties by side-chain modification [69]
Nucleotide deletion[c]	Conceivable: nucleotide insertion	Insertion of non-templated nts [12]; template slippage*	Insertion-RNA editing [5]**	– 1 frameshift suppressor tRNA [15]*	Conceivable: amino acid insertion[c]
Nucleotide insertion[c]	Conceivable: nucleotide deletion	Skipping of nts in gene [12]*	Deletion-RNA editing [5]*	+ 1 frameshift suppressor tRNA [15]*	Conceivable: amino acid deletion[d]
Sequence deletion[c]	Sequence insertion	Template back slipping [70]*	Conceivable: retro-transposition	?	?
Sequence insertion[c]	Conceivable: sequence deletion	Skipping of a stretch [71]*	Intron cis-splicing [72]*	Ribosome hopping [19]*	Protein splicing[d] (intein) [16]*
Gene fragmentation[e]	Fragment joining [20]*	Transcriptional template switching [23]*	Trans-splicing [21]**	Translational template switching (tmRNA) [73]*	Intein trans-splicing [16]*. Conceivable: transpeptidase/sortase [74]
Gene overlap[f]	Conceivable: gene duplication and restauration of two separate genes	Separate/alternative transcription initiation and termination [75]*	Conceivable: alternative RNA cleavage[g]	Conceivable: translational stop and outframe reinitiation upstream[g]	Not applicable
Gene fusion	Conceivable: splitting of fusion	Conceivable: transcription stop and inframe reinitiation downstream	Conceivable: RNA cleavage	Conceivable: translational stop and inframe reinitiation downstream	Proteolytic cleavage of polyproteins, e.g., in viruses [76]*

[a] Observed cases (*); instances occurring in diplonemids (**); no instances observed as to our knowledge (no asterisk).

[b] 'Cotranscriptional', 'cotranslational' refer here to an event that occurs during incorporation of the respective nucleotides/residues (occasionally, these terms are also used to state that an event happens prior to the completion of mRNA or protein synthesis).

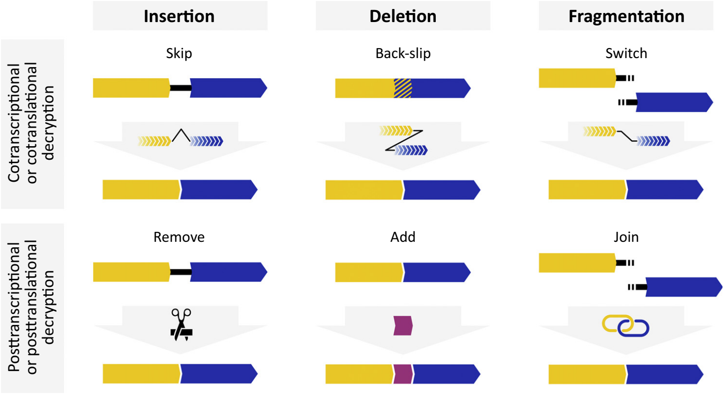
[c] 'Nucleotide' insertions and deletions as defined here are 1–3 nt long. 'Sequence' insertion and deletion refer to stretches of > 3 nt.

[d] This stratagem requires that the number of inserted/deleted nucleotides is a multiple of three.

[e] Fragmented genes with fragmented products are not considered in this table.

[f] Full or partial gene overlap on same or complementary strand.

[g] Strategy conceivable if overlapping genes are encoded on the same strand.


Insertion
Deletion
Fragmentation
Cotranscriptional or cotranslational decryption
Skip
Back-slip
Switch
Posttranscriptional or posttranslational decryption
Remove
Add
Join

Trends in Genetics

Figure 1. Major Types of Gene Encryption and Cellular Deciphering Strategies at the RNA and Protein Levels. Deciphering can act either during synthesis of RNA/protein or thereafter. When encryption is due to an insertion, deciphering involves either skipping of the inserted sequence during transcription/translation, or subsequent removal from the RNA or protein product. In the case of a deletion in the gene, decryption takes place either by cotranscriptional/cotranslational back-slipping, or the missing portion is inserted ('Add') subsequently in the product. When a gene is fragmented and pieces are spread over several physical locations, the information is put together either by template switching during RNA/protein synthesis or by synthesizing separate fragments of precursor transcripts/proteins with subsequent joining of the pieces.

