RNA BIOLOGY
2016, VOL. 0, NO. 0, 1–8
http://dx.doi.org/10.1080/15476286.2016.1240143

Taylor & Francis
Taylor & Francis Group

POINT OF VIEW

Post-transcriptional mending of gene sequences: Looking under the hood of mitochondrial gene expression in diplonemids

Matus Valach [a], Sandrine Moreira [a], Drahomíra Faktorová [b], Julius Lukeš [b,c], and Gertraud Burger [a]

[a]Department of Biochemistry and Robert-Cedergren, Center for Bioinformatics and Genomics, Université de Montréal, Montreal, Canada; [b]Institute of Parasitology, Biology Center and Faculty of Sciences, University of South Bohemia, České Budějovice, Czech Republic; [c]Canadian Institute for Advanced Research, Toronto, Canada

ABSTRACT

The instructions to make proteins and structural RNAs are laid down in gene sequences. Yet, in certain instances, these primary instructions need to be modified considerably during gene expression, most often at the transcript level. Here we review a case of massive post-transcriptional revisions *via* trans-splicing and RNA editing, a phenomenon occurring in mitochondria of a recently recognized protist group, the diplonemids. As of now, the various post-transcriptional steps have been cataloged in detail, but how these processes function is still unknown. Since genetic manipulation techniques such as gene replacement and RNA interference have not yet been established for these organisms, alternative strategies have to be deployed. Here, we discuss the experimental and bioinformatics approaches that promise to unravel the molecular machineries of trans-splicing and RNA editing in *Diplonema* mitochondria.

Abbreviations: A-to-I, adenosine to inosine; C-to-U, cytidine to uridine; HMM, hidden Markov model; MS, mass spectrometry; mtDNA, mitochondrial DNA; mt-SSU rRNA, mitochondrial small subunit rRNA; mt-LSU rRNA, mitochondrial large subunit rRNA; RNA, ribonucleic acid; ADAR, adenosine deaminase acting on RNA; ADAT, adenosine deaminase acting on tRNA; APOBEC, apolipoprotein B mRNA editing catalytic component; REL, RNA editing ligase; TUTase, terminal uridylyl transferase

ARTICLE HISTORY
Received 31 August 2016
Accepted 20 September 2016

KEYWORDS
Cryptic genes; diplonemids; gene fragmentation; multipartite mtDNA; RNA editing; trans-splicing

Introduction

Genomes tell only part of the story: Types of post-transcriptional amendments

The living cell manufactures its own building blocks by carrying out the instructions specified by its genes. Yet, the final gene product, whether a protein or a structural RNA, is not always the literal implementation of its blueprint. Alterations of the genetic information can occur at any stage of the manufacturing process: during and after transcription into RNA, or during and after translation for protein-coding genes.[1] Changes that take place post-transcriptionally are most frequently observed in nature. These include nucleotide substitutions, insertions or deletions (known as RNA editing[2]), and removal of larger stretches (e.g. during intron splicing) or insertions of a mobile genetic element (e.g., by retrotransposition). In the following, we will review a recently discovered system with massive and diverse post-transcriptional processes that are associated with a remarkably jumbled mitochondrial genome as observed in diplonemids.

From insignificance to fame

Diplonemids (Euglenozoa, Excavata) are the sistergroup of kinetoplastids that include the human-pathogenic trypanosomes and leishmanias.[3,4] Being 'innocent' free-living ocean creatures with only two recognized genera (*Diplonema* and *Rhynchopus*), diplonemids had been considered an insignificant taxonomic group –until not long ago. Recent world-wide ecological studies revealed that diplonemids are one of the genetically most diverse, most cosmopolitan and most abundant eukaryotic groups in the sunlit ocean.[5-9]

The eccentric mitochondrial genome of Diplonema

Following a preliminary description by Simpson's group,[3] the mitochondrial DNA (mtDNA) of the type species *Diplonema papillatum* has now been characterized in detail. The gene content is quite conventional, specifying proteins of the respiratory chain and oxidative phosphorylation, and the ever-present large and small-subunit rRNAs, mt-LSU and mt-SSU rRNA[10,11] (Table 1). Genome architecture and gene structure, however, are most unusual (Fig. 1A). The genome, estimated at ~600 kbp, consists of at least 81 distinct circular chromosomes (present in multiple copies) that fall into two classes based on molecule size (6 and 7 kbp).[12] About 95% of a chromosome's length shares its sequence with that of the other members of its class, thus only 5% of its sequence is distinctive. This unique region, referred to as 'cassette', includes coding sequence, notably a piece ('module') of a gene.[13] All genes except *rns* (specifying mt-SSU rRNA) are fragmented in up to 11 pieces, and each module (40–550 nt long) resides on its own chromosome (Table 1).

CONTACT Matus Valach matus.a.valach@gmail.com; Gertraud Burger Gertraud.Burger@umontreal.ca

Table 1. Mitochondrial gene complement, gene structure and RNA editing sites in *D. papillatum*.

Gene	Gene product	Number of modules	Number of editing sites Deaminations	U-additions (length)
atp6	ATP synthase subunit 6	3	0	0
cob	Apocytochrome b	6	0	1 (3 nt)
cox1	Cytochrome oxidase subunit 1	9	0	1 (6 nt)
cox2	Cytochrome oxidase subunit 2	4	0	1 (3 nt)
cox3	Cytochrome oxidase subunit 3	3	0	1 (1 nt)
nad1	NADH dehydrogenase subunit 1	5	0	1 (16 nt)
nad4	NADH dehydrogenase subunit 4	8	29	1 (2 nt)
nad5	NADH dehydrogenase subunit 5	11	0	0
nad7	NADH dehydrogenase subunit 7	9	1	0
nad8	NADH dehydrogenase subunit 8	3	0	0
rnl	mt-LSU rRNA	2	0	1 (~26 nt)
rns	mt-SSU rRNA	1	45	1 (8 nt)
y1	Protein of unknown function	2	11	1 (4 nt)
y2	Protein of unknown function	4	3	2 (18 nt; 11 nt)
y3	Protein of unknown function	5	7	3 (~27 nt; 17 nt; 1 nt)
y4	Protein of unknown function	2	0	2 (~28 nt; 12 nt)
y5.3	Product of unknown function	3	18	1 (>30 nt)
y6	Product of unknown function	2	0	1 (6 nt)

Due to the extensive fragmentation combined with sequence divergence, not a single gene was recognized initially in the mitochondrial genome sequence of *Diplonema*.[1] Only transcriptome data revealed the coding content of mtDNA, as well as the post-transcriptional processes involved in building functional products from the fragmented genes.

