University of South Bohemia

Faculty of Science

Master thesis

Vojtěch David **2015**

University of South Bohemia

Faculty of Science

High-throughput analysis of uridine insertion and deletion RNA editing in *Perkinsela*

Master thesis

Bc. Vojtěch David

Supervisor: Prof. RNDr. Julius Lukeš, Csc.

Supervisor specialist: MSc. Pavel Flegontov, PhD.

České Budějovice 2015

David, V., 2015: High-throughput analysis of uridine insertion and deletion RNA editing in *Perkinsela*. Master thesis, in English. – 39p., Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic.

Annotation:
This thesis is a follow-up of my Bachelor thesis about the mitochondrial genome of kinetoplastid protist *Perkinsela sp*. This work introduces a novel approach in high-throughput analysis method of uridine insertion and deletion RNA editing, describes its background and proposes its further development. Its effect on the interpretation of U-indel editing, both in *Perkinsela* and in general, is demonstrated via attached manuscript which also introduces other biologically relevant aspects of Perkinsela mitochondrion.

I hereby declare that this bachelor thesis is entirely the result of my own work except where otherwise indicated. I have only used the resources given in the list of references.

I hereby declare that, in accordance with Article 47b of Act No. 111/1998 in the valid wording, I agree with the publication of my master thesis, in full form to be kept in the Faculty of Science archive, in electronic form in publicly accessible part of the STAG database operated by the University of South Bohemia in České Budějovice accessible through its web pages. Further, I agree to the electronic publication of the comments of my supervisor and thesis opponents and the record of the proceedings and results of the thesis defence in accordance with aforementioned Act No. 111/1998. I also agree to the comparison of the text of my thesis with the Theses.cz thesis database operated by the National Registry of University Theses and a plagerism detection system.

Date .....

Signature .....

Manuscript authorship declaration

I hereby declare, that I have dominantly contributed to the attached manuscript in all aspects except for: a) microscopy, b) cell culture, c) majority of sequencing libraries preparations d) northern blotting e) Bowtie2 source code modification and d) T-aligner source code. The novel methodology has been co-developed by me, Pavel Flegontov and Evgeny Gerasimov, who has mainly contributed by programming and early software testing. Remaining bioinformatic analyses and data visualization had been carried out almost completely by me under the inspiring supervision of Pavel Flegontov. The text of the manuscript is a product of collaborative effort of all the authors, depending on their contribution and area of specialization.

Table of contents

1.1 Introduction..... 1
2.1 Guide to high-throughput analysis of transcriptomic sequences with extensive uridine insertion and deletion RNA editing..... 2
2.2 Input data..... 3
2.3 Seeding..... 4
2.4 Alignment..... 5
2.5 Main product reconstruction and alternative editing..... 6
2.6 Discussion..... 8
2.7 Literature.....11
3.1 Biological significance of uridine insertion and deletion RNA editing analysis.....12
3.2 Gene loss and error-prone RNA editing in the mitochondrion of Perkinsela, an endosymbiotic kinetoplastid (manuscript beginning).....13
3.3 Introduction.....14
3.4 Materials and methods.....16
3.5 Results and discussion.....22
3.6 References.....35
4.1 Suplementary data.....40

Introduction

The first case of RNA editing has been described as early as in the eighth decade of the last century. In a kinetoplastid protozoan, *Trypanosoma brucei*, insertion of uridines not encoded in their mitochondrial genome, has been published (Benne et al., 1986). Ever since, several other types of other RNA editing have been described, across all domains of life (Maas, 2012).

After the initial discovery of *cox2* (subunit 2 of cytochrome oxidase, COII) editing, RNA editing was further described in other mitochondrial genes of *T.brucei* (Feagin et al., 1987; Shaw et al., 1988) and closely related kinetoplastids (Lukeš et al., 1994); Blom et al., 1998). In summary, kinetoplastid uridine insertion/deletion (U-indel) RNA editing affects transcript domains or whole transcripts, with up to half of nucleotides in the final transcript being created by RNA editing (Koslowsky et al., 1990). In transcripts not edited over their entire length, 3' region is typically edited. That is true for trypanosomatids, the most extensively studied kinetoplastid clade. In contrast, some early branching kinetoplastids have a pattern of two edited domains at both ends of the transcript (Lukeš et al., 1994 ; Blom et al., 1998; David, 2013).

The amazing discovery of RNA editing has been possible because of emerging techniques of nucleic acid sequencing. Since 1984, sequencing technologies have undergone rapid evolution, outperforming the older strategies more than ten hundred times (Kircher and Kelso, 2010). Increasing amount of transcriptomic data currently available for indel-edited transcripts has turned out to be difficult to analyze with common bioinformatic tools, developed mostly for traditional model species (David, 2013; Koslowsky et al., 2014; Ochsenreiter and Hajduk, 2006).

Today, the mechanism of U-indel editing is arguably well defined (Fig. 1). It starts with transcription of the pre-edited mRNA and short guide RNAs (gRNAs) (Koslowsky et al., 2014). Editing of the mRNA proceeds from 3' to 5' (Halbig et al., 2004, attached manuscript) of the edited domain using information encoded within short guiding domains of gRNAs. Uridines are inserted into or excised from mRNA until the edited region pairs perfectly with the gRNA, with G to U pairing allowed (Aphasizhev and Aphasizheva, 2014). This is performed by a large protein complex called editosome (Worthey, 2003). The next gRNA acts further upstream, thus forming a cascade of editing (Aphasizhev and Aphasizheva,

2014). Yet, several aspects of U-indel RNA editing remain poorly understood (Ochsenreiter et al., 2008; Aphasizheva et al., 2011; Ridlon et al., 2013).

My bachelor thesis project on the mitochondrial genome and RNA editing of *Perkinsela*, which is a unique case of intracellular endosymbiotic kinetoplastid living within a parasitic amoebozoan *Paramoeba* (see the attached manuscript) have turned out to be the first investigation of U-indel edited mitochondrial transcripts based on transcriptome sequencing. For that purpose, we have written a novel software called T-aligner, and modified the existing Bowtie2 read mapper. This work should serve as a guide explaining our general methodology for analysis of U-indel editing with second-generation sequencing.

Part 1:

Guide to high-throughput analysis of transcriptomic sequences with extensive uridine insertion and deletion RNA editing

On the importance of optimal analysis of RNA editing

The importance of proper data analysis and optimization can be illustrated by my Bachelor thesis (David, 2013), in which two abundant extensively edited transcripts were misidentified as a divergent mitochondrial rRNA, a case not unprecedented in kinetoplastids and their relatives (Sharma et al., 2009; Valach et al., 2014). The fully edited sequence, easily identifiable as *cox2*, was not reconstructed at that time due to technical problems with mapping large sets of extensively edited reads. In this part of my thesis, I am showing how various parameters affect outcomes of RNA editing analysis, introducing a novel software developed for that purpose, and speculate on a concept for even more sophisticated and less labor-intensive analysis of U-indel RNA editing.

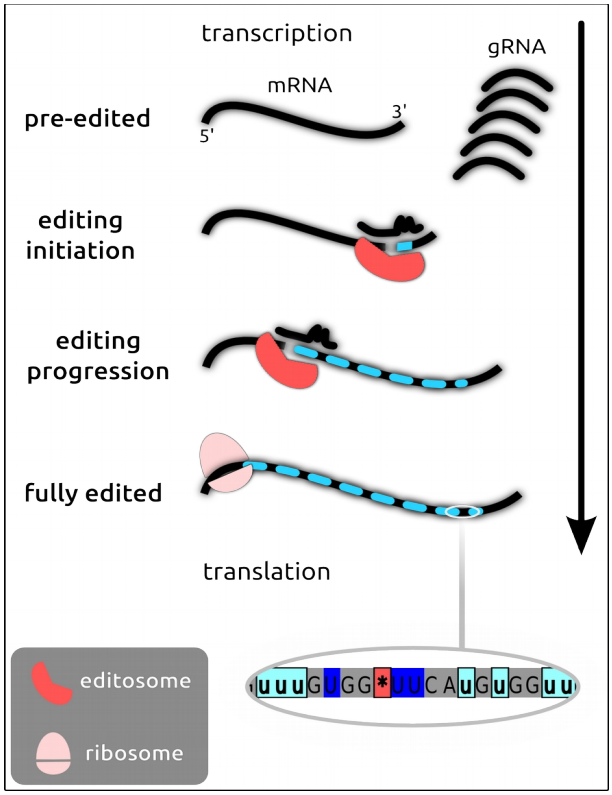


Figure 1:Schematic representation of RNA editing in kinetoplastids. Within the edited sequence sample, the inserted uridines are shown in small lowercase, deleted uridine is represented by an asterisk.

Input data

Currently the mainstream second-generation sequencing platform, Illumina, provides sequencing reads up to 300 nucleotides (nt) long. Usually, the first step of trancriptomic analysis is read mapping, a high-throughput alignment process which works in a BLAST-like manner (David, 2013). Unfortunately, other contemporary platforms (454, Ion Torrent) tend to introduce indel errors in homopolymer tracts (Kircher and Kelso, 2010), which makes them not suitable for U-indel editing analysis. Although Illumina reads contain a relatively small amount of single nucleotide mismatch errors (Nakamura et al., 2011), this platform remains the best option for investigation of U-indel RNA editing.

As the first model case, paired 250nt strand-specific RNA-seq reads from the *Paramoeba-Perkinsela* symbiotic system (see attached manuscript for details) were merged into pseudo-reads with a minimum overlap of 10 nt (Fig. 2a). The second model case is illustrated by much shorter RNA-seq reads (Fig. 2b), obtained via the iCLIP procedure schematically

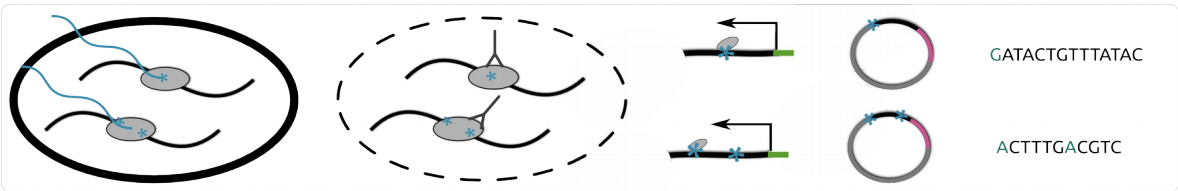


Figure 2: Schematic representation of the iCLIP protocol. First, RNA is UV-crosslinked to the protein (protein-RNA interaction is represented by asterisk). Second, cell is lysed and the protein of interest is purified with imunoprecipitation. Third, excessive RNA is removed by SDS-PAGE, the protein is digested by protease K, and cDNA is prepared. Fourth, size selected cDNA is ligated with sequencing adapters, circularized, amplified wit PCR and sequenced.