Gene Fragments are Joined at the Transcript Level

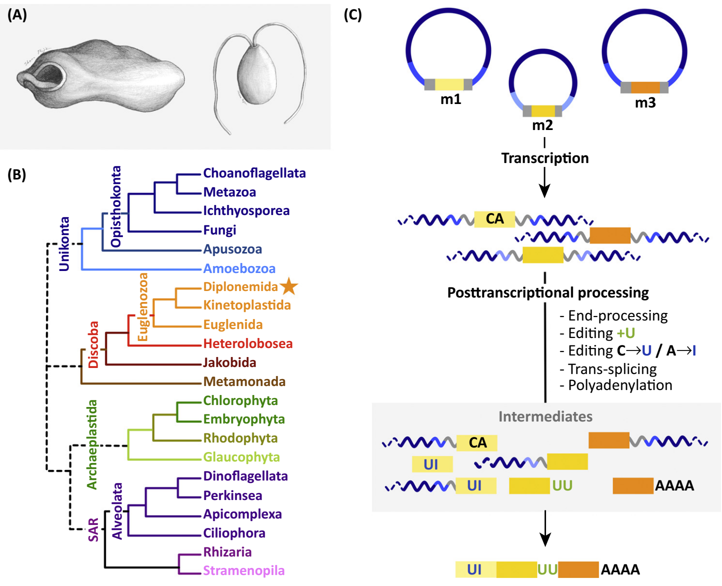
At the outset of the initial study on *Diplonema*, when only partial mtDNA sequence was available, a single gene module had been recognized, as it contains a highly conserved region of *cox1* (whose gene product is cytochrome oxidase subunit 1) [40]. Later on, high-coverage genome sequencing unveiled nearly all cassettes, but only transcript sequences allowed us to first pinpoint modules and then to assign the modules to their cognate genes [38,41].

Deep sequence data of the transcriptome further revealed RNA precursors at various stages, allowing us to reconstruct the transcript processing cascade. Gene modules are separately transcribed, together with long adjacent stretches reaching into the chromosome's constant regions (Figure 2C). Subsequent end-processing yields bare coding regions, except for modules that define the 5′ end of mRNAs and contain a short 5′ **untranslated region (UTR)**. In the following step, modules that will form a transcript's 3′ terminus are 3′-polyadenylated, except mt-SSU rRNA to which a few Us are attached. Diplonemids are among the few taxa whose mitochondrial mRNAs carry an A-tail. Finally, end-processed module transcripts are assembled into contiguous RNAs by a process we refer to as trans-splicing. Module joining appears to proceed in no particular order, starting simultaneously at multiple points in a given transcript and following various assembly paths, until completion [42].

Transcript assembly is astoundingly accurate. Less than 1% of the transcripts are mix-ups of modules from different genes. This raises the question of how a module chooses its correct bonding partner from among approximately 80 candidates. Bioinformatics analyses of modules and adjacent sequences did not reveal sequence or secondary structure motifs that would allow partner recognition [42]. Furthermore, elements archetypal for intron (trans-) splicing or **ribozymes** [hammerhead, hepatitis delta virus (HDV) ribozymes, etc.] are absent as well [39].