Post-transcriptional decoding of encrypted mitochondrial genes in *Diplonema*

Trans-splicing of fragmented genes

The first question addressed by mitochondrial transcriptome sequencing was how *Diplonema*'s fragmented genes lead to contiguous proteins and structural RNAs. We found that gene pieces are transcribed separately together with long adjoining regions on the corresponding chromosome (Fig. 1B).[14] These precursor RNAs are then processed to yield short module transcripts. Subsequently, modules belonging to a given gene are joined (trans-spliced) to form a full-length mRNA or rRNA.[10,12,13] A plethora of intermediates are observed: single and multi-module transcripts with precursors either fully end-processed, or still carrying flanking non-coding sequence. What exactly assures the accurate joining of cognate modules remains unclear, since recurrent sequence or secondary structure motifs in or adjacent to modules appear to be absent.[15,16]

Insertion and substitution RNA editing

Trans-splicing is not the only surprising post-transcriptional processing step in *Diplonema* mitochondria. Analysis of transcript sequences also revealed that gene expression involves RNA editing (Fig. 1B). Most noticeable are insertions of non-encoded uridines (Us) in mRNAs and rRNA. First detected in

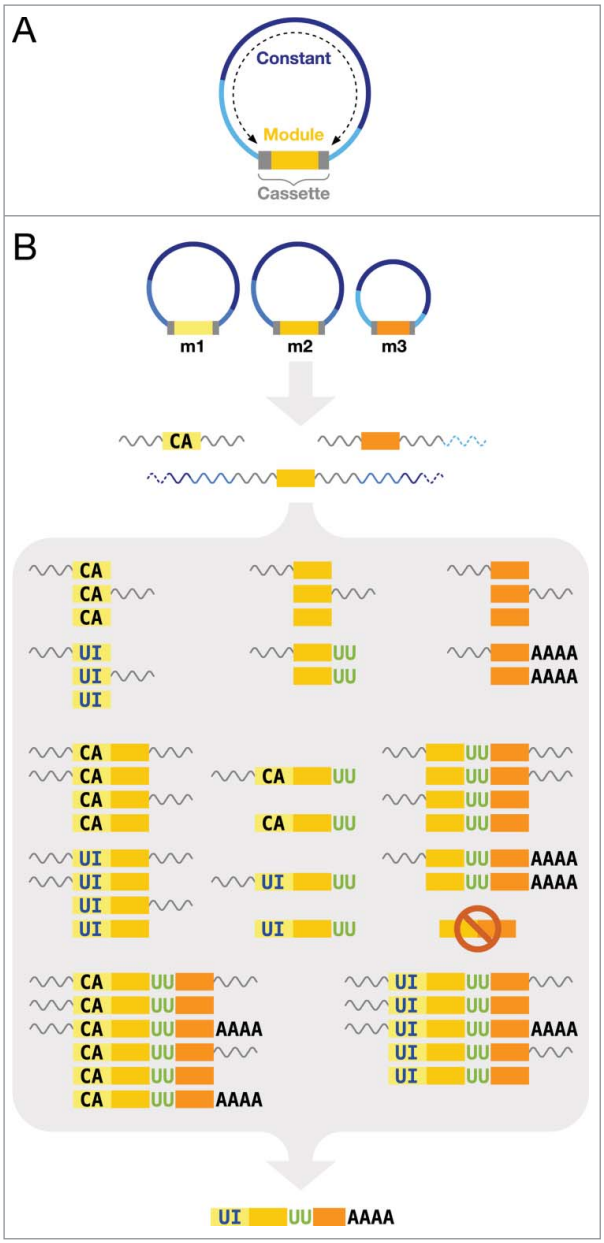

Figure 1. Genome architecture, gene structure, and RNA processing in *D. papillatum* mitochondria. (A) Canonical structure of mitochondrial chromosomes composed of a unique cassette that includes a gene fragment (module) and a constant region whose sequence is shared between all members of the respective chromosome class (A, or B, see text). (B) From gene to transcript. The depicted gene consists of 3 modules, which are all transcribed separately. The primary transcript includes long stretches of the constant region. Primary transcripts are processed in various steps proceeding in parallel: removal of non-coding regions 5′ and 3′ of modules, substitution and U-appendage RNA editing of select modules, polyadenylation of the terminal modules, and joining of module ends that are fully processed (gray background). Us are added prior to trans-splicing; U insertions into joined modules have not been observed (as indicated by crossing out).

cox1[13] (see Table 1) and initially considered a rare event, U insertions are now known to occur in about 80% of all mature mitochondrial transcripts, always at module boundaries or the transcripts 3′ ends. Among the most extreme cases is mt-LSU rRNA with ~26 Us intercalating between modules 1 and 2 (Fig. 2).[17] We demonstrated that Us are appended to the module upstream of the junction, before this terminus engages in trans-splicing.[14]