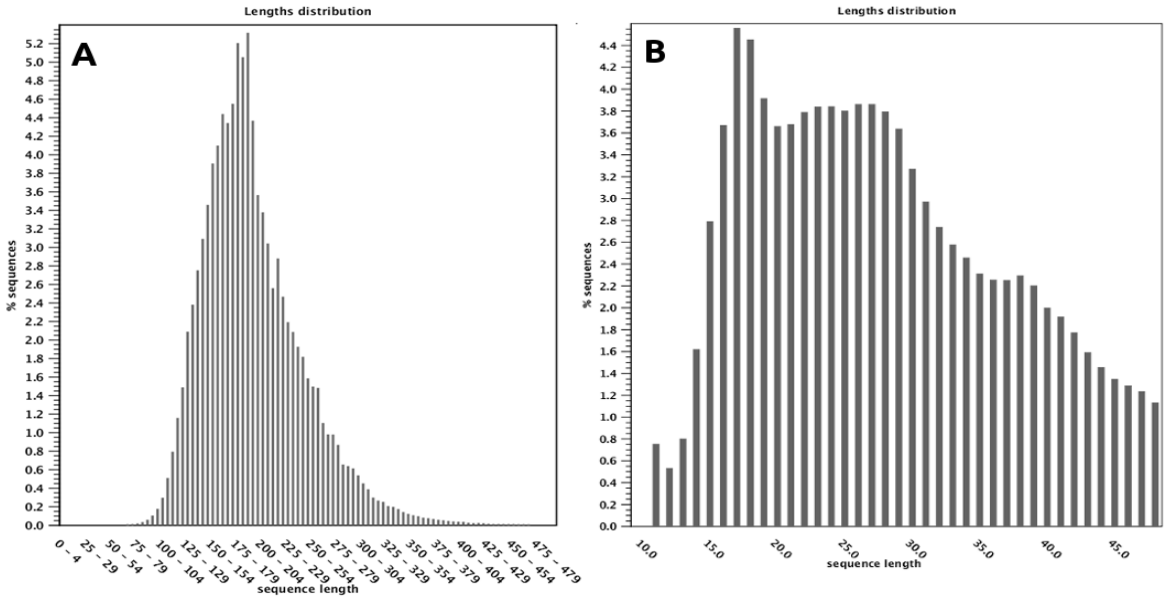


Figure 3: Length distribution for RNA-seq reads of Paramoeba pemaquidensis *(merged pseudo-reads obtained from trimmed and filtered paired reads).* **B.** *Length distribution for* Trypanosoma brucei *iCLIP reads (adapters and barcodes removed)*

shown in Fig. 3 (see also König et al., 2010). The latter set of reads represents mRNA isolated from a UV-crosslinked RNA-binding protein involved in U-indel RNA editing (MRP1 protein of *T.brucei*) and, unlike the first case, informative sequencing errors, which are dependent on protein RNA interactions during UV irradiation (Pandit et al., 2013).

Seeding

The first part of read mapping process starts with so-called seeding, an heuristic step which makes mapping much faster. During seeding, short sub-strings (seeds) of the read and its complement sequence are taken and usually searched for a perfect match within the reference sequence (Hoffmann et al., 2009). Only the reads containing a seed are allowed to proceed to the regular alignment, greatly reducing computation time at the cost of some false negative results (Langmead and Salzberg, 2012).

Read loss caused by improper seeding counts twice for heavily edited U-indel data. Due to closely spaced indels, a fraction of reads will contain no seed perfectly matching the reference. This problem becomes especially severe if an average read is much shorter than the edited domain. This problem further escalates in editing domains more than two times longer than the average read length, where no edited read can be mapped regardless the coverage. The solution for the seeding problem is to use a few partially edited references for mapping, or to remove Us completely from both the reads and the reference at the cost of information loss and increase in false positive results discussed below.

L	Δt (min)	N (reads)
21	0	1220
19	0	1235
16	6	1475
13	16	1760
10	928	3301

Table 1:Number of reads mapped with the modified Bowtie2 using U-indel optimized settings depends on seed length: cob *transcript,* Perkinsela *strain GillNOR1/I .*

In the case of seeding, Bowtie2 works in a traditional way and respects read length during the seed preparation step. The effect of seed length on read coverage in the modified Bowtie2 is shown in Table 1. On the other hand, T-aligner uses a short fixed seeding region on the reference. If a seeding region is carefully chosen in a never-edited region, such strategy can significantly alleviate the seeding problem. To avoid incorrect seeding, while keeping high sensitivity, the smallest seed length ever used in our work was 10 nt. and the longest seed was 20 nt. The optimal seed length used for the final version of read mapping has been 14 nt in case of *Perkinsela* mitochondrial reads and 10nt for iCLIP data. These values reflect a tradeoff between amount of data being mapped, reference length, read length expected sequencing errors and acceptable computation time.

Alignment

Reads with a positive seed match are subjected to a full alignment. The alignment is achieved through Smith-Waterman-like algorithms which starts match-making between the read and reference from the aligned seed and move towards both ends of the read (Smith, 1981). Each misaligned nucleotide is penalized, and the highest scoring alignment is then accepted as long as the best score does not fall below a rejection threshold. Mismatches, gap openings and gap extensions are penalized separately.

The penalty ratios and the rejection threshold require special attention in U-indel edited data as default settings will only map a small fraction of relevant reads. Uridine-specific gap opening and extension penalties cannot be set in contemporary read mappers, therefore this small addition has been introduced into the Bowtie2 mapper (attached manuscript). Generally, penalty values should follow this scheme: gaps containing U only << mismatches < gaps containing C/A/G. Otherwise, due to densely spaced inserts, reads will be misaligned with long heteropolymer insertions .

Optimization of the modified Bowtie2 mapper for correct alignment of *Perkinsela* indel-edited reads has resulted in the following settings: 1) mismatch penalty of 18, 2) gap opening and extension penalties of 10 for C/A/G, 3) gap opening and extension penalties of 1 for U, 4) original gap policy was changed to allow indels at the very end of the read. Sensitivity for mapping millions of reads on the mitochondrial genome has been enhanced using seed length of 14. A number of attempts to align each read when looking for the best alignment has been increased to 20. A number of seeding attempts performed in case a duplicate seed is found has been increased to 3. The rejection threshold in Bowtie2 can be dependent on a function read length, which is especially helpful when merged pseudo-reads with a broad length distribution are used. In order to align heavily edited reads, the sum of penalty values each read can have has been risen to 0.5x read length.

An alternative alignment solution has been introduced in the form of a novel software, T-aligner (attached manuscript). T-aligner exploits the biological nature of RNA editing by aligning reads only in 3' to 5' direction from a pre-defined seeding region and allowing only U-indels and a small number of mismatches. The latest version of T-aligner also counts and categorizes reads into main and alternative editing pathways, if a fully edited sequence is provided (see below).

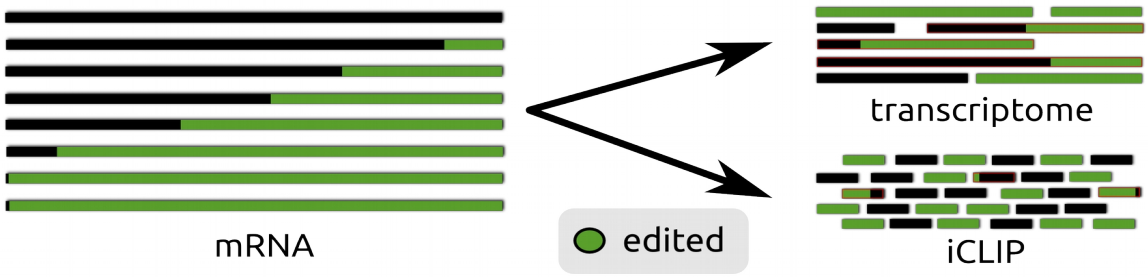


Figure 4: Fraction of reads containing partially edited sequence increases with read length. The bars represent individual sequences, green and black color stands for edited and pre-edited sequence, respectively. The fraction of partially edited reads (edited and pre-edited sequence 'hybrid') within transcriptomic libraries consisting of long reads is significantly large, whereas iCLIP libraries contain only a tiny fraction of 'hybrids', unless an investigated protein specifically binds partially edited region.

Since iCLIP analysis uses small reads (Fig. 2b), and the fraction of partially edited reads is usually low (Fig. 4), a pre-set 'very sensitive' program of Bowtie2 is sufficient for mapping iCLIP reads either on pre-edited or fully edited references. For a minor fraction of reads, being 'hybrids' of edited and non-edited sequences, the U-indel-optimized program derived from the *Perkinsela* project works well. Due to small read sets generated by iCLIP experiments, seed length of 10 can be routinely used without huge computation costs. The iCLIP read mappings (Fig. 5) require further downstream processing and statistical analysis (König et al., 2010).

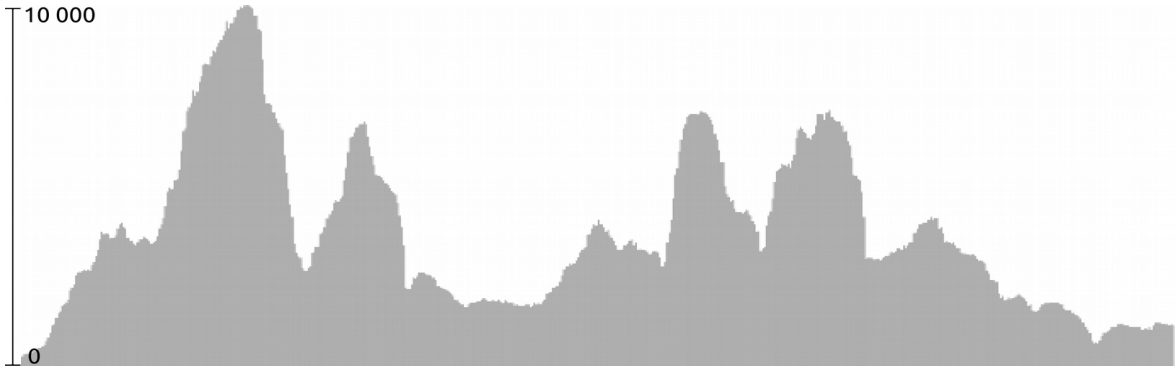


Figure 5:An example of iCLIP reads mapped on fully edited mRNA of cox3 in T.brucei.

Main product reconstruction and alternative editing

In order to reconstruct fully edited sequences *de novo* with transcriptomic data and a genome at hand, it is important to run Bowtie 2 mappings with 10-14 nt seeds, low U-indel or A-indel (for antisense transcripts) penalties and loose rejection thresholds. A coverage profile of strand-specific reads, can guide preliminary annotation, and edited transcript domains can be identified as indel-rich regions of the multiple read alignment. With the knowledge of transcript loci and editing domain borders, seeding regions for T-aligner can be set at the 3' end of each edited domain. The fully edited sequence is then reconstructed by recursively adjusting the reference sequence based on the most supported sequence inferred

from the T-aligner output and shifting a seeding region towards the 5' of the domain. Finally, a predicted fully edited sequence can be verified by conceptual translation and building a protein phylogenetic tree.

However, a few cases requiring special attention can occur and have to be solved individually at the moment. First, the edited region can be so close to the 3' of the transcript that almost no edited reads can be mapped with Bowtie2 due to seeding problems, and hence a seeding region for a subsequent run of T-aligner cannot be inferred. Second, reads at the very 3' usually contain U- and AU-rich tails which can bias read mapping and aggravate the effect described before. These issues can usually be solved by manual analysis of Bowtie 2 alignment, from which first edited sites can be derived and included into the seeding region for T-aligner.