Key Figure

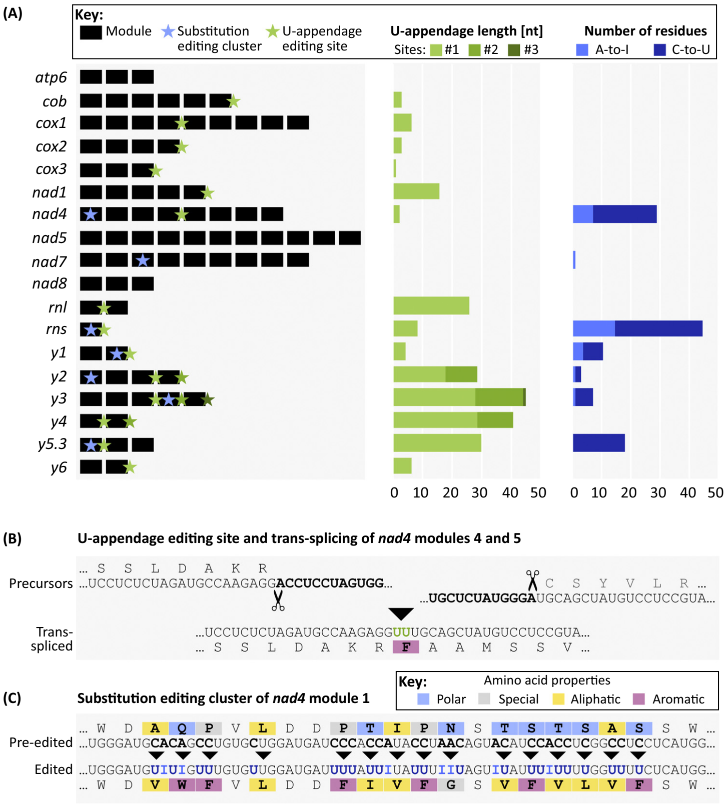
Diplonemid Phylogeny, Appearance and Mitochondrial Gene Expression



Trends in Genetics

Figure 2. (A) Appearance of the diplonemid *Rhynchopus euleeides* (drawing based on scanning microscopy, kindly provided by S. Teijeiro). Left, trophic state with large subapical flagellar pocket ('mouth'; the two flagella do not emerge). Right, a free-swimming 'swarmer' [77]. (B) Diplonemids' placement in the eukaryotic tree. The topology is taken from [78]. (C) mtDNA structure and gene expression. The constant region (dark blue) is common to the two classes of mitochondrial chromosomes. The class-specific regions (bright and pale blue) are adjacent to the cassette (light gray). Each cassette contains a gene fragment (module m1 to m3, yellow shades). Transcription starts and ends within the constant region, thus transcripts require extensive end processing. Substitution editing at clustered sites is illustrated for module 1, symbolized by the dinucleotide CA that is converted to UI. U-appendage RNA editing is shown for module 2, to which two Us are added at its 3′ end prior to trans-splicing. Module 3 undergoes polyadenylation.

In the apparent absence of **cis elements** in or adjacent to modules, we postulate match-making **trans factors** that bind cognate module partners and align them for covalent joining. An obvious candidate is antisense RNA, either multiple short molecules each serving an individual junction, or full-length transcripts reverse-complementary to mRNAs and rRNAs. Reverse transcription PCR (RT-PCR) and transcriptome data indicate short antisense RNAs, but only at low abundance and for a limited number of junctions [42,43]; these RNAs probably represent experimental artifacts or fortuitous transcripts. Conversely, trans factors could also be proteins. Promising matchmaker candidates are proteins that interact simultaneously with multiple RNAs, such as scaffold proteins assisting in ribonucleoprotein particle assembly [44].


(A)
Key:
Module
Substitution editing cluster
U-appendage editing site
U-appendage length [nt]
Sites: #1 #2 #3
Number of residues
A-to-I
C-to-U
atp6
cob
cox1
cox2
cox3
nad1
nad4
nad5
nad7
nad8
rnl
rns
y1
y2
y3
y4
y5.3
y6
0 10 20 30 40 50
0 10 20 30 40 50
(B) U-appendage editing site and trans-splicing of nad4 modules 4 and 5
... S S L D A K R
Precursors ...UCCUCUCUAGAUGCCAAGAGGACCUCCUAGUGG...
C S Y V L R ...
...UGCUCUAUGGGAUGCAGCUAUGUCCUCCGUA...
Trans-spliced ...UCCUCUCUAGAUGCCAAGAGGUUUGCAGCUAUGUCCUCCGUA...
... S S L D A K R F A A M S S V ...
(C) Substitution editing cluster of nad4 module 1
Key: Amino acid properties
Polar Special Aliphatic Aromatic
... W D A Q P V L D D P T I P N S T S T S A S S W ...
Pre-edited ...UGGGAUGCACAGCCUGUGCUGGAUGAUCCCACCAUACCUAACAGUACAUCCACCUCGGCCUCCUCAUGG...
Edited ...UGGGAUGUIUIGUUUGUGUUGGAUGAUUUUAUUIUAUUUIIUAGUIUAUUIUUUGGUUUUCUCAUGG...
... W D V W F V L D D F I V F G S V F V L V F S W ...