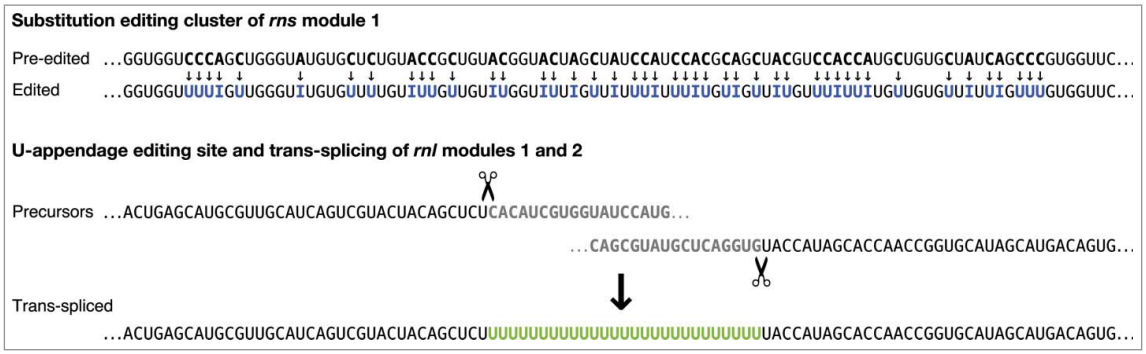

Figure 2. RNA editing types in *Diplonema* mitochondria. Top, C-to-U and A-to-I substitutions in module 1 of *rns* specifying mt-SSU rRNA. Edited nucleotides are depicted in blue. Bottom, U-appendage RNA editing between modules 1 and 2 of *rnl* specifying mt-LSU rRNA. After 3′ end processing (symbolized by scissors) of module 1, about 26 Us are attached to this terminus. Module 2 with an end-processed 5′ end is then joined with the U-tail of module 1.

U-based RNA editing occurs also in trypanosome mitochondria, where sometimes more than half of a transcript sequence is created via post-transcriptional U insertions and deletions (indels).[18,19] Yet, in *Diplonema*, U deletions were not observed (nor addition/removal of other nucleotides),[11] nor the cut/reseal strategy seen in trypanosomes.[14]

A second type of RNA editing occurs in *Diplonema* mitochondria –nucleotide substitutions– and affects about one third of the transcripts. We observe adenosine-to-inosine (A-to-I) and cytidine-to-uridine (C-to-U) changes.[11] These types of substitutions indicate *in situ* deamination of nucleotides in the transcripts. C-to-U exchanges are frequent in organelles, especially mitochondria and chloroplasts from plants[2], while A-to-I editing is a first in organellar non-tRNA transcripts.

Substitution editing sites in *Diplonema* mitochondria congregate in a few defined regions and are narrowly spaced. For example, *nad4* (encoding NADH dehydrogenase subunit 4) contains nearly 30 such sites packed in a 55-nt long stretch, and *rns* has 45 sites in an 85-nt long segment (Fig. 2). Again, there are no recognizable cis-motifs specifying the insertion or substitution editing sites.[11]

Both RNA editing types are crucial for mitochondrial function in *Diplonema* and make the transcript more similar to homologs of other eukaryotes. In protein-coding genes for instance, U-appendage rectifies the reading frame, or alternatively adds codons missing in the *Diplonema* gene, while nucleotide substitutions change codon identity.[11]

The above-depicted eccentric features of *D. papillatum* mitochondria are found as well in 3 other diplonemids,[15,20-22] and in *Hemistasia phaeocysticola*,[23] which previously has been placed within Kinetoplastida, but appears to be more closely related to diplonemids.[24]

Methods that turned out useful in characterizing *Diplonema*'s mitochondrial genome and transcriptome

The study of RNA editing and trans-splicing in *Diplonema* mitochondria was far from straight-forward, because many assumptions underlying conventional procedures did not apply to this system. Breaking the secrets of *Diplonema* mitochondria required specific adaptations or even custom approaches described in the following sections.

Experimental/biochemistry approaches

Diplonema mitochondria are not discrete, tiny organelles, but rather a large fragile network that lines the inside of the cell membrane.[12] Therefore, a dedicated cell disruption procedure had to be developed that efficiently breaks the cells, but keeps the delicate organelle intact. As a test for mitochondrial intactness, we used citrate synthase activity, which is catalyzed by a soluble matrix enzyme that easily escapes perforated mitochondria. Cell disruption by the nitrogen cavitation method[25] was a breakthrough. This method (used before for trypanosomes[26,27]), sets a cell suspension under high gas pressure and then releases the pressure abruptly exerting physical stress on the cells. Obviously, parameters such as treatment time, gas pressure, and buffer osmolarity had to be optimized as for any other species.

Not all experiments required isolated mitochondria. For example, a large portion of *Diplonema*'s mtDNA was sequenced with material separated from nuclear DNA by CsCl-density centrifugation.[12,28] Mitochondrial DNA forms a band that (unlike in most other systems) is more G+C-rich and more prominent than nuclear DNA. Another, simpler method is phenol-extraction of whole-cell lysates at low pH, whereby the small closed DNA circles behave like RNA.[10] Yet, since the *Diplonema* mitochondrial genome is multipartite, we also extracted DNA from whole mitochondria, to assure that chromosomes departing from the majority G+C content or topology are not overlooked.

Finally, *Diplonema* belongs to the few taxa whose mitochondrial mRNAs (and the mt-LSU-rRNA) are poly-adenylated.[3] Therefore, we enriched mitochondrial transcripts by oligo(dT) purification, which facilitated their characterization via targeted RT-PCR[10,13,14] and transcriptomics.[11,17]

In silico approaches

A standard procedure in eukaryotic genomics is to generate reads by whole-genome/random-fragment approach and assemble these reads (ideally yielding chromosome-size contigs). Mapping of transcriptome (RNA-Seq) reads against the genome locates genes and introns and allows one to spot potential polymorphisms and RNA editing sites. A plethora of bioinformatics tools exist for performing these steps, but many of them failed when applied to the *Diplonema* mitochondrial genome.

For example, assembly engines are not capable of reconstructing genomes with such long and abundant repeats as in *Diplonema* mtDNA. Initial assemblies resulted in hybrid contigs that consisted of repeat-containing reads from different chromosomes, numerous contigs much smaller than chromosomes, and contaminations with nuclear sequences. The few completely assembled chromosomes were obtained from libraries made from individual circles,[13] but obviously, this is not an efficient way of characterizing a genome consisting of several dozen distinct chromosomes.

As complete assembly of the *Diplonema* multi-partite mtDNA is too difficult and costly, we chose to focus on cassettes. The corresponding contigs were pulled out from an assembly by their characteristic constant regions that flank both sides of the cassettes. Yet, since *Diplonema* mitochondrial genes are highly derived and modules are short, the exact start and end of modules within cassettes was difficult to recognize in the genome sequence alone. Precise location of gene modules and assignment of modules to genes required transcript information, which again was challenging to obtain.