Once a fully edited sequence is reconstructed and confirmed, it can be provided to T-aligner for alternative editing analysis. With the fully edited sequence loaded (referred to as the main editing pathway), T-aligner detects reads matching the genomic sequence and the main pathway, as well as partially edited reads, which are not in disagreement with the main pathway. The latter group of reads is defined in the following way: a) no additional edited sites are present compared to the main pathway, b) indels are shorter or equal in size to those in the main pathway, c) all sites are edited in the same direction as in the main pathway (insertion vs insertion, deletion vs deletion). Remaining sequences are marked as alternatively edited reads and this group is further narrowed down by merging shorter reads with exactly matching longer ones. Support values (average read counts) are assigned to each non-redundant editing intermediate obtained this way. The whole process of RNAediting analysis is demonstrated in Fig. 6.

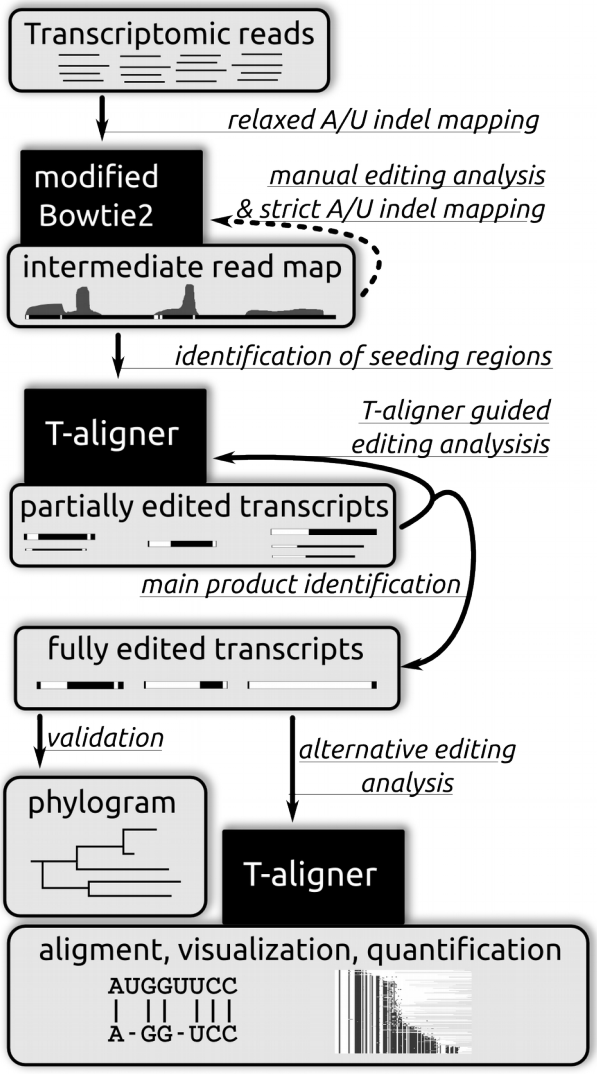


Figure 6:RNA editng analysis flowchart

Discussion

Despite the advantages of the U-indel editing analysis method described here, there is always room for improvements and tinkering. Here I propose a concept for the second-generation of 'U-indel editing solver', which could be easily used to analyze U-indel editing across kinetoplastid diversity and under various experimental conditions in model trypanosomatids. The hallmarks of this approach are: requirement for a single reference sequence for the whole analysis, usage of T-less reads for seeding, and an algorithm for reconstruction of editing variants somewhat similar to recent transcriptomic assembly approaches. The proposed software would be composed of two independent modules, compatible with other tools through commonly used formats (Fig 7).

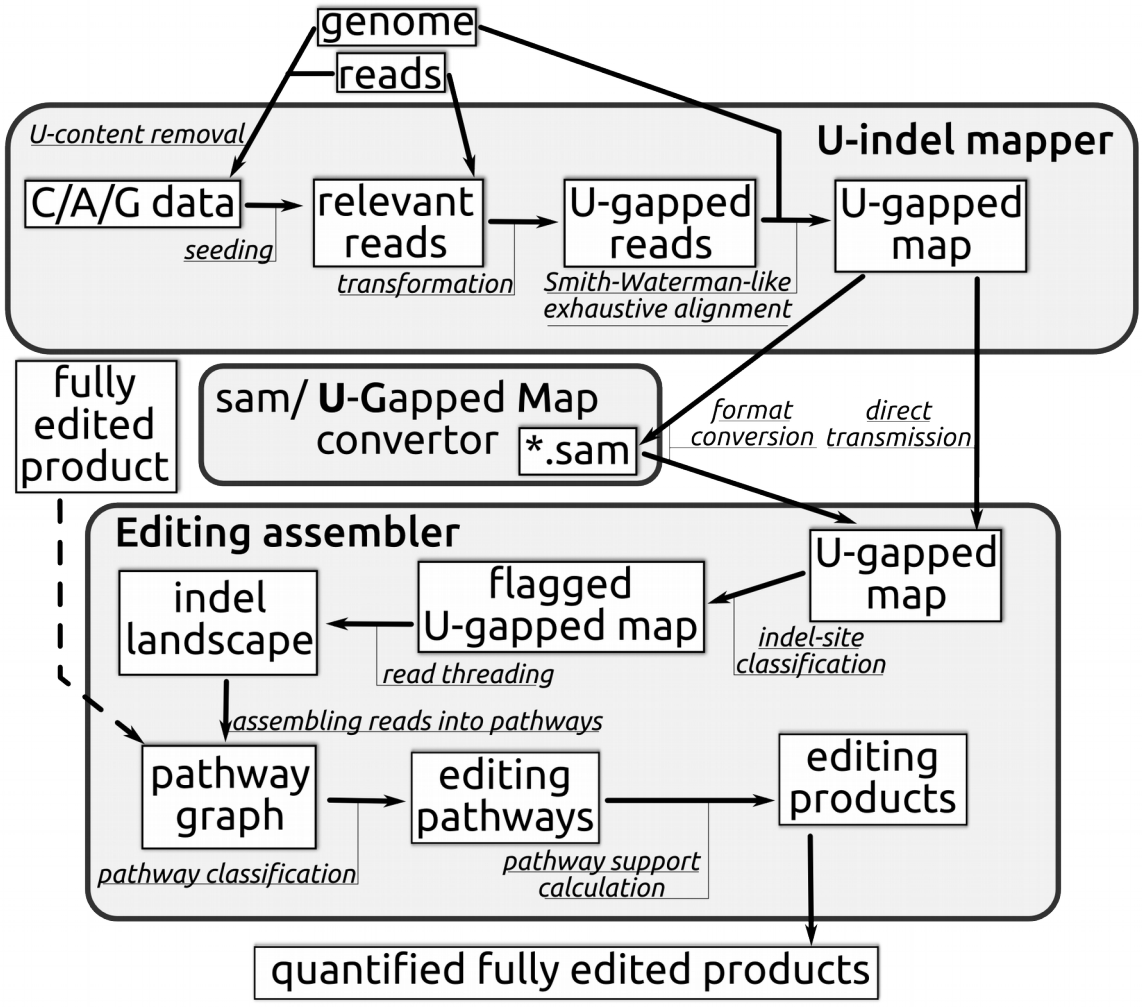


Figure 7: Flowchart of proposed software.

The first module, U-indel mapper, would take a genomic template without U nucleotides. By removing U content, this reference now resembles all possible versions and editing intermediates including the fully edited product. On this reference, U-less seeding with a

long seed would quickly scan through reads saving only names of any positive matches. This will reduce read count to several thousands per gene (assuming standard library size), which will undergo transformation into a 'U-gapped format' composed of non-T(U) nucleotides interspersed with numbers of Us between them, for example 0G1A0G2 instead of GTAGTT. These transformed reads will be than exhaustively aligned to the likewise transformed reference in a Smith-Waterman manner with high indel penalties, assuming indels to be mis-sequenced uridines. At this step, an alignment in this novel format could be either transformed to the standard SAM output format, or passed to the second module.

The workflow of the second module, Editing assembler, can be divided into 3 steps. A reference genomic sequence and mapped reads are required as well as a fully edited sequence. In case the final product remains unknown, the software will assume the most abundant translatable sequence assembled to be a main product, and it will be marked in the output.

At the first step, aligned reads in the 'U-gapped format' will be compared with reference and U-indel sites will be flagged in each read based on the number of uridines as ‘edited site’, ‘terminally edited site’ or ‘pre-edited site’. An 'indel landscape' of all possible U-indels will be also created. The 'indel landscape' is basically a two-dimensional graph of edited positions on the x-axis, and number of inserted/deleted uridines, ΔU on the y-axis. Based on the 'indel landscape', an assembly-like graph (Fig. 8) will be created, representing consecutive editing steps supported by individual reads threaded through them.

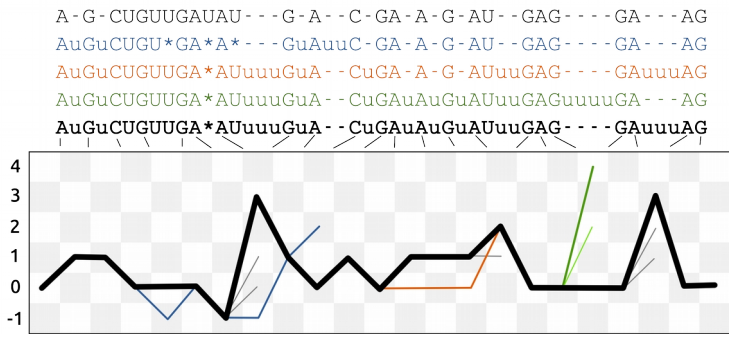


Figure 8: Graphical representation of an ‘indel landscape’ and multiple sequence alignment of respective transcripts. The alignment shows genomic reference (on top), 3 alternative color-coded pathways and the main pathway in bold (at the bottom). A lowercase 'u' stands for an inserted nucleotide, an asterisk for post-transcriptional deletion. In the graph, each column represents an editing site and each line represents a change in uridine content. Relative support is represented by pathway thickness. Thin lines are examples of editing intermediates

At the second step, paths in the graph composed of overlapping reads will be identified, in an assembly-like manner. Paths will be then classified into ‘main’ and ‘alternative’ using the knowledge of ‘terminal editing states’ at each site, simply speaking, the number of Us at each site in the final edited product (introducing these into the initial graph would greatly

complicate its structure). By comparing the number of reads supporting each pathway, the main pathway would be identified, in case it was not provided prior to the analysis.

At the third step, each pathway is assigned a relative abundance estimate. Initially, each alternative pathway has a score representing a number of mapped reads (1 read equals support value of 1) which do not belong to any other pathway. A read supporting several alternative pathways at once can occur, and such a shared read will contribute to the final support score only a fraction of its value, proportional to the number of unique read per pathway. For example, a read shared among two pathways, one supported by a single unique read and the second by 3, will add 0.25 to the first pathway and 0.75 to the second. Pathways departing from the main one, will be further given a share of support value of the main pathway reflecting the length of shared sequence. Finally, all sequences will be reported along with their support values.