Trends in Genetics

Figure 3. Mitochondrial Gene Complement, Gene Fragmentation and RNA Editing Sites in *Diplonema papillatum*. (A) Modular structure of genes and positions of RNA editing sites. Gene products are: *atp6*, subunit 6 of ATP synthase; *cob*, apocytochrome b; *cox1-3*, subunits of cytochrome c oxidoreductase; *nad1–8*, subunits of NADH dehydrogenase; *rnl*, mt-LSU rRNA; *rns*, tentative mt-SSU rRNA. Gene products of *y1–y6* are unknown; *y1–y4* code for proteins. The charts indicate the number of appended Us per RNA editing site and the number of substituted A and C residues in a cluster. Note that for the gene *y5*, only the most abundant isoform (*y5.3*) is depicted. (B) Functional consequences of U-appendage RNA editing in the *nad4* module 2. The scissor symbol represents module end-processing. U-addition corrects the reading frame of the mRNA. (C) functional consequences of substitution RNA editing in the *nad4* module 1. Substitution RNA editing mainly converts codons for polar and special amino acids into codons specifying hydrophobic residues.

The other question is how cognate modules are joined, either by a proteinaceous RNA ligase or a trans-acting ribozyme. Eukaryotic cells contain several RNA ligase enzymes, and one of them could have been co-opted by diplonemid mitochondria for module-transcript ligation ([45] and references therein). Further candidates are RNA-ligase ribozymes that are able to ligate free RNA ends in trans, similar to those engineered recently by *in vitro* evolution [46]. Work is underway to unravel the biochemical mechanisms involved in *Diplonema* mitochondrial trans-splicing.

Insertion RNA Editing in *Diplonema* Mitochondria

A comprehensive comparison of the mitochondrial genome and transcriptome revealed that as much as three-quarters of the transcripts carry nucleotides not specified within the genome, exclusively uridines, summing to more than 220 posttranscriptionally added Us [41] (Figure 3A, right panel).

U-insertion (and deletion) RNA editing is well known from kinetoplastid mitochondria [11]. But instead of cleaving the precursor transcripts prior to insertion, Us in diplonemids are added to module 3′ ends prior to trans-splicing or polyadenylation (Figure 2C). One example is the *cox1* module-4 transcript, which is extended by six Us that are retained in the trans-spliced mRNA, marking the module-4/module-5 junction [36,39]. An even more spectacular U-appendage occurs during maturation of the mitochondrial large subunit ribosomal RNA (mt-LSU rRNA). A total of 26 Us are added in a row at the *rnl* module-1/module-2 junction prior to trans-splicing [43].

The appended Us fulfill a crucial biological role. In *cox3* (encoding cytochrome oxidase subunit 3), U addition together with polyadenylation generate the stop codon of the reading frame. In *cox1*, the added Us give rise to two additional residues in the protein sequence, filling positions that are present in homologs of essentially all other species. Amino acids generated by U addition (cUU-Leu, UUU-Phe, Ucg-Ser) restore the conserved protein three-dimensional (3D) structure, but with a twist: U-addition causes a polarity shift in the corresponding protein region of the *Diplonema* protein compared to Cox1 homologs. However, this shift is counterbalanced by a compensatory change in a downstream region that engages in a 3D-structure interaction with Leu-Phe-Ser [39]. In mt-LSU rRNA, the posttranscriptionally added U tract is thought to restore the molecule's secondary structure by reconstituting parts of the otherwise conserved helices 61 and 26a that are missing from the *Diplonema* gene [43]. Both examples strongly suggest that RNA editing of mitochondrial transcripts in *Diplonema* is function-critical.

For appendage RNA editing in mitochondria, Us appear better suited than other nucleotides. A homopolymer of U has weaker base-stacking properties than homopolymers of other nucleotides [47] and is thus structurally more amenable to folding into a particular shape to compensate deletions in structural RNAs. Furthermore, since UUU-codons specify the apolar residue phenylalanine, U addition can replace the loss of hydrophobic protein regions, which are predominant in mitochondrion-encoded genes.