We attempted *de novo* transcriptome assembly (without a reference genome) using several tools (e.g. refs[29,30] and rnaSPAdes (http://bioinf.spbau.ru/en/rnaspades)), each time resulting in a large and confusing array of contigs corresponding to a multitude of intermediates from end-processing, trans-splicing and RNA editing, intermingled with mature transcripts. Still, mitochondrial transcripts could be retrieved from these assemblies. Some were tracked down by simple BLAST comparison of the conceptual translation with known mitochondrial proteins; others by more sensitive searches using hidden Markov Model (HMM) profiles[31] that we built from a collection of homologous mitochondrial proteins from excavates. Mitochondrial rRNAs, because of their excessive sequence divergence, escaped detection by HMM and covariance analysis.[32,33] Only comparative manual secondary structure modeling and biochemical approaches revealed these transcripts in the end.[11,17]

To validate inferred intron splice-junctions, researchers often map RNA-Seq reads back to the genome (e.g., using Bowtie[34] or STAR[35]). However, when applied to *Diplonema*, these tools retrieved only a handful of junctions due to the assorted transcript population. We therefore developed an in-house tool that maps paired-end RNA-Seq reads to the genome in local mode, and then scrutinizes not-aligning portions of reads (soft-clipped regions) to define module neighbors and exact junctions.[9]

Detection of RNA editing in *Diplonema* mitochondria posed hurdles as well. The RNA-to-DNA mapping step of classical tools discarded reads containing clustered edited sites, because the mapping quality (which depends on the number of mismatches per length) was below the default threshold. On the other hand, lowering the mapping-quality threshold or employing tools (such as ref.[36]) that use a degenerated alphabet, for example Y for T or C and R for A or G, returned many false positives. Therefore, Stadler's group (University of Leipzig) kindly adapted their read mapper, originally designed for bisulfite sequencing[36], for *Diplonema*. The tool segemehl_Diplonema, which displays superior performance, employs a reduced alphabet during the alignment step –not the seeding step– and in addition filters alignment quality based on accuracy, minimum read length and minimum error-free length for the seed.

Hypotheses on the trans-splicing and RNA editing machineries in *Diplonema* mitochondria

The processes of RNA editing and trans-splicing in *Diplonema* mitochondria are now fairly well characterized, but we do not know the players. The proteins involved are almost certainly encoded by the nuclear genome, translated in the cytoplasm and imported into the mitochondrion, as is the predominant portion of the mitochondrial proteome in all other eukaryotes.[37] Further, we postulate that RNA editing and trans-splicing are catalyzed by proteinaceous enzymes. Specifically, substitution RNA editing is probably performed by homologs of RNA- or tRNA-specific adenosine deaminases (ADAR and ADAT) or the apolipoprotein B mRNA editing enzyme (APOBEC), which are involved in A-to-I and C-to-U editing, respectively, of animal nuclear mRNAs and tRNAs,[38] or by enzymes of the RNA metabolism. U-appendage RNA-editing, in turn, might involve a homolog of terminal uridylyl transferase 2 (TUTase 2) participating in kinetoplastid indel RNA editing, or potentially other terminal transferases.[39] Finally, trans-splicing might be catalyzed by an RNA Editing Ligase (REL)-like RNA enzyme (the type that reseals RNA indel editing sites in kinetoplastid mitochondria), or alternatively one of the RtcB-type ligases that join the ends of tRNA halves after intron removal.[40]

How would these enzymes recognize the sites to be edited and the modules to be trans-spliced? In the apparent absence of cis-motifs, we suggested earlier the existence of trans-factors that bind to the corresponding transcript regions, and coopt an enzyme that catalyzes the required reaction (Fig. 3). Our first guess for trans-factors was RNAs.[13] However, molecules resembling RNA-editing guides from trypanosome mitochondria are not detectable in *Diplonema*,[14] and searches for RNA trans-splicing guides have been inconclusive.[14,15] Trans-factors may also be proteins such as the PentatricoPeptide Repeat (PPR) proteins in plant organelles.[41] Given the various possibilities of directors and actors involved, strategies to test these hypotheses should be based on a minimum of preconceived expectations.

How to detect and dissect the RNA editing and trans-splicing machineries?

Identification of the molecular machineries that conduct RNA editing (the "editosome") and trans-splicing (the "joinosome") in *Diplonema* mitochondria will require *in silico* and experimental approaches. The methods that appear feasible in this

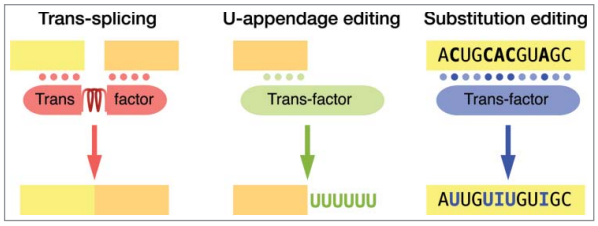

Figure 3. Hypotheses on RNA editing and trans-splicing mechanisms in *Diplonema* mitochondria. We postulate that trans-splicing, substitution and U-appendage RNA editing are guided by trans-acting factors that align cognate modules, direct U-additions, and specify deamination sites, respectively. These factors are probably proteins or RNAs.

system and that we envisage to employ are summarized in Fig. 4. A more detailed description follows in the sections below.

Bioinformatics approaches to identify proteins involved in RNA editing and trans-splicing

As a first step, we will screen the *Diplonema* nuclear genome for genes with the propensity to bind RNA. The PFAM database provides HMMs for characterized protein domains such as PPR and zinc-finger motifs. In a second step, we will seek domains characteristic of enzymatic activities that we postulate to act in RNA editing and trans-splicing. Proteins carrying a domain of interest will then be scrutinized for mitochondrial target signals. Predicted mitochondrial localization will add supporting evidence for a candidate.