In summary, this module should produce a list of edited transcripts incompatible with the main editing product, but without their own partially edited versions. Moreover, a much more realistic estimate of relative proportions of such products compared to the fully edited transcript and its precursors (i.e. the main pathway) will be made.

Within this part and the attached manuscript, I have showed that the software solutions for RNA editing analysis developed during my master studies are capable of handling large second generation sequencing data-sets. Above-mentioned software and settings produce high-quality read mapping and their usage highlights the complexity of editing errors and alternatively edited variants in a novel way. In addition, I have proposed here an even more exhaustive solution for mapping U-indel-edited reads, which could overcome a few limitations of the current approach and operate in a more straightforward and time-efficient manner.

Literature

Aphasizhev, R., and Aphasizheva, I. (2014). Mitochondrial RNA editing in trypanosomes: Small RNAs in control. Biochimie *100*, 125–131.

Aphasizheva, I., Maslov, D., Wang, X., Huang, L., and Aphasizhev, R. (2011). Pentatricopeptide Repeat Proteins Stimulate mRNA Adenylation/Uridylation to Activate Mitochondrial Translation in Trypanosomes. Mol. Cell *42*, 106–117.

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

Blom, D., De Haan, A., Van Den Berg, M., Sloof, P., Benne, R., Jirku, M., and Lukeš, J. (1998). RNA editing in the free-living bodonid Bodo saltans. Nucleic Acids Res. *26*, 1205–1213.

David, V. (2013). Assembly and annotation of a mitochondrial genome of kinetoplastid protist Perkinsela. Thesis in English,Faculty of Science, University of South Bohemia., 33p.

Feagin, J.E., Douglas, P.J., and Stuart, K. (1987). Developmentally regulated addition of nucleotides within apocytochrome b transcripts in Trypanosoma brucei. Cell *49*, 337–345.

Halbig, K., de Nova-Ocampo, M., and Cruz-Reyes, J. (2004). Complete cycles of bloodstream trypanosome RNA editing in vitro. RNA *10*, 914–920.

Hoffmann, S., Otto, C., Kurtz, S., Sharma, C.M., Khaitovich, P., Vogel, J., Stadler, P.F., and Hackermüller, J. (2009). Fast Mapping of Short Sequences with Mismatches, Insertions and Deletions Using Index Structures. PLoS Comput. Biol. *5*, e1000502.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D.J., Luscombe, N.M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. Nat. Struct. Mol. Biol. *17*, 909–915.

Kircher, M., and Kelso, J. (2010). High-throughput DNA sequencing - concepts and limitations. BioEssays *32*, 524–536.

Koslowsky, D., Bhat, G., Perrolaz, A., Feagin, J.E., and Kenneth, S. (1990). The MURF3 gene of T. brucei contains multiple domains of extensive editing and is homologous to a subunit of NADH dehydrogenase. Cell *62*, 901–911.

Koslowsky, D., Sun, Y., Hindenach, J., Theisen, T., and Lucas, J. (2014). The insect-phase gRNA transcriptome in Trypanosoma brucei. Nucleic Acids Res. *42*, 1873–1886.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lukeš, J., Arts, G., van den Burg, J., De Haan, A., Opperdoes, F., Sloof, P., and Benne, R. (1994). boreli.pdf. Nov. Pattern Ed. Reg. Mitochondrial Transcr. Cryptobiid Trypanoplasma Borreli *13*, 5089–5098.

Maas, S. (2012). Posttranscriptional recoding by RNA editing. Adv. Protein Chem. Struct.

Biol. *86*, 193–224.

Nakamura, K., Oshima, T., Morimoto, T., Ikeda, S., Yoshikawa, H., Shiwa, Y., Ishikawa, S., Linak, M.C., Hirai, A., Takahashi, H., et al. (2011). Sequence-specific error profile of Illumina sequencers. Nucleic Acids Res. *39*, e90–e90.

Ochsenreiter, T., and Hajduk, S.L. (2006). Alternative editing of cytochrome c oxidase III mRNA in trypanosome mitochondria generates protein diversity. EMBO Rep. *7*, 1128–1133.

Ochsenreiter, T., Anderson, S., Wood, Z.A., and Hajduk, S.L. (2008). Alternative RNA Editing Produces a Novel Protein Involved in Mitochondrial DNA Maintenance in Trypanosomes. Mol. Cell. Biol. *28*, 5595–5604.

Pandit, S., Zhou, Y., Shiue, L., Coutinho-Mansfield, G., Li, H., Qiu, J., Huang, J., Yeo, G.W., Ares, M., and Fu, X.-D. (2013). Genome-wide Analysis Reveals SR Protein Cooperation and Competition in Regulated Splicing. Mol. Cell *50*, 223–235.

Ridlon, L., Skodova, I., Pan, S., Lukeš, J., and Maslov, D.A. (2013). The Importance of the 45 S Ribosomal Small Subunit-related Complex for Mitochondrial Translation in Trypanosoma brucei. J. Biol. Chem. *288*, 32963–32978.

Sharma, M.R., Booth, T.M., Simpson, L., Maslov, D.A., and Agrawal, R.K. (2009). Structure of a mitochondrial ribosome with minimal RNA. Proc. Natl. Acad. Sci. *106*, 9637–9642.

Shaw, J.M., Feagin, J.E., Stuart, K., and Simpson, L. (1988). Editing of kinetoplastid mitochondrial mRNAs by uridine addition and deletion generates conserved amino acid sequences and AUG initiation codons. Cell *3*, 401–411.

Smith, T.F. (1981). Identification of common molecular subsequences. J. Mol. Biol. *147*, 195–197.

Valach, M., Moreira, S., Kiethega, G.N., and Burger, G. (2014). Trans-splicing and RNA editing of LSU rRNA in Diplonema mitochondria. Nucleic Acids Res. *42*, 2660–2672.

Worthey, E.A. (2003). Comparative analysis of editosome proteins in trypanosomatids. Nucleic Acids Res. *31*, 6392–6408.

Part 2:

Biological significance of uridine insertion and deletion RNA editing analysis

Hereby mentioned novel method has been so far used for analysis of mitochondrial genome of *Perkinsela*, which is, besides being an unique non-photosyntetic eukaryotic endosymbiont, an early branching kinetoplastid with the oldest U-indel editing system studied so far. Results of this project are summarized in following manuscript, which has been submitted for publication.

Gene loss and error-prone RNA editing in the mitochondrion of *Perkinsela*, an endosymbiotic kinetoplastid

Vojtěch David [1,2,§], Pavel Flegontov [1,3,§], Evgeny Gerasimov [4], Goro Tanifuji [5,†], Hassan Hashimi [1,2], Maria D. Logacheva [6], Shinichiro Maruyama [5,@], Naoko T. Onodera [5], Michael W. Gray [5,7], John M. Archibald [5,7], and Julius Lukeš [1,2,7,*]

[1] *Institute of Parasitology, Biology Centre, Czech Academy of Sciences, České Budějovice (Budweis), Czech Republic*

[2] *Faculty of Sciences, University of South Bohemia, České Budějovice (Budweis), Czech Republic*

[3] *Life Science Research Centre, Faculty of Science, University of Ostrava, Ostrava, Czech Republic*

[4] *Faculty of Biology, Lomonosov Moscow State University, Moscow, Russia*

[5] *Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Canada*

[6] *Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia*

[7] *Canadian Institute for Advanced Research, Toronto, Canada*

[§] These authors contributed equally

[†] Present address: Faculty of Life and Environmental Sciences, University of Tsukuba, Tsukuba, Japan

[@] Present address: Graduate School of Life Sciences, Tohoku University, Sendai, Japan

*Author for correspondence: Institute of Parasitology, Branišovská 31, 37005 České Budějovice, Czech Republic; e-mail: jula@paru.cas.cz

Abstract

Perkinsela is an enigmatic early-branching kinetoplastid protist that lives as an obligate endosymbiont inside *Paramoeba* (Amoebozoa). We have sequenced the highly reduced mitochondrial genome of *Perkinsela*, which possesses only six protein-coding genes (*cox1*, *cox2*, *cox3*, *cob*, *atp6*, and *rps12*), despite the fact that the organelle itself contains more DNA than is present in either the host or endosymbiont nuclear genomes. An *in silico* analysis of two *Perkinsela* strains showed that mitochondrial RNA editing and processing machineries typical of kinetoplastid flagellates are generally conserved, and all mitochondrial transcripts undergo U-insertion/deletion editing. Canonical kinetoplastid mitochondrial ribosomes are also present. We have developed software tools for accurate and exhaustive mapping of RNA-seq reads having extensive U-insertions/deletions, allowing a detailed investigation of RNA editing via deep sequencing. With these methods we show that up to 50% of reads for a given edited region

contain errors of the editing system or, less likely, correspond to alternatively edited transcripts.

Key words: mitochondrion, *Perkinsela*, *Paramoeba*, RNA editing, alternative editing, NADH dehydrogenase

Introduction

Kinetoplastids are a diverse, widespread, and ecologically significant group of protists, some of which are devastating human parasites. Kinetoplastids have been the focus of intense research mainly because of the medical importance of *Leishmania* and *Trypanosoma* species, and have been shown to exhibit a variety of unique cellular and molecular features, including RNA editing, mRNA *trans*-splicing, and genes arranged in polycistronic arrays (Verner et al. 2015). However, relatively little is known about the origin and evolution of these features across the full breadth of kinetoplastid diversity, despite the fact that there is tremendous species richness in both terrestrial, obligatorily parasitic trypanosomatids (Maslov et al. 2013) and free-living marine bodonids (de Vargas et al. 2015).

Insertion and/or deletion of uridine (U) residues into/from the mitochondrial (mt) mRNAs of kinetoplastids was the first type of RNA editing to be discovered (Benne et al. 1986). A plethora of post-transcriptional modifications has subsequently been described in organisms ranging from bacteria to plants and humans (for review see Maas 2012). RNA editing events include various insertions and deletions of single or multiple residues as well as base modifications and replacements, and occur in both non-coding and protein-coding RNAs transcribed from nuclear and/or organellar genomes (Gott and Emeson 2000; Gray 2003). Numerous types of conversion editing have been implicated in a wide range of cellular processes including embryonic development of the brain (Li and Church 2013) and cancer (Avesson and Barry 2014).

While RNA editing seems to be particularly abundant in the mitochondria and plastids of land plants (Takenaka et al. 2013), U insertion/deletion (U-indel) RNA editing is at present confined to the mitochondria of kinetoplastids (for review see Hashimi et al. 2013; Aphasizhev and Aphasizheva 2014; Verner et al. 2015) and their sister clade Diplonemea (Marande and Burger 2007; Kiethega et al. 2013; Valach et al. 2013). U-indel editing is the most complex form of RNA editing known. Multiple sites within most transcripts are edited, with some mRNAs

edited over their entire length (so-called pan-editing). In the model kinetoplastid *Trypanosoma brucei*, more than 70 different proteins have been shown to be incorporated into numerous dynamic editing complexes (Hashimi et al. 2013; Aphasizhev and Aphasizheva 2014), and up to a thousand different small RNA molecules, called guide (g) RNAs, act as templates that define editing sites along a cognate mRNA (Kozlowsky et al. 2013).