At the moment, it is unknown which enzyme performs the U-appending activity in *Diplonema* mitochondria. It might be a protein related to the mitochondrial template-independent poly(U) polymerase [or **terminal uridylyl transferase (TUTase)**] RET2 from the trypanosome editosome [48]. Alternatively, it might be a relative of cytosolic TUTases involved in RNA maturation, turnover, and quality control [49].

Substitution RNA Editing in *Diplonema* Mitochondria

A comprehensive comparison of the *Diplonema* mitochondrial genome and transcriptome uncovered nucleotide substitution as a second type of posttranscriptional RNA editing in this system. Certain cytidines (Cs) are replaced by Us, well known from mitochondrial and plastid mRNA maturation, especially in plants [5], and certain adenosines (As) by inosines (Is), otherwise a hallmark of nuclear transcripts [50]. The latter substitution has not been seen in organellar mRNAs or rRNAs before (except one instance of an inosine derivative [51]), but is frequent in *Diplonema* mitochondria, with 29 A-to-I sites out of 114 substitutions (Figure 3A). It was demonstrated experimentally that As in precursor transcripts are indeed replaced by Is [41], pointing to an activity that performs *in situ* base deamination.

Like U-appendage, substitution RNA editing affects both mRNAs and rRNAs in diplonemid mitochondria. The mRNA with the largest number of substitution editing sites is *nad4* (subunit 4 of NADH dehydrogenase), where over a stretch of 55 nt in module 1, all Cs and half of all As are edited (Figure 3B). Editing renders the inferred protein sequence significantly more similar to Nad4 homologs from other eukaryotes.

Diplonema mitochondrial SSU rRNA is the first case of an rRNA containing Is (only one instance of an inosine derivative has been reported, notably in cytosolic LSU rRNA from a kinetoplastid [51]). Inosines in structural RNAs will have a considerable impact on the molecule's secondary structure, because this nucleotide not only pairs with C but also reasonably well with U, A, and G. As a consequence, the number of possible alternative RNA structures expands dramatically. We presume that folding of *Diplonema* mt-SSU rRNA into its functional shape therefore requires assistance by **RNA chaperones**, as do the A + T-rich trypanosome, yeast, and human mt-rRNAs, which are stabilized by pentatricopeptide repeat (PPR) proteins ([52] and references therein).

The postulated deaminase that catalyzes substitution RNA editing in *Diplonema* mitochondria might well be related to enzymes involved in nucleotide metabolism or base modification/editing of tRNAs. We favor the idea of an evolutionary link between this deaminase and the tRNA-editing family, **adenosine deaminases acting on tRNA (ADAT)**, because one of its members is able to deaminate both, As and Cs ([53] and references therein).

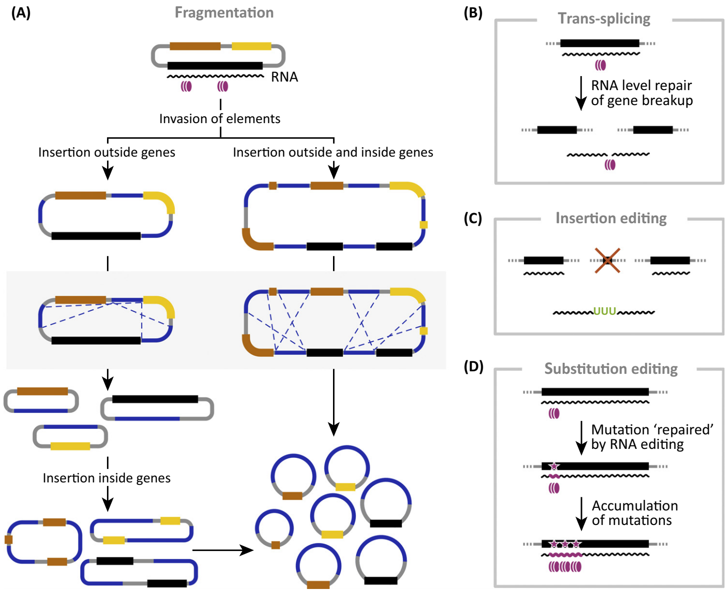
Conservatism versus Liberalism in Mitochondrial RNA Editing across Diplonemids