So far, we have examined RNA ligase candidates. Among the *Diplonema* nuclear genes carrying RNA/DNA ligase domains, the hypothetical protein DpRNL is a valid candidate for catalyzing trans-splicing. Domain composition and features of the catalytic domain make it a member of RNA ligases 2, an assignment corroborated by 3D structure modeling, molecular dynamic simulations and phylogenetic analyses.[42] Yet, whether or not DpRNL performs mitochondrial trans-splicing remains to be demonstrated experimentally. Although bioinformatics is rarely able to convincingly predict a molecule's specific biological role, it is a powerful means to prioritize candidates for experimental validation.

Experimental approaches

In model systems, elegant methods are available for validating the predicted function of a protein. For example, gene manipulations add tags to proteins, which allow quick affinity purification and testing of the catalytic activity *in vitro*. Alternatively, targeted genes in the genome are inactivated for probing the consequence *in vivo*. These methodologies are not yet available for *Diplonema*. Still, there are a number of alternative ways to experimentally investigate the mitochondrial editosome and joinosome in this protist. The only prerequisite is an efficient protocol for purifying intact mitochondria in sufficient amounts, which is now in place.

Several strategies seem promising. One is to isolate mitochondrial macromolecular complexes, and analyze their protein components by mass spectrometry; candidates for the editosome and joinosome are complexes that include proteins with the postulated catalytic domains. A second avenue is to seek out proteins that physically interact with the substrates for RNA editing or trans-splicing, and then to dissect the corresponding complexes. Yet another is the binding of a co-factor that is indicative of an enzyme class of interest, for instance ATP and GTP in the case of RNA ligases. The ultimate validation demonstrates the postulated catalytic activity, i.e. ligation, U-addition, or nucleotide deamination of synthetic RNA substrates. In the following section, we will look in more detail into approaches capitalizing on these aspects in the detection of the postulated RNA editosome and joinosome in *Diplonema* mitochondria.

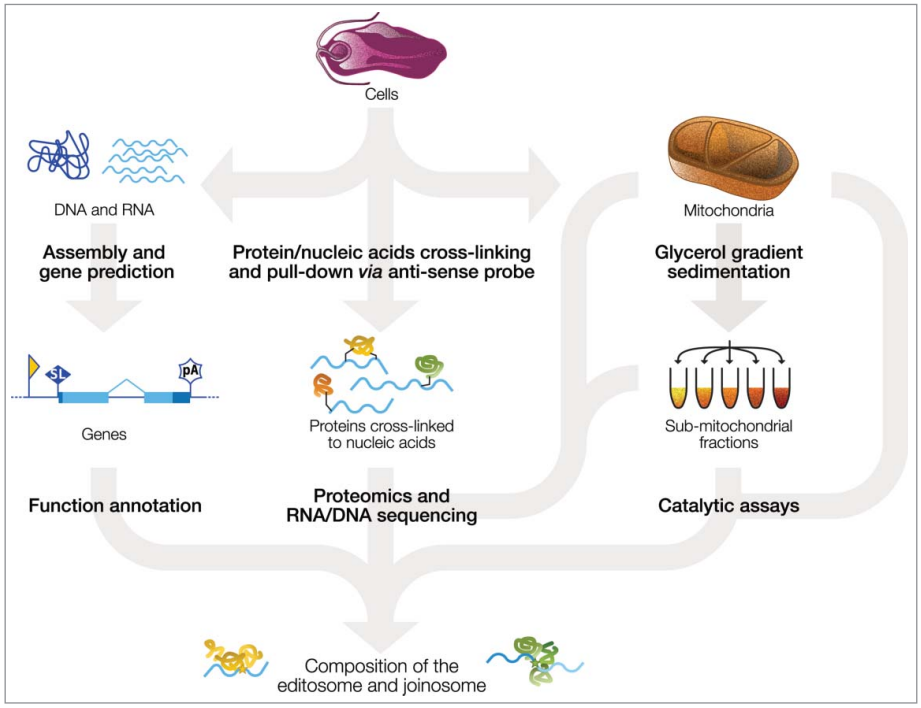

Figure 4. Envisaged experimental approaches for dissecting the RNA editing and trans-splicing machineries of *Diplonema* mitochondria. (Left section) Genomic DNA and mRNA isolated from *Diplonema* cells are sequenced and assembled. Predicted genes are then functionally annotated by similarity searches, domain prediction, and subcellular localization prediction to filter out nuclear gene candidates. The flag represents the transcription start site; SL, spliced-leader attachment site; pA, polyadenylation site. (Central section) Starting from whole cells, proteins are cross-linked (by UV light or chemically) to their mtRNA substrates and then selectively pulled down from lysates by biotin-labeled probes, which are anti-sense to a subset of mtRNA processing intermediates. The prey are analyzed by mass spectrometry (MS) and high-throughput nucleic acids sequencing. (Right section) Purified mitochondria are fractionated on a glycerol gradient and assayed for catalytic activities associated with RNA editing and joining. MS and RNA/DNA sequencing are used to examine the components of the mitochondrial lysate, as well as its fractions. All approaches converge on genes whose products are potentially involved in substitution or U-appendage RNA editing or in trans-splicing.

Proteomics analysis

Tandem mass spectrometry (MS) is a method of choice in the study of RNA editing and trans-splicing machineries in *Diplonema* mitochondria. Since the proteins of interest will be most likely present in much lower proportion than mitochondrial metabolic enzymes, whole organelle lysates will have to be fractionated by rate zonal sedimentation, generally on a glycerol density gradient. The resulting submitochondrial fraction can be further subdivided by electrophoretic separation of mitochondrial complexes under native and denaturating conditions.[43,44] MS analysis of such a material determines which of the candidate proteins encoded by the nuclear genome are indeed located in the mitochondrion. In addition, analysis of isolated complexes informs about the partners with which a given protein candidate is associated, directly or indirectly. In the case of *Diplonema*, this opens a way to explore potential RNA or DNA factors implicated in RNA editing and trans-splicing. Combination of MS and nucleic acid sequencing has been used with success in dissecting the composition of (ribonucleo)protein complexes such as the mitochondrial ribosome,[45] the kinetoplastid editosome[46] and the conventional spliceosome.[47]