Another unusual feature of kinetoplastid mitochondria is the structure and composition of their ribosomes. In *T. brucei*, 129 mitochondrial ribosomal proteins are nucleus-encoded and targeted to the organelle post-translationally (Zíková et al. 2008a). Only a single ribosomal protein, RPS12 (Aphasizheva et al. 2013), and two rRNAs are encoded in the mitochondrial genome. The bulk of the mtDNA (or kinetoplastid (k) DNA) of kinetoplastids is made up of minicircles encoding gRNA genes (Aphasizhev and Aphasizheva 2014). The 9S and 12S mitochondrial rRNAs of *T. brucei* are highly truncated and lack several conserved domains that are functionally significant in other eukaryotes (Sloof et al. 1985). Their transcription is developmentally regulated and they are 3′-polyuridylylated (Adler et al. 1991). Determination of the high-resolution three-dimensional structure of a protein-rich, rRNA-poor mitochondrial ribosome of a related species, *Leishmania tarentolae*, was instrumental in explaining the shrunken mitochondrial rRNAs (Sharma et al. 2009).

We are studying the molecular biology and evolution of the early-branching kinetoplastid *Perkinsela* sp. Members of this morphologically divergent, flagellum-lacking genus live as obligate endosymbionts inside amoebae (Dyková et al. 2000), to our knowledge the only known example of a co-evolving endosymbiotic relationship between two non-photosynthetic eukaryotes. The kinetoplastid-amoeba symbiotic system appears to have emerged early in the evolution of the genus *Paramoeba* (Young et al. 2014). The closest known relative of *Perkinsela* is the fish ectoparasite *Ichthyobodo necator*, with both of these kinetoplastids belonging to the Prokinetoplastina clade (Moreira et al. 2004), currently represented by a relatively small number of species in rRNA databases (Lukeš et al. 2014). Within the confines of the host amoeba cytoplasm, *Perkinsela* is sometimes referred to as the ‘parasome’ or ‘*Perkinsela amoebae*-like organism (PLO)’. Amoeba hosts include free-living and facultatively parasitic marine amoebae of the genera *Paramoeba* and *Janickina* (Dyková et al. 2008; Kudryavtsev et al. 2011; Feehan et al. 2013; Young et al. 2014). The *Perkinsela* strains

studied here are associated with *Paramoeba pemaquidensis*, the causative agent of amoebic gill disease, which results in considerable mortality at marine fish farms (Young et al. 2008; Mitchell and Rodger 2011).

Using *Perkinsela* and *Paramoeba* genomic and transcriptomic data (Tanifuji et al. unpubl. data), we have assembled the mitochondrial genomes of *Perkinsela* strains CCAP1560/4 and GillNOR1/I, and characterized their overall structure and expression with particular attention to RNA editing. Furthermore, we have predicted the composition of their respiratory chain complexes, as well as proteins involved in RNA editing, processing, and translation. We show that the mitochondrial genome of *Perkinsela*, composed of a huge number of fragments with terminal repeats, has undergone considerable reduction in gene content, and that all detected protein-coding transcripts undergo extensive U-indel RNA editing. While most proteins associated with RNA editing and with mitochondrial ribosomes in *T. brucei* are recognizable *in Perkinsela*, mitochondrial rRNAs were not found despite an exhaustive search, suggesting that they are fragmented and/or extremely divergent, similar to the situation observed in the related diplonemid, *Diplonema papillatum* (Valach et al. 2013).

Importantly, we have conducted what is to our knowledge the first investigation of U-indel-edited mitochondrial transcripts based on deep transcriptome sequencing, and developed software tools for accurate mapping of extensively edited reads. Since the discovery of this type of RNA editing in 1986, editing mechanisms have been unraveled via targeted sequencing on a clone-by-clone basis (Blum et al. 1990; Maslov and Simpson 1992; Landweber et al. 1993). Recently, deep sequencing of gRNA libraries in *T. brucei* (Koslowsky et al. 2013; Madina et al. 2014) has uncovered an unexpected degree of complexity and disorder inherent in gRNA-mediated editing. By deep sequencing of mRNAs, we have unveiled an even greater level of complexity in the form of 'misediting' (Sturm et al. 1992; Arts et al. 1993; Maslov et al. 1994), although we have not detected alternative translatable mRNAs of considerable abundance.

Materials and Methods

Cell culture

Paramoeba pemaquidensis strain CCAP1560/4 was obtained from the CCAP (Culture

Collection of Algae and Protozoa). Cells were grown on MYS medium (0.01% malt extract and 0.01% yeast extract in artificial seawater, solidified with 1.5% agar) (Page 1973). *P. pemaquidensis* strain GillNOR1/l was obtained from the culture collection of the Institute of Parasitology, Czech Academy of Sciences, and was grown on MY75S medium (0.01% malt extract and 0.01% yeast extract in artificial seawater, solidified with 2.0% agar). Both strains were grown in the dark at 20ºC.

Microscopy

P. pemaquidensis GillNOR1/I strain, carrying *Perkinsela* and feeding on diverse bacteria, was grown on agar plates, and the cells were collected as described previously (Dyková et al. 2000). Cells were prepared for phase contrast, DAPI epifluorescence light microscopy, and high pressure freezing transmission electron microscopy following protocols described elsewhere (Yurchenko et al. 2008; Votýpka et al. 2014).

Paramoeba pemaquidensis *sequencing*

Two strains of *P. pemaquidensis* with their respective *Perkinsela* endosymbionts were used in this study. Strains CCAP1560/4 (Page 1976) and GillNOR1/I were isolated from gills of Atlantic salmon captured in the waters of Wales and Tasmania, respectively (Dyková et al. 2005). As it is currently impossible to separate *Perkinsela* from its amoeba host, or to separate their DNAs, we prepared and sequenced total genomic and polyA-enriched transcriptomic libraries (Supporting Table 1) from the strains CCAP1560/4 and GillNOR1/I.

Mitochondrial genome assembly

Raw DNA sequence reads from all sequencing platforms were filtered and trimmed to ensure quality, depleted of adapter sequences, and paired-end reads were merged using the CLC Genomics Workbench v.6.5 (Supporting Table 1). The mitochondrial genomes of both *Perkinsela* strains were assembled from combined next generation sequencing reads with the Newbler assembler (GS De Novo Assembler v.2.9): from single 454, mate pair and paired-end Illumina HiSeq reads in the case of strain CCAP1560/4 and from paired-end Illumina MiSeq

reads for strain GillNOR1/I (Supporting Table 1). A number of assembly parameters were tested with the goal of maximizing mitochondrial contig size. Manual analysis of a graph of alternative contig connections (produced by Newbler) with an in-house visualizing script was used to close gaps and assemble long repetitive regions. RNA-seq assemblies were performed with Trinity software (Haas et al. 2013).

Gene identification

Proteins predicted from the *Perkinsela* mitochondrial contigs and translated transcriptomic assemblies were initially identified using the HMMER3 software. Available kinetoplastid and diplonemid mitochondrial protein sequences were used for the construction of Hidden Markov models (HMMs), which were subsequently used as queries against conceptual translations of the genome and transcriptome assemblies, with an E-value cutoff of 10^{-1}. Best-scoring hits were compared to the NCBI (nr) protein database in order to filter out host and bacterial proteins. Additional mitochondrial contigs were then identified with BLASTn using typical repetitive regions from contigs identified in the first step. *Perkinsela* nucleus-encoded proteins associated with mitochondrial oxidative phosphorylation, RNA editing and processing machineries, and mitochondrial translation were identified using HMMs (with an E-value cutoff of 10^{-10}) based on the corresponding orthologous groups from the OrthoMCL database re-aligned using MUSCLE (Edgar 2004). We classified as 'missing' all *Perkinsela* hits with an E-value > 10^{-50} that did not recover the corresponding *T. brucei* ortholog as the best hit in reciprocal BLASTp (with an E-value cutoff of 10^{-3}). *Perkinsela* hits with an E-value < 10^{-50} and without a suitable reciprocal BLASTp hit were aligned with their supposed orthologs in trypanosomatids. All protein alignments were performed using MUSCLE with default settings, and checked manually.

Searching for rRNAs

The following approaches were used to identify mitochondrial rRNA genes in *Perkinsela*. First, BLAST searches with known kinetoplastid and diplonemid homologs as queries were performed with an E-value cutoff of 10^{-5}. Second, transcribed regions on contigs not assigned to the host or *Perkinsela* nuclear genomes (Tanifuji et al. unpubl. data) were selected for further inspection. Third, reads containing the large subunit (LSU) peptidyl transferase core sequence

(ACCTCGNTGT) conserved in *Diplonema* (Valach et al. 2013) were assembled separately using the CLC Genomics Workbench v.6.5. The top candidates from each of these searches were subjected to manual secondary structure folding, with terminal hairpin prediction performed using the mFOLD thermodynamic folding application (http://mfold.rit.albany.edu/?q=mfold/RNA-Folding-Form). Default options were used to construct guiding graphs for manual secondary structure prediction (except for the 'Loop max' option, which was restricted to 10, 20, and 30 nucleotides). Structures were assessed by similarity to those of *Leishmania* LSU and SSU and *Diplonema* LSU rRNAs (Sharma et al. 2009; Valach et al. 2013).

Bowtie2 modification

Bowtie2 is an open-source fast and accurate short read mapper written in the C++ programming language (Langmead and Salzberg 2012). It uses a fast multiseeding procedure to find candidate alignment locations, and then proceeds with the Smith-Waterman algorithm to create the best gapped alignment. For additional speed, Bowtie2 implements the Smith-Waterman alignment algorithm with SIMD (single instruction, multiple data), allowing it to fill several dynamic programming table cells by executing a single instruction (Farrar 2007)**.** However, Bowtie2 uses a scoring system with equal gap open and extension penalties for the four nucleotides, A, G, T, and C. We modified Bowtie2 to facilitate accurate alignment of U insertion/deletion-edited RNA reads, while preserving mapping speed and accuracy. Edited reads of the mitochondrial genomes of kinetoplastids have U-indels only, therefore they can be aligned correctly when gap penalties for T (corresponding to U in RNA) are different from those for A, G, and C.

We modified the Bowtie2 v.2.0.2 source code and implemented a more complex nucleotide-specific gap scoring system that allows separate penalty values for A, G, T, and C using the --rdg-X and --rfg-X options on the command-line (for gaps in the read and reference, respectively, where X can be A, T, G, or C). Source code modifications were made both in the aligner module, which fills the dynamic programming table, and in the backtrack module of the program, which reconstructs the alignment using the filled dynamic programming table. Branch and array access instructions were minimized for each step, ensuring minimal time cost for more complex scoring. Using this scoring matrix, U-indel edited reads can be successfully mapped and accurately aligned with a low T-indel penalty and high penalties for other nucleotides.

Additional modifications of the alignment procedure were necessary in order to let reads have a gap/mismatch after the last nucleotide of the read (option --gbar 0). This option allows the seeding of more extensively edited reads on a pre-edited RNA sequence and prevents a significant fraction of edited reads from being discarded.