In phylogenetic trees based on mitochondrion-encoded proteins, *Diplonema* and *Rhynchopus* species are far apart from one another. For example, the phylogenetic distance between Cox1 proteins from these two diplonemids is as large as that between rhodophytes and fungi (see Figure 2B). This poses the question of how much RNA editing patterns are conserved across diplonemids. When comparing the positions and number of posttranscriptionally inserted Us in seven genes from four taxa (*D. papillatum, Diplonema ambulator, Diplonema* sp.2, *Rhynchopus euleeides* [39]), preservation of sites (together with the gene breakpoints) was observed throughout the four species. The situation for substitution RNA editing is the opposite. For example, when comparing *nad4* across diplonemids, C-to-U and A-to-I editing sites, although congregating in the same module, do not have a single position in common [41]. Apparently, U-appendage editing is maintained together with the modular gene structure, whereas substitution RNA editing diverges rapidly, as do gene sequences in diplonemid mitochondria.

Evolution of Multipartite mtDNA, Cryptic Genes and the Deciphering Machinery of Diplonemids

As described earlier, the mitochondrial genome architecture, gene structure and expression mechanisms in diplonemids are most unusual, raising questions about how this all has evolved from an ancestral genome that was most likely monopartite and densely packed with genes, like mtDNA of heteroloboseans (NC_002573.1 [54,55]; see Figure 2B). We speculate that the ancestral single-chromosome mtDNA has been invaded by a transposon or an intron-like mobile element. Alternatively, the mitochondrial replication origin may have undergone sudden and massive transposition [56], leaving multiple copies scattered across the genome (Figure 4A). Genome fragmentation is probably a consequence of recombination among these dispersed repeats, as proposed for a number of other eukaryotes [57–62].

In all cases except diplonemids, minichromosomes of organellar multipartite genomes carry (one or a few) complete genes (or no gene at all). If chromosomes do contain gene fragments, as for example in *Euglena*, dinoflagellate, and *Amoebidium* mtDNA, a complete gene version is always



Trends in Genetics

Figure 4. Hypotheses on the Origin and Emergence of Fragmented Genome and Genes, and RNA Editing in Diplonemid Mitochondria. (A) Genome fragmentation. Autogenous or foreign genetic elements invade the genome. Left-hand scenario, elements insert between ('outside') genes, followed by intrachromosomal recombination that leads to a multipartite genome with intact genes. Recombination among chromosomes may revert the genome into a single, yet rearranged chromosome. After genome fragmentation, elements also insert within genes. Finally, recombination between elements leads to single-module-bearing chromosomes. Right-hand scenario, the alternative path involves a single step where elements insert outside as well inside genes. (B–D) Posttranscriptional processes that may have arisen at any stage of genome/gene fragmentation. We postulate that sequence-specific RBPs serve as recognition factors in trans-splicing and RNA editing, assuming that breakpoints, deletions and substitutions in genes fall in the binding motif of a given RBP. RBPs likely recruit catalytic activities to perform module ligation, U-addition, or deamination. (B) Matchmaking in trans-splicing is thought to be performed by an RBP that is able to bind a sequence motif constituted by two molecules, thus bringing cognate transcript pieces together. (C) Nucleotide deletions may have arisen by loss of a micro-module-bearing chromosome. (D) Clusters of substitution sites may have emerged because a machinery responsible for RNA editing of one position would allow accumulation of additional mutations in its proximity. Black, yellow and orange boxes, genes; blue lines, insertion elements; dashed lines, recombination; undulated black lines, transcripts; purple-filled ovals, RNA binding proteins (RBPs).

present in one of the other chromosomes [24,61]. Similarly, the multipartite mtDNA of kinetoplastids contains, in addition to the 'maxicircle' on which (encrypted) protein-coding and rRNA genes reside, hundreds of 'minicircles'. The minicircle-encoded guide RNA genes, which direct U-insertions and deletions in precursor transcripts [11], are probable evolutionary derivatives of gene fragments [24]. The situation in diplonemids is different and unique: element insertion within genes must have been permissible thanks to a unique trans-splicing machinery, and may have occurred prior to or after mtDNA disintegration (Figure 4A).