Capture of RNA-binding proteins

Proteins involved in *Diplonema*'s mitochondrial RNA editing and trans-splicing are expected to bind to their RNA substrate, i.e., the pre-edited transcripts and mono-module transcripts, respectively. The use of RNA baits for protein pull-down comes to mind, but runs the risk of non-specific interactions that occur secondarily during experimental manipulations. One method to preserve native associations is *in vivo* covalent RNA-protein cross-linking, generally by UV. The RNA-protein couple is then pulled down via hybridization to an oligonucleotide probe whose sequence is reverse-complementary to the RNA substrate, and the protein partner is analyzed by MS.[48-50]

A similar approach probes for the enrichment of diagnostic RNA processing intermediates across mitochondrial density gradient fractions. For example, an above-average ratio of trans-splicing intermediates *versus* mature transcripts might signal that the corresponding density fraction is also enriched in the hypothetical joinosome.

Co-factor binding as a proxy for catalytic activity

Several large families of RNA ligase enzymes require distinct co-factors.[51-53] Certain RNA ligases self-adenylate during their catalytic cycle, for example the prototypical T4 RNA ligases and the editosomal ligases of kinetoplastids. Other RNA ligases such as RtcB and ligT use GTP as a co-factor. Therefore, submitochondrial fractions can be screened by incubation with radioactive ATP or GTP. Fractions that incorporate the labeled nucleotides will likely contain the corresponding RNA ligases.

Isolation of catalytically active macromolecular complexes

The conceptually most straightforward approach toward identification of the postulated editosome and joinosome in *Diplonema* mitochondria is to assay the conversion of synthetic RNA substrates into expected products. Sub-mitochondrial fractions can be assayed in this way. Specifically, substrate 3′ uridylylation, deamination of As and Cs, or joined substrates would signal the presence of TUTase, RNA deaminase or RNA ligase, respectively. This strategy was successfully used in the characterization of the kinetoplastid editosome,[54] but is in practice rather difficult, because the appropriate reaction conditions need to be determined first. For example, RNA ligases require defined substrate ends, in addition to specific pH and ion concentrations. Some ligate only moieties with 3′-hydroxyl and 5′-phosphate termini, while others join 3′-phosphate and 5′-hydroxyl, or even 5′-hydroxyl and 2′,3′-cyclic phosphate termini.[53] MS identification of the RNA ligase class to which candidate joinosome ligases belong will help finding suitable reaction conditions for screening sub-mitochondrial fractions.

Development of transformation *in Diplonema*

If gene manipulation were feasible in *Diplonema*, it would be the approach of choice to dissect molecular machineries. Therefore, we recently started to develop transformation procedures for this organism, initially using protocols applied for *Trypanosoma brucei*, which involve antibiotics as selectable markers. At present, 2 crucial issues are solved: *Diplonema* takes up DNA via electroporation, and several antibiotics can be used for selecting transformants. We succeeded, although with low efficiency, in introducing at various random positions into the nuclear genome a construct containing an antibiotic resistance gene framed by 5′ and 3′ untranslated regions (UTRs) from a *Trypanosoma* gene (Faktorová et al., unpublished).

Work is in progress to optimize constructs for high transformation frequency, homologous integration, and simple detection of introduced genes. The new generation of constructs makes use of long UTRs from highly transcribed *Diplonema* genes, and, in addition to the selective marker, a fluorescent protein gene as a reporter.

Conclusion and outlook

At the time of writing, we are about to embark on the experimental identification of editosome and joinosome components. Once candidate (ribonucleo)protein complexes are isolated and the catalytic activity demonstrated *in vitro*, validation *in vivo* will be required. The latter will be in close reach when genetic engineering methods are fully established in *Diplonema*.

In summary, exploration of this exotic non-model organism was a risky endeavor at first, but has been most lucrative in unearthing more of nature's astounding molecular innovations.

Disclosure of potential conflicts of interest

No potential conflicts of interest were disclosed.

Acknowledgments

We thank M. Aoulad Aissa for excellent technical assistance, B. F. Lang and M. Sarrasin (Université de Montréal) for critically reading the manuscript, and B. Kaur (Institute of Parasitology, České Budějovice) for sharing unpublished data.

Funding

This work was supported by the Canadian Institute of Health Research (CIHR; operating grant MOP-79309 to G.B.), the Gordon and Betty Moore Foundation (grant GBMF4983 to J.L. and G.B.), and the Czech Grant Agency (15-21974S to J.L.).

ORCID

Matus Valach http://orcid.org/0000-0001-8689-0080
Sandrine Moreira http://orcid.org/0000-0002-4580-3566
Drahomíra Faktorová http://orcid.org/0000-0001-9623-2233
Julius Lukeš http://orcid.org/0000-0002-0578-6618
Gertraud Burger http://orcid.org/0000-0002-8679-8812