T-aligner

T-aligner is a new software program written for the purpose of this study and using the C++ programming language with the source code posted online (Github). T-aligner combines the optimal but time-consuming Smith-Waterman alignment with fast hash-based exact matching. The algorithm is specially designed to map extensively edited RNA-seq reads on pre-edited transcript references, also called cryptogenes. Exact matches between short substrings (seeds) are first found using a hash table. A local optimal alignment is then produced with the Smith-Waterman algorithm, allowing 'T,-' and '-,T' gaps with zero penalty, thus taking into account the biological mechanism of U-indel RNA editing. The general T-aligner workflow is as follows (Supporting Fig. 1): a fixed seed is chosen in a never-edited or universally edited 3′-terminus of the transcript (or editing domain in appropriate cases). Reads are then mapped if they satisfy the following criteria: (i) they contain the seed; (ii) at least part of the read lies 5′ to the seed; (iii) the alignment may contain any number of U-indels of any length; (iv) the alignment contains no other indels and no or few mismatches. After the alignments are produced, T-aligner classifies all editing events (U insertion or U deletion) and clusters the reads into three groups: (i) those matching the reference sequence, (ii) those matching the putative main 'editing pathway' (i.e., the user-defined final edited product) and (iii) all other reads containing alternative editing events. Reads matching the main pathway are defined as follows: (i) those with no additional edited sites compared to the main pathway; (ii) reads with insertions/deletions that are shorter or equal in size to those in the main pathway; and (iii) reads in which all sites are edited in the same direction as in the main pathway (e.g., insertion in the main pathway versus insertion in a sequence read). Reads in violation of any of these conditions are placed in the 'alternative editing' group. Sequence reads that are exact substrings of other reads are then merged into 'editing intermediates'. The support value associated with an editing intermediate can be used to determine the most abundant sequences, which is useful when examining alternative editing. All

sequences clustered into the ‘reference’ and ‘main pathway’ groups are assigned a support value equal to the number of reads in each group. For each sequence from the ‘alternative’ group, support is determined as follows: reads falling into the ‘reference’ and ‘main pathway’ groups are excluded; if a read is unique – i.e., can be included as a substring in at most one longer read – it adds 1 to the support value; if a read supports k>1 alternative sequences, it adds 1/k to a support value for each sequence.

Read mapping and analysis of U-indel RNA editing

Bowtie2 v.2.0.2 or v.2.1.0 mapping software was used for both DNA and RNA-seq reads utilizing the end-to-end mapping mode, the ‘very sensitive’ options, and default alignment scoring. In order to produce precise alignments in extensively edited regions, we used a modification of Bowtie2 v.2.0.2 with the base-specific indel penalties described above. The following set of options was routinely used: (i) high gap opening and extension penalties of 10 for A, G, C in the reference and individual sequence reads (--rfg 10,10 --rdg 10,10); (ii) minimal gap opening and extension penalties of 1 for T or A (depending on transcript orientation) in the reference and reads (--rfg-T 1,1 --rdg-T 1,1 or --rfg-A 1,1 --rdg-A 1,1); (iii) high mismatch penalty equal to 18 (--mp 18); (iv) options allowing terminal mismatches (--gbar 0 --dpad 50), and (v) other options (--end-to-end -D 20 -R 3 -N 1 -L 14 -i S,1,0.50 --score-min L,0,-2). Reads mapped to the edited regions were manually checked before further processing. Poor-quality alignments, especially those introducing large gaps, were not considered. Alignments made with Bowtie2 were cut into overlapping windows, and examined to find sequences appropriate for seeding further read mappings with T-aligner.

One to three iterations of read mapping with T-aligner (with the original seed shifting in the 3′ to 5′ direction) were enough to cover the whole transcript or its edited region, and then reconstruct the main editing pathway. Repeating T-aligner-assisted read mapping with prior knowledge of the main edited product allowed us to reveal and quantify alternative editing products.

Northern blotting

Northern analysis of *cox2* was performed as previously described (Kafková et al. 2012). Briefly,

10 µg of RNA isolated from *Perkinsela* strain GillNOR1/I and *T. brucei* strain 29-13 was run on a high resolution 4%-acrylamide/7M urea gel and transferred onto a Zeta-probe membrane (Bio-Rad). The membrane was subsequently probed with 5′-^{32}P-end-labelled oligonucleotides corresponding to the antisense (5′-CCCTTTCAACACGTCAAAACAAGC-3′) and sense (5′-GCTTGTTTTGACGTGTTGAAAGGGC-3′) pre-edited sequence of the 5′-end – i.e., the last to be processed – of the larger 3′-edited domain. The oligonucleotides were also used to probe dot blots of serially-diluted, denatured PCR products amplified from this same region to demonstrate that the two probes are equally sensitive.

Results and Discussion

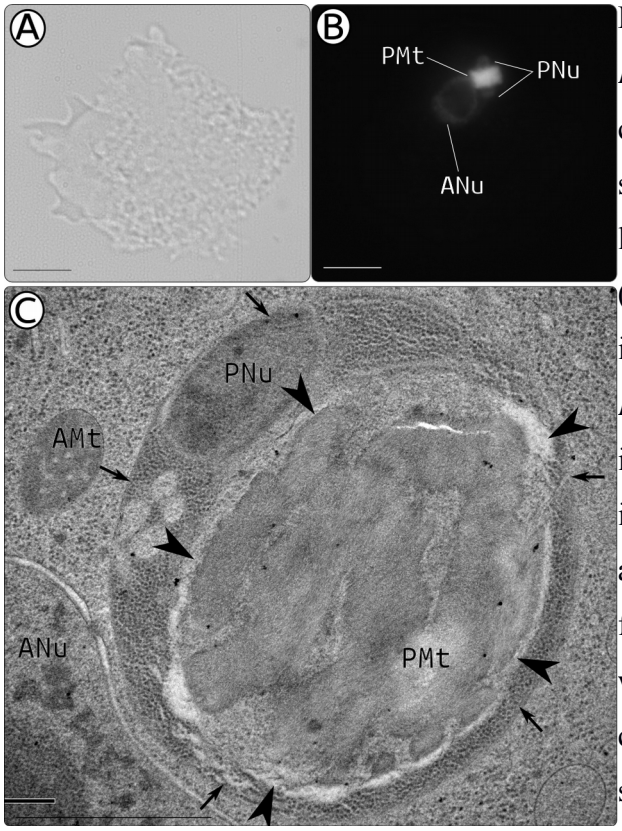


*Figure 1. Phase contrast (**A**), DAPI staining (**B**), and high-pressure freezing transmission electron microscopy (**C**) of Perkinsela strain GillNOR1/I. The scale bars in panels A and B correspond to 10 µm, the bar in panel C corresponds to 1 µm. Small arrows mark the single membrane separating Perkinsela from the amoeba host cytoplasm, and arrowheads mark the outer mitochondrial membrane of Perkinsela. Abbreviations: ANu, amoeba nucleus; PNu, Perkinsela nuclei; PMt, Perkinsela mitochondrion.*

Perkinsela *mitochondrial genome structure*

Perkinsela can be visualized in the *Paramoeba* cell (Fig. 1A) by DAPI staining of DNA, which shows that the endosymbiont is invariably located in the perinuclear region of the amoeba (Figs. 1B and 1C). Interestingly, based on the intensity of DAPI staining, it appears that *Perkinsela* harbors a larger amount of DNA in its mitochondrion (= kDNA) than in the rather inconspicuously stained nuclei of *Perkinsela* and *Paramoeba* (Fig. 1B). High-pressure freezing transmission electron microscopy, which optimally preserves fine structure, confirmed an earlier observation obtained by standard electron microscopy (Dyková et al. 2000; Tanifuji et al. 2011), namely that the single mitochondrion of *Perkinsela* is packed with kDNA strands arranged in parallel electron-dense layers (Fig. 1C). Indeed, since both DAPI-staining and electron microscopy

show that the kDNA and the single mitochondrion occupy most of the *Perkinsela* cell volume and that the organellar genome constitutes the most abundant DNA in this endosymbiont-host system, it is likely that this inflated genome has a very high copy number.

Trypanosomatid mtDNAs studied so far invariably have a complement of 18 protein-coding genes and two rRNA genes (Verner et al. 2015). However, individual flagellate species differ in gene regions at which post-transcriptional U-indel editing takes place (Lukeš et al. 1994; Simpson and Maslov 2006). Out of this conserved gene set, we identified just six protein-coding genes (*cox1, cox2, cox3, cob, atp6, rps12*) on three assembled mitochondrial contigs in *Perkinsela*, which are similar in both studied strains (Fig. 2). Due to the presence of highly repetitive sequences at the ends of these contigs, we were unable to extend them significantly or connect them with other non-repetitive contigs using next generation sequencing reads, even with manual analysis of a contig graph produced by the assembler software GS De Novo Assembler v.2.9 (Newbler).

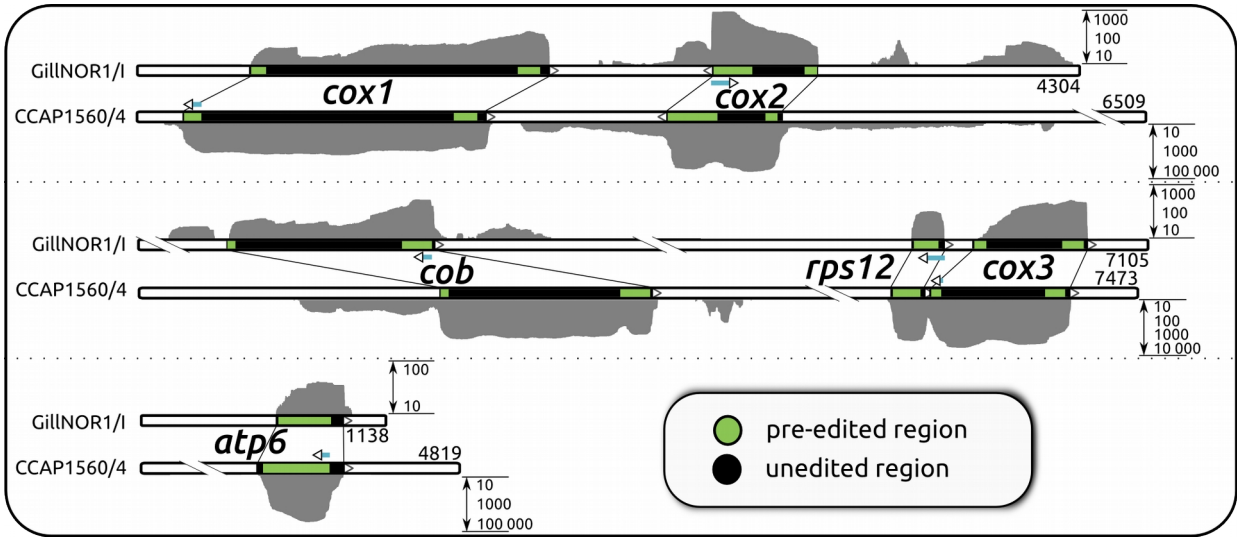


Figure 2.Gene-bearing mitochondrial scaffolds identified in Perkinsela strains CCAP1560/4 and GillNOR1/I. Transcript regions undergoing RNA editing are shown in green. Scaffold 1 contains cox1 and cox2 genes in reverse orientation; scaffold 2 contains cob, and closely spaced rps12 and cox3 genes in the same orientation; scaffold 3 contains only the atp6 gene. While most transcripts are edited in separate regions at their ends, rps12 and atp6 are edited over almost their entire length, i.e., pan-edited. Gene regions used for detailed mapping of alternatively edited reads (Supporting Fig. 8) are shown with teal arrows, also indicating the direction of RNA editing in these regions. For the GillNOR1/I strain, coverage with strand-specific RNA-seq reads (with 'U-indel optimized' settings) for each transcript is shown in the sense orientation only; for both sense and antisense reads plotted, see Supporting Fig. 2. For CCAP1560/4, RNA-seq reads were non-strand-specific. Coverage (gray blocks) is plotted in logarithmic scale. Absolute values of coverage are markedly different for the two strains due to different sequencing approaches used (Supporting Table 1).