The molecular processes that rescue fragmented and mutated genes in present-day diplonemids (Figure 4B–D) are startlingly sophisticated. Such complexity may seem a wasteful means of gene expression, challenging traditional thinking that seeks selective advantage in each evolutionary change. However, more recent theories advocate that selection is not the only driving force of evolution. Fixation of ‘unnecessarily’ complex processes may likewise arise by **genetic drift** (if the new traits are neutral or only marginally deleterious [63]). Further gradual complication occurs by ongoing ratchet-like trends, as postulated by the model of **constructive neutral evolution** [64–66].

The main feature of this latter theory is that solutions were already in place before the problem arose. In the context of hidden genes in *Diplonema*, the conventional ancestor most likely had pre-existing deaminases, TUTases, and RNA-binding proteins offering the opportunity for deployment in RNA editing and trans-splicing as soon as gene fragmentation and mutations occurred. We expect to gain intriguing insights into evolutionary trajectories by tracing back diplonemid's gene decryption arsenal to basic cellular machineries (e.g., those involved in DNA repair [67]).

Concluding Remarks and Future Perspectives

The example of diplonemid mitochondrial gene expression underscores two important points. First, available genome information is underexploited, as annotation procedures conservatively predict what is expected and known (see Outstanding Questions), thereby overlooking unconventional genes, unanticipated genetic elements, and novel molecular mechanisms. Obviously, detection of hidden genetic information needs data complementary to genome sequence, most importantly transcriptomics and proteomics data, but also knowledge about the diversity of gene structure and expression modes across living organisms, and a sense for nature's resourceful escapades.

The second point is that the remarkable gene structure and expression mode described above was discovered in a barely known and seemingly insignificant taxon of protistan eukaryotes. Since **protists** encompass the overwhelming portion of eukaryotic diversity, tapping this largely unexploited resource promises discovery of many more novelties, ranging from innovative molecular devices to inventive biochemical pathways and unsuspected biocompounds. Despite the heralded postgenomic era, genomics has yet to transit into adulthood.

Acknowledgments

We thank B.F. Lang and M. Sarrasin (Université de Montréal) for critically reading the manuscript. This work was supported by the Canadian Institute of Health Research (CIHR; operating grant MOP-79309 to G.B.).

Outstanding Questions

We are not aware of examples in nature where deletions in genes are compensated by posttranscriptional retrotransposition, or where split genes are restored posttranslationally by a transpeptidase. Could such reactions be engineered?

What is the biological role of the unassigned mitochondrial genes *y1–y6* in *D. papillatum* and are they conserved throughout diplonemids?

What is the extent of genome and gene fragmentation and RNA editing in the newly described diplonemid clades?

Diplonemid mt-rRNAs stand out for their short length, deviant secondary structure and inosine content. Is the reduction of structural rRNA features counterbalanced by proteins or additional RNA species or does the volume vacated by rRNA reduction remain unfilled as in mammalian mito-ribosomes [68]?

What kind of molecules are the hypothetical matchmakers in module trans-splicing: RNA, protein, or DNA? Once brought together, is ligation of module transcripts catalyzed by proteinaceous enzymes or rather by ribozymes?

How did nucleotide substitutions in diplonemid mitochondrial genes arise and why are C-to-U and A-to-I sites so massively clustered? Furthermore, are both deamination reactions catalyzed by the same enzyme?

Is U-tailing of diplonemid mt-SSU rRNAs performed by the same enzyme as U-appendage RNA editing of module transcripts? Does the postulated diplonemid terminal U-transferase share a common ancestry with the trypanosome TUTase involved in insertion RNA editing?

Since mitochondrial RNA editing in diplonemids is very different from that in kinetoplastids, might the machineries have been acquired horizontally rather than by vertical descent?