References

1. Burger G, Moreira S, Valach M. Genes in hiding. Trends Genet 2016; 32:553-65; PMID:27460648; http://dx.doi.org/10.1016/j.tig.2016.06.005
2. Knoop V. When you can't trust the DNA: RNA editing changes transcript sequences. Cell Mol Life Sci 2010; 68:567-86; PMID:20938709; http://dx.doi.org/10.1007/s00018-010-0538-9
3. Maslov DA, Yasuhira S, Simpson L. Phylogenetic affinities of *Diplonema* within the Euglenozoa as inferred from the SSU rRNA gene and partial COI protein sequences. Protist 1999; 150:33-42; PMID:10724517; http://dx.doi.org/10.1016/S1434-4610(99)70007-6
4. Simpson AGB, Roger AJ. Protein phylogenies robustly resolve the deep-level relationships within Euglenozoa. Mol Phylogenet Evol 2004; 30:201-12; PMID:15022770; http://dx.doi.org/10.1016/S1055-7903(03)00177-5
5. Lara E, Moreira D, Vereshchaka A, López-García P. Pan-oceanic distribution of new highly diverse clades of deep-sea diplonemids. Environ Microbiol 2009; 11:47-55; PMID:18803646; http://dx.doi.org/10.1111/j.1462-2920.2008.01737.x
6. de Vargas C, Audic S, Henry N, Decelle J, Mahé F, Logares R, Lara E, Berney C, Le Bescot N, Probert I, et al. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. Science 2015; 348:1261605; PMID:25999516; http://dx.doi.org/10.1126/science.1261605
7. Lukeš J, Flegontova O, Horák A. Diplonemids. Curr Biol 2015; 25: R702-4; http://dx.doi.org/10.1016/j.cub.2015.04.052
8. Flegontova O, Flegontov P, Malviya S, Audic S, Wincker P, de Vargas C, Bowler C, Lukeš J, Horák A. (2016) Unexpected diversity and abundance of planktonic diplonemids in the world ocean. Curr. Biol. (in press).
9. Gawryluk RMR, del Campo J, Okamoto N, Strassert JFH, Lukeš J, Richards TA, Worden AZ, Santoro AE, Keeling PJ. (2016) Morphological identification and single-cell genomics of marine diplonemids. Curr. Biol. (in press).
10. Vlcek C, Marande W, Teijeiro S, Lukeš J, Burger G. Systematically fragmented genes in a multipartite mitochondrial genome. Nucleic Acids Res 2011; 39:979-88; PMID:20935050; http://dx.doi.org/10.1093/nar/gkq883
11. Moreira S, Valach M, Aoulad-Aissa M, Otto C, Burger G. Novel modes of RNA editing in mitochondria. Nucleic Acids Res 2016; 44:4907-19; PMID:27001515; http://dx.doi.org/10.1093/nar/gkw188
12. Marande W, Lukeš J, Burger G. Unique mitochondrial genome structure in diplonemids, the sister group of kinetoplastids. Eukaryot Cell 2005; 4:1137-46; PMID:15947205; http://dx.doi.org/10.1128/EC.4.6.1137-1146.2005
13. Marande W, Burger G. Mitochondrial DNA as a genomic jigsaw puzzle. Science 2007; 318:415; PMID:17947575; http://dx.doi.org/10.1126/science.1148033
14. Kiethega GN, Yan Y, Turcotte M, Burger G. RNA-level unscrambling of fragmented genes in *Diplonema* mitochondria. RNA Biol 2013; 10:301-13; PMID:23324603; http://dx.doi.org/10.4161/rna.23340
15. Kiethega GN, Turcotte M, Burger G. Evolutionarily conserved *cox1* *trans*-splicing without *cis*-motifs. Mol Biol Evol 2011; 28:2425-8; PMID:21436119; http://dx.doi.org/10.1093/molbev/msr075
16. Moreira S, Breton S, Burger G. Unscrambling genetic information at the RNA level. Wiley Interdiscip Rev RNA 2012; 3:213-28; PMID:22275292; http://dx.doi.org/10.1002/wrna.1106
17. Valach M, Moreira S, Kiethega GN, Burger G. Trans-splicing and RNA editing of LSU rRNA in *Diplonema* mitochondria. Nucleic Acids Res 2014; 42:2660-72; PMID:24259427; http://dx.doi.org/10.1093/nar/gkt1152
18. Feagin JE, Abraham JM, Stuart K. Extensive editing of the cytochrome c oxidase III transcript in *Trypanosoma brucei*. Cell 1988; 53:413-22; PMID:2452697; http://dx.doi.org/10.1016/0092-8674(88)90161-4
19. Read LK, Lukeš J, Hashimi H. Trypanosome RNA editing: the complexity of getting U in and taking U out. Wiley Interdiscip Rev RNA 2016; 7:33-51; PMID:26522170; http://dx.doi.org/10.1002/wrna.1313
20. Roy J, Faktorová D, Benada O, Lukeš J, Burger G. Description of *Rhynchopus euleeides* n. sp. (Diplonemea), a free-living marine euglenozoan. J Eukaryot Microbiol 2007; 54:137-45; PMID:17403154; http://dx.doi.org/10.1111/j.1550-7408.2007.00244.x
21. Roy J, Faktorová D, Lukeš J, Burger G. Unusual mitochondrial genome structures throughout the Euglenozoa. Protist 2007; 158:385-96; PMID:17499547; http://dx.doi.org/10.1016/j.protis.2007.03.002
22. Flegontov P, Gray MW, Burger G, Lukeš J. Gene fragmentation: a key to mitochondrial genome evolution in Euglenozoa? Curr Genet 2011; 57:225-32; PMID:21544620; http://dx.doi.org/10.1007/s00294-011-0340-8
23. Yabuki A, Tanifuji G, Kusaka C, Takishita K, Fujikura K. Hyper-eccentric structural genes in the mitochondrial genome of the algal parasite *Hemistasia phaeocysticola*. Genome Biol Evol 2016; evw207; http://dx.doi.org/10.1093/gbe/evw207
24. Yabuki A, Tame A. Phylogeny and reclassification of *Hemistasia phaeocysticola* (Scherffel) Elbrächter & Schnepf, 1996. J Eukaryot Microbiol 2015; 62:426-9; PMID:25377132; http://dx.doi.org/10.1111/jeu.12191
25. Wallach DFH, Kamat VB. Plasma and cytoplasmic membrane fragments from Ehrlich ascites carcinoma. Proc Natl Acad Sci USA 1964; 52:721-8; PMID:14212548; http://dx.doi.org/10.1073/pnas.52.3.721
26. Hauser R, Pypaert M, Häusler T, Horn EK, Schneider A. In vitro import of proteins into mitochondria of *Trypanosoma brucei* and *Leishmania tarentolae*. J Cell Sci 1996; 109:517-23; PMID:8838675
27. Schneider A, Charrière F, Pusnik M, Horn EK. Isolation of mitochondria from procyclic *Trypanosoma brucei*. Methods Mol Biol 2007; 372:67-80; PMID:18314718; http://dx.doi.org/10.1007/978-1-59745-365-3_5
28. Burger G, Lavrov DV, Forget L, Lang BF. Sequencing complete mitochondrial and plastid genomes. Nat Protoc 2007; 2:603-14; PMID:17406621; http://dx.doi.org/10.1038/nprot.2007.59
29. Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S, et al. SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. Bioinformatics 2014; 30:1660-6; PMID:24532719; http://dx.doi.org/10.1093/bioinformatics/btu077
30. Schulz MH, Zerbino DR, Vingron M, Birney E. *Oases*: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 2012; 28:1086-92; PMID:22368243; http://dx.doi.org/10.1093/bioinformatics/bts094
31. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol 2011; 7:e1002195; PMID:22039361; http://dx.doi.org/10.1371/journal.pcbi.1002195
32. Gautheret D, Lambert A. Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles. J Mol Biol 2001; 313:1003-11; PMID:11700055; http://dx.doi.org/10.1006/jmbi.2001.5102
33. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 2013; 29:2933-5; PMID:24008419; http://dx.doi.org/10.1093/bioinformatics/btt509
34. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012; 9:357-9; PMID:22388286; http://dx.doi.org/10.1038/nmeth.1923
35. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013; 29:15-21; PMID:23104886; http://dx.doi.org/10.1093/bioinformatics/bts635
36. Otto C, Stadler PF, Hoffmann S. Fast and sensitive mapping of bisulfite-treated sequencing data. Bioinformatics 2012; 28:1698-704; PMID:22581174; http://dx.doi.org/10.1093/bioinformatics/bts254