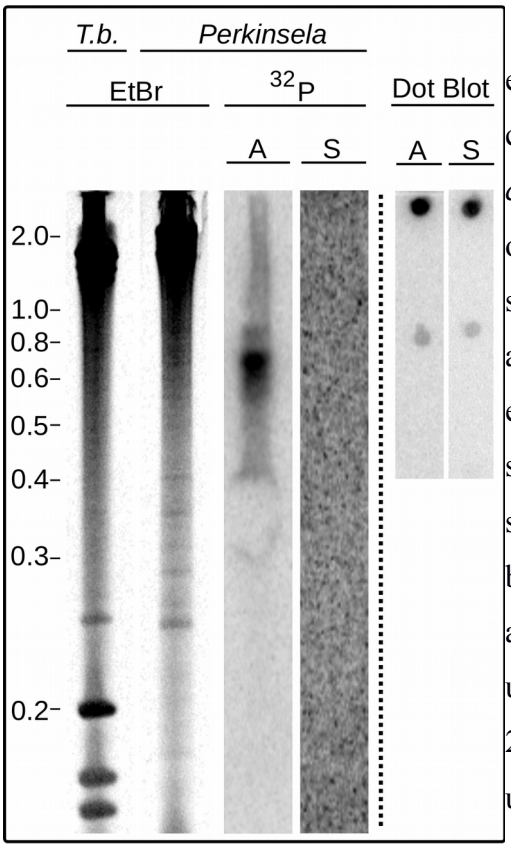


Figure 3. Northern blot with anti-sense and sense probes for the cox2 transcript, showing that only a single-strand of the cox2 gene is transcribed. Total RNA from T. brucei (T.b.) and Perkinsela resolved on a denaturing gel is visualized by ethidium bromide (EtBr) stain. The signal from the anti-sense cox2 probe is shown in the lane labelled "^{32}P, A", while the sense-probed Northern membrane is shown in the lane labelled "^{32}P, S". Dot blots simultaneously probed with anti-sense (A) and sense (S) probes are shown on the right, with increasing dilution from top to bottom of a denatured plasmid bearing an insert corresponding to the probed sequence.

This highly reduced set of a half-dozen genes encodes subunits of three respiratory complexes: *cob* of complex III (ubiquinone-cytochrome *c* oxidoreductase); *cox1*, *cox2* and *cox3* of complex IV (cytochrome *c* oxidase); and *atp6* of complex V (ATP synthase), suggesting a functional respiratory chain. The apparent absence of respiratory complex I in *Perkinsela* (mtDNA-encoded subunits *nad1* thru *nad9* are missing) is further supported by the absence of the nucleus-encoded subunits of this complex (Supporting Table 2; see below). All six mtDNA-encoded protein-coding genes are transcribed (with varying transcript abundance) and undergo U-indel editing to slightly different degrees (Fig. 2, Table 1 and below). Long antisense transcripts were undetectable by Northern blotting, at least in the case of *cox2* (Fig. 3). Due to the extremely slow growth of *Paramoeba* in culture, we were not able to accumulate enough RNA for testing antisense transcription of other *Perkinsela* mitochondrial genes by Northern blotting, but mapping of strand-specific RNA-seq reads revealed no significant antisense transcripts in strain GillNOR1/I (Supporting Fig. 2).

Despite extensive searching, rRNA genes could not be identified by BLAST using known kinetoplastid and *Diplonema* rRNA genes as queries. Further candidate sequences were obtained from transcribed regions of the assembly not assigned to the *Perkinsela* or host nuclear genomes (Tanifuji et al. unpubl. data). In addition, *Perkinsela* reads containing the peptidyl transferase core motif ACCTCGNTGT conserved even in the highly diverged *Diplonema* LSU rRNA (Valach et al. 2013) were assembled, and resulting contigs were added to the list of putative rRNA sequences. All top candidates were subjected to a careful

											edited regions combined				5′ edited region (or full transcript for pan-edited genes)					3′ edited region						
strain	gene	contig ID (Fig. 2)	start	end	pre-edited length, nt	edited length, nt	size increase, %	Us in ORF	protein length	length, nt*	U insertions	U deletions	edited sites	start*	end*	length, nt*	U insertions	U deletions	edited sites	start*	end*	length, nt*	U insertions	U deletions	edited sites	edited regions length, 3′/5′
CCAP1560/4	*cob*	2	1350	2314	964	1136	18%	45%	370	169	187	16	82	16	50	35	41	3	14	797	930	134	146	13	68	3.8
GillNOR1/I			1019	1955	936	1125	20%	47%	370	170	186	12	81	12	47	36	45	4	15	794	927	134	141	8	66	3.7
CCAP1560/4	*cox1*	1	210	1584	1374	1567	14%	40%	521	185	198	6	85	18	95	78	77	3	31	1217	1323	107	121	3	54	1.4
GillNOR1/I			513	1878	1365	1589	16%	42%	521	171	202	5	83	21	88	68	77	2	30	1219	1321	103	125	3	53	1.5
CCAP1560/4	*cox2*	1	2950	2464	486	661	36%	51%	209	282	208	34	125	27	81	55	57	6	23	261	487	227	151	28	102	4.1
GillNOR1/I			3112	2625	487	710	46%	49%	209	234	206	34	122	6	62	57	58	8	24	291	467	177	148	26	98	3.1
CCAP1560/4	*cox3*	2	6515	7171	656	814	24%	44%	256	134	162	4	78	20	64	45	69	2	31	539	627	89	93	2	47	2.0
GillNOR1/I			6195	6819	624	801	28%	49%	255	151	164	5	78	18	71	54	71	3	31	549	645	97	93	2	47	1.8
CCAP1560/4	*rps12***	2	6510	6353	157	268	72%	55%	80	123	123	13	52	15	137											
GillNOR1/I			6015	6165	150	257	72%	55%	80	110	123	12	52	13	122											
CCAP1560/4	*atp6***	3	3747	4110	363	651	80%	0.613	197	309	318	30	154	29	337											
GillNOR1/I			623	944	321	625	94%	0.62	197	298	311	28	152	13	310											

* coordinates and length values correspond to pre-edited sequences

** pan-edited transcript, *i.e.,* edited throughout most of its length

Table 1. Statistics for edited mitochondrial mRNAs in Perkinsela. Only the main edited products are taken into account.

manual secondary structure prediction with the help of the mFOLD terminal hairpin prediction software, but no SSU or LSU rRNA-like folds were found (data not shown).

In light of the recent discovery of a split and edited LSU rRNA in *Diplonema*, a relative of *Perkinsela*, and the fact that the SSU rRNA of *Diplonema* remains unidentified (Valach et al. 2013), it seems likely that extreme divergence and/or fragmentation render the mitochondrial rRNAs of *Perkinsela* unrecognizable. We consider it highly improbable that the mitochondrial rRNA is genuinely absent, as upon RNA editing, detected transcripts have evolutionarily conserved open reading frames, implying the requirement of a functional ribosome to translate them into protein. Moreover, both universal and kinetoplastid-specific mitochondrial ribosomal proteins are generally conserved in *Perkinsela* (Supporting Table 2), and a ribosomal subunit gene (*rps12*) is also present in its organellar genome (Fig. 2).

Nucleus- and mitochondrion-encoded respiratory chain subunits

Using Hidden Markov models (HMM) constructed on the basis of trypanosomatid orthologs, the *Perkinsela* genomic contigs (Tanifuji et al. unpubl. data) were searched for mitochondrial proteins (see Materials and Methods for details). Since none of the nucleus-encoded subunits of the respiratory complex I (NADH dehydrogenase) were detected, we consider this component of

the respiratory chain missing in *Perkinsela* (Fig. 4; Supporting Table 2). This inference is in agreement with our failure to detect any of the mtDNA-encoded subunits of complex I in the mitochondrial contigs. The other respiratory complexes (II through V) that together mediate oxidative phosphorylation are apparently present in *Perkinsela* (Fig. 4 and Supporting Table 2). We conclude that in the mitochondrion of *Perkinsela* the respiratory chain is functional, with the missing complex I likely replaced by an as-yet-unidentified alternative NADH dehydrogenase. Although the distantly related *T. brucei* possesses both mitochondrial- and nucleus-encoded subunits of complex I, its function remains elusive, with a highly active alternative dehydrogenase substituting for the canonical biochemical activity (Verner et al. 2011; Surve et al. 2012). It thus seems that in kinetoplastids complex I is prone to loss and was eliminated in the early-branching *Perkinsela*.