References

1. Mungpakdee, S. *et al.* (2014) Massive gene transfer and extensive RNA editing of a symbiotic dinoflagellate plastid genome. *Genome Biol. Evol.* 6, 1408–1422
2. Nash, E.A. *et al.* (2008) Dinoflagellates: a mitochondrial genome all at sea. *Trends Genet.* 24, 328–335
3. Dorrell, R.G. *et al.* (2016) Diversity of transcripts and transcript processing forms in plastids of the dinoflagellate alga *Karenia mikimotoi*. *Plant Mol. Biol.* 90, 233–247
4. Grewe, F. *et al.* (2011) A unique transcriptome: 1782 positions of RNA editing alter 1406 codon identities in mitochondrial mRNAs of the lycophyte *Isoetes engelmannii*. *Nucleic. Acids Res.* 39, 2890–2902
5. Knoop, V. (2011) When you can't trust the DNA: RNA editing changes transcript sequences. *Cell Mol. Life Sci.* 68, 567–586
6. Rosenberg, B.R. *et al.* (2011) Transcriptome-wide sequencing reveals numerous APOBEC1 mRNA-editing targets in transcript 3′ UTRs. *Nat. Struct. Mol. Biol.* 18, 230–236
7. Blanc, V. *et al.* (2014) Genome-wide identification and functional analysis of Apobec-1-mediated C-to-U RNA editing in mouse small intestine and liver. *Genome Biol.* 15, R79
8. Alon, S. *et al.* (2015) The majority of transcripts in the squid nervous system are extensively recoded by A-to-I RNA editing. *Elife* 4, e05198
9. Nishikura, K. (2016) A-to-I editing of coding and non-coding RNAs by ADARs. *Nat. Rev. Mol. Cell Biol.* 17, 83–96
10. Ling, J. *et al.* (2015) Genetic code flexibility in microorganisms: novel mechanisms and impact on physiology. *Nat. Rev. Microbiol.* 13, 707–721
11. Read, L.K. *et al.* (2016) Trypanosome RNA editing: the complexity of getting U in and taking U out. *Wiley interdisciplinary Rev. RNA* 7, 33–51
12. Gott, J.M. (2013) Mechanisms and functions of RNA editing in Physarum polycephalum. In *RNA Editing: Current Research and Future Trends* (Maas, S., ed.), Horizon Press
13. Dunkle, J.A. and Dunham, C.M. (2015) Mechanisms of mRNA frame maintenance and its subversion during translation of the genetic code. *Biochimie* 114, 90–96
14. Caliskan, N. *et al.* (2015) Changed in translation: mRNA recoding by -1 programmed ribosomal frameshifting. *Trends Biochem. Sci.* 40, 265–274
15. Atkins, J.F. and Bjork, G.R. (2009) A gripping tale of ribosomal frameshifting: extragenic suppressors of frameshift mutations spotlight P-site realignment. *Microbiol. Mol. Biol. Rev.* 73, 178–210
16. Gogarten, J.P. *et al.* (2002) Inteins: structure, function, and evolution. *Annu. Rev. Microbiol.* 56, 263–287

71. Sztuba-Solinska, J. *et al.* (2011) Subgenomic messenger RNAs: mastering regulation of (+)-strand RNA virus life cycle. *Virology* 412, 245–255

72. Papasaikas, P. and Valcarcel, J. (2016) The spliceosome: the ultimate RNA chaperone and sculptor. *Trends Biochem. Sci.* 41, 33–45

73. Giudice, E. *et al.* (2014) Trans-translation exposed: understanding the structures and functions of tmRNA-SmpB. *Front. Microbiol.* 5, 113

74. Spirig, T. *et al.* (2011) Sortase enzymes in Gram-positive bacteria. *Mol. Microbiol.* 82, 1044–1059

75. Shabalina, S.A. *et al.* (2010) Connections between alternative transcription and alternative splicing in mammals. *Genome. Biol. Evol.* 2, 791–799

76. Yost, S.A. and Marcotrigiano, J. (2013) Viral precursor polyproteins: keys of regulation from replication to maturation. *Curr. Opin. Virol.* 3, 137–142

77. Roy, J. *et al.* (2007) Description of *Rhynchopus euleeides* n. sp. (Diplonemea), a free-living marine euglenozoan. *J. Eukaryot. Microbiol.* 54, 137–145

78. Derelle, R. *et al.* (2015) Bacterial proteins pinpoint a single eukaryotic root. *Proc. Natl. Acad. Sci. U.S.A.* 112, E693–E699