37. Neupert W, Herrmann JM. Translocation of proteins into mitochondria. Annu Rev Biochem 2007; 76:723-49; PMID:17263664; http://dx.doi.org/10.1146/annurev.biochem.76.052705.163409
38. Gerber AP, Keller W. RNA editing by base deamination: more enzymes, more targets, new mysteries. Trends Biochem Sci 2001; 26:376-84; PMID:11406411; http://dx.doi.org/10.1016/S0968-0004(01)01827-8
39. Kwak JE, Wickens M. A family of poly(U) polymerases. RNA 2007; 13:860-7; PMID:17449726; http://dx.doi.org/10.1261/rna.514007
40. Tanaka N, Meineke B, Shuman S. RtcB, a novel RNA ligase, can catalyze tRNA splicing and *HAC1* mRNA splicing *in vivo*. J Biol Chem 2011; 286:30253-7; PMID:21757685; http://dx.doi.org/10.1074/jbc.C111.274597
41. Sun T, Bentolila S, Hanson MR. The unexpected diversity of plant organelle RNA editosomes. Trends Plant Sci 2016; S1360-1385; http://dx.doi.org/10.1016/j.tplants.2016.07.005
42. Moreira S, Noutahi E, Lamoureux G, Burger G. Three-dimensional structure model and predicted ATP interaction rewiring of a deviant RNA ligase 2. BMC Struct Biol 2015; 15:20
43. Wittig I, Braun HP, Schägger H. Blue native PAGE. Nat Protoc 2006; 1:418-28; PMID:17406264; http://dx.doi.org/10.1038/nprot.2006.62
44. Petrov A, Wu T, Puglisi EV, Puglisi JD. RNA purification by preparative polyacrylamide gel electrophoresis. Meth Enzymol 2013; 530:315-30; PMID:24034329; http://dx.doi.org/10.1016/B978-0-12-420037-1.00017-8
45. Greber BJ, Ban N. Structure and function of the mitochondrial ribosome. Annu Rev Biochem 2016; 85:103-32; PMID:27023846; http://dx.doi.org/10.1146/annurev-biochem-060815-014343
46. Göringer HU. "Gestalt," composition and function of the *Trypanosoma brucei* editosome. Annu Rev Microbiol 2012; 66:65-82; PMID:22994488; http://dx.doi.org/10.1146/annurev-micro-092611-150150
47. Papasaikas P, Valcárcel J. The spliceosome: the ultimate RNA chaperone and sculptor. Trends Biochem Sci 2016; 41:33-45; PMID:26682498; http://dx.doi.org/10.1016/j.tibs.2015.11.003
48. Engreitz JM, Sirokman K, McDonel P, Shishkin AA, Surka C, Russell P, Grossman SR, Chow AY, Guttman M, Lander ES. RNA-RNA interactions enable specific targeting of noncoding RNAs to nascent pre-mRNAs and chromatin sites. Cell 2014; 159:188-99; PMID:25259926; http://dx.doi.org/10.1016/j.cell.2014.08.018
49. McHugh CA, Chen CK, Chow A, Surka CF, Tran C, McDonel P, Pandya-Jones A, Blanco M, Burghard C, Moradian A, et al. The *Xist* lncRNA interacts directly with SHARP to silence transcription through HDAC3. Nature 2015; 521:232-6; PMID:25915022; http://dx.doi.org/10.1038/nature14443
50. Chu C, Zhang QC, da Rocha ST, Flynn RA, Bharadwaj M, Calabrese JM, Magnuson T, Heard E, Chang HY. Systematic discovery of Xist RNA binding proteins. Cell 2015; 161:404-16; PMID:25843628; http://dx.doi.org/10.1016/j.cell.2015.03.025
51. Ho CK, Wang LK, Lima CD, Shuman S. Structure and mechanism of RNA ligase. Structure 2004; 12:327-39; PMID:14962393; http://dx.doi.org/10.1016/j.str.2004.01.011
52. Tanaka N, Chakravarty AK, Maughan B, Shuman S. Novel mechanism of RNA repair by RtcB via sequential 2′,3′-cyclic phosphodiesterase and 3′-phosphate/5′-hydroxyl ligation reactions. J Biol Chem 2011; 286:43134-43; PMID:22045815; http://dx.doi.org/10.1074/jbc.M111.302133
53. Popow J, Schleiffer A, Martinez J. Diversity and roles of (t)RNA ligases. Cell Mol Life Sci 2012; 69:2657-70; PMID:22426497; http://dx.doi.org/10.1007/s00018-012-0944-2
54. Aphasizheva I, Aphasizhev R. U-insertion/deletion mRNA-editing holoenzyme: definition in sight. Trends Parasitol 2016; 32:144-56; PMID:26572691; http://dx.doi.org/10.1016/j.pt.2015.10.004