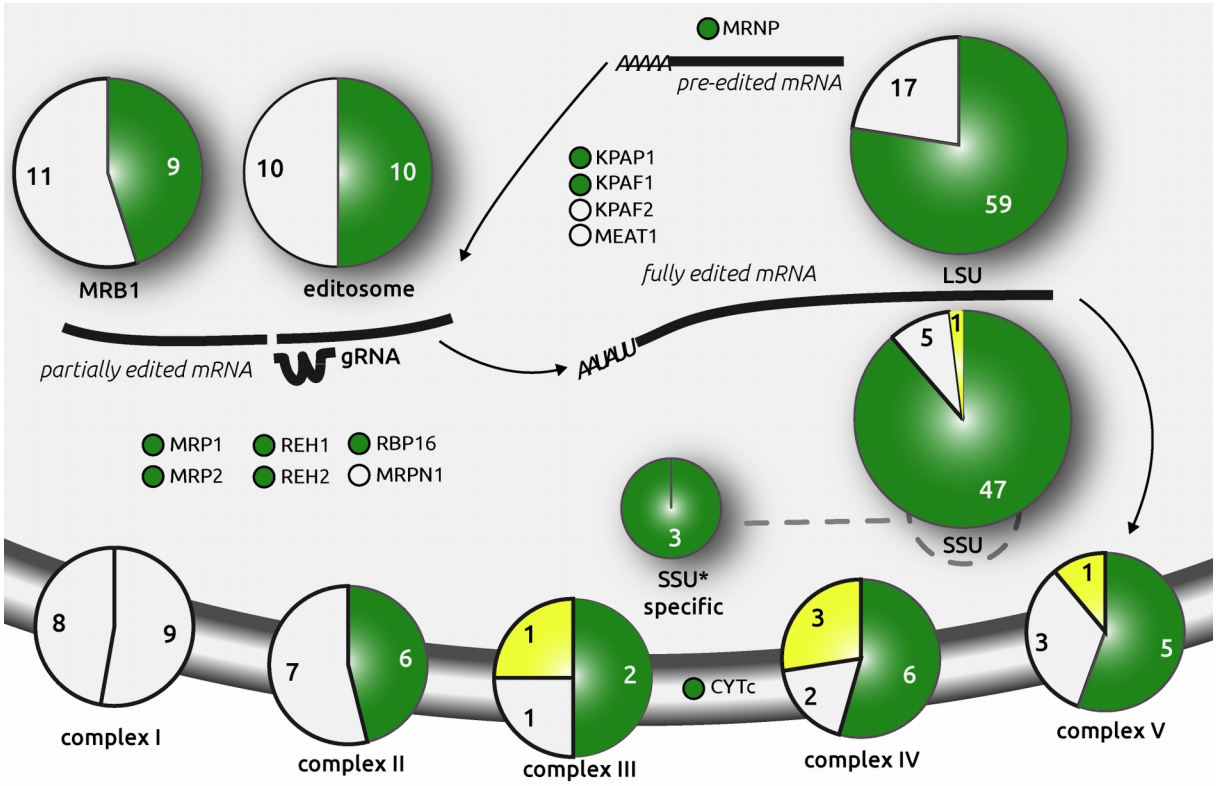


Figure 4. Conservation of respiratory chain subunits, RNA editing and processing factors, and mitochondrial ribosomal proteins in Perkinsela. Each complex is represented as a pie chart, and numbers indicate subunits analyzed in this study. Green color marks proteins identified in the Perkinsela genome (also listed in Supporting Table 2). Missing proteins are shown in white and proteins encoded in the mitochondrial genome in bright yellow areas. The left-hand section of the pie chart for respiratory chain complex I represents subunits encoded in the mitochondrial genomes of trypanosomatids but missing in Perkinsela. The following complexes are shown: respiratory chain complexes I-V, RNA Editing Core Complex (editosome), Mitochondrial RNA-binding complex (MRB1), large (LSU) and small (SSU) subunits of the mitochondrial ribosome, and proteins unique to the SSU subunit. A number of other proteins involved in mRNA/gRNA processing are also shown.*

RNA editing and processing complexes, mitochondrial ribosomes

Next, we verified the presence of nucleus-encoded genes for proteins imported into the *Perkinsela* mitochondrion using *T. brucei* as a reference. Despite its endosymbiotic lifestyle and large evolutionary distance from other kinetoplastid flagellates, *Perkinsela* has generally conserved kinetoplastid mitochondrial transcription and translation machineries, as well as a complex RNA editing machinery (Fig. 4; Supporting Table 2). The composition of these protein complexes is described below.

Transcription of the mitochondrial genome is performed by a dedicated single-subunit, phage T3/T7-like RNA polymerase (Grams et al. 2002), which is present in *Perkinsela*. In trypanosomatids, the formation of short A-tails on pre-edited mRNAs and long A/U-tails on fully-edited transcripts is controlled by kinetoplast poly(A) polymerase 1 (KPAP1), 3′-terminal uridylyl transferase (TUTase) KRET1, and their accessory factors KPAF1 and KPAF2, which together regulate mRNA translatability and stability (Aphasizheva et al. 2011). Except for KPAF2, all these nucleus-encoded and mitochondrion-targeted proteins are present and well conserved in *Perkinsela*. KRET1 also appends 3′-oligo(U) tails to rRNAs and gRNAs in trypanosomatids (Aphasizheva and Aphasizhev, 2010), so it seems reasonable to assume that this enzyme performs the same function in *Perkinsela*.

The core set of editing reactions in trypanosomatid mitochondria is executed by the RNA Editing Core Complex (RECC), also called the 20S editosome (Aphasizhev and Aphasizheva 2014). In the first step of the editing reaction, the cleavage of the mRNA at a mismatch between it and a hybridizing gRNA yields 5′- and 3′-mRNA fragments bridged by the gRNA, and is performed by one of three RECC endonucleases (Carnes et al. 2008). Remarkably, among these three endonucleases, only a homolog of the U-insertion-specific enzyme KREN2 was found in *Perkinsela*; KREN1 (the deletion-specific endonuclease) and KREN3 were not detected. KREN3 is known to act on the *cox2* transcript edited by a *cis*-gRNA located in its 3′-UTR in trypanosomatids (Golden and Hajduk 2005). Of the KREPB proteins (KREPB6 thru 8), which within RECC form dimers with the KREN endonucleases (Carnes et al. 2011), only KREPB6, which in *T. brucei* interacts with KREN3, was found. KREPB8 and KREPB7, which dimerize with KREN1 and KREN2, respectively, are apparently absent in *Perkinsela*. With regard to the

deletion of extraneous Us from the 5′ mRNA fragment (Ernst et al. 2009), the dedicated exonucleases KREX1 and KREX2 have predicted orthologs in *Perkinsela*. The KRET2 TUTase, responsible for adding Us to the 5′ mRNA fragment, and the insertion-specific RNA ligase KREL2, which reseals the two RNA fragments, were also found (Ernst et al. 2003; Aphasizhev and Aphasizheva 2011). The deletion-specific RNA ligase KREL1 is missing in *Perkinsela*. Of the accessory and structural RECC subunits (KREPA1 thru 6 and KREPB4 and 5), three are present whereas five seem to be missing (Supporting Table 2). The undetected orthologs were presumably replaced or have evolved beyond recognition in *Perkinsela*, or they are normally essential for editing transcripts encoding the numerous complex I subunits, which have been lost in this kinetoplastid.

In addition to RECC, which provides the core editing enzymatic activities, various other proteins and macromolecular complexes have been shown to play vital roles in editing. One example is the mitochondrial RNA-binding complex 1 (MRB1), a dynamic structure that binds and recruits gRNAs into the editing complex, processes massively edited mRNAs that require several gRNAs, and links RNA editing with mRNA tailing and translation machineries (Hashimi et al. 2013). Of six invariably recovered MRB1 subunits (Hashimi et al. 2008; Panigrahi et al. 2008; Weng et al. 2008; Ammerman et al. 2012), four are found in *Perkinsela*, including the crucial gRNA-binding subunits GAP1 and GAP2. The missing core subunits are MRB5390 and MRB8620 (Supporting Table 2). However, of 14 other putative editing complex members, only five are found, whereas TbRGG1, TbRGG2, MRB8170 and MRB4160 (Ammerman et al. 2012; Kafková et al. 2012) are missing in *Perkinsela* (Supporting Table 2). However, TbRGG3, which associates with MRB1 as well as other mitochondrial RNA binding proteins (McAdams et al. 2015), yields a hit. Hence, the same picture emerges as for the 20S editosome: the functional core of the MRB1 complex is mostly conserved between *Perkinsela* and its trypanosomatid relatives.

A separate small complex, a heterotetramer of RNA-binding proteins 1 and 2 (MRP1/MRP2) that stimulates annealing of gRNA and mRNA molecules (Schumacher et al. 2006; Zíková et al. 2008b), is also present in *Perkinsela*. The same is true for the RNA-binding protein 16 (RBP16), which interacts with both mRNA and gRNA and has a multifunctional role in mitochondrial RNA metabolism (Fisk et al. 2009). However, RNA processing endonuclease

mRPN1, involved in cleavage of long gRNA precursor transcripts (Madina et al. 2011), was not detected, suggesting that gRNA transcription patterns may profoundly differ between trypanosomatids and *Perkinsela* (we did not attempt to identify gRNA genes in the latter). Finally, both RNA editing helicases KREH1 and KREH2, likely required for unwinding the gRNA:mRNA duplex (Hashimi et al. 2008; Hernandez et al. 2010; Li et al. 2011), are detected in *Perkinsela*.

Ribosomes in trypanosomatid mitochondria contain extremely reduced rRNAs and have acquired a multitude of novel proteins, apparently to compensate for the loss of RNA domains (Sharma et al. 2009), or through protein 'accretion' by a neutral evolutionary mechanism (Lukeš et al. 2011). Thus, both ribosomal LSU and SSU contain dozens of trypanosomatid-specific proteins, but lack some of the universally conserved ones (Zíková et al. 2008a). Of 27 mitochondrial LSU proteins conserved throughout eukaryotes, 25 are found in *Perkinsela* (Supporting Table 2). Of 49 trypanosomatid-specific mitochondrial LSU subunits, only 15 could not be detected in *Perkinsela* (Supporting Table 2). This significant conservation between *Perkinsela* and trypanosomatids is also seen for the SSU ribosomal proteins: in the case of 10 subunits universally present in mitoribosomes, only one is missing, whereas just four out of 43 trypanosomatid-specific proteins could not be detected in *Perkinsela* (Supporting Table 2).

Another peculiar feature of the mitochondrial translation system in trypanosomatids is a separate 45S complex containing 9S SSU rRNA, termed SSU* (Maslov et al. 2007). The role of this complex remains elusive: it possibly provides an interface between the editing and translation machineries and is indispensable for the translation of some (e.g., *cob* and *cox1*) but not all (e.g., *rps12*) edited mRNAs (Ridlon et al. 2013). In *T. brucei*, the protein compositions of SSU and SSU* overlap substantially: 25 SSU* proteins are shared with SSU, with just three being unique. With the exception of two apparently missing subunits, all proteins shared between SSU and SSU* have been found in *Perkinsela*, and the same applies for all three SSU*-specific ones (Fig. 4, Supporting Table 2). In summary, proteins incorporated into RNA editing and processing as well as translation machineries are generally conserved in *Perkinsela*, despite its deep evolutionary separation from *T. brucei* and other trypanosomatids (Lukeš et al. 2014).

Analysis of edited RNA molecules

We carried out an in-depth analysis of RNA editing based on thousands of Illumina reads per gene for *Perkinsela* strain CCAP1560/4, greatly surpassing the limits of traditional methods. We also took advantage of lower coverage but longer read sequence data (up to 450 bp long) generated for the GillNOR1/I strain (see Supporting Fig. 3, Supporting Table 1). Preliminary analyses revealed that read mapping with such a high fraction of U-indels is problematic, as publicly available read mapping software was not designed for such applications. Our initial approach using the Bowtie2 v.2.0.2 mapper with low indel penalties resulted in alignments that required extensive manual improvement due to misalignments in regions with closely spaced U-indel sites (data not shown). In order to improve mapping of U-indel-rich reads, we modified the Bowtie2 v.2.0.2 software, introducing nucleotide-specific gap opening and gap extension penalties into the Smith-Waterman alignment module (see Materials and Methods). Mapping reads with strict penalties for gaps containing A, C or G but with relaxed penalties for gaps containing only U dramatically reduced the number of misalignments and improved the yield of edited reads (see Supporting Fig. 4). In the case of pan-edited transcripts or long editing domains, extra runs of mapping on partially edited templates were necessary to reconstruct the final edited product, as reads edited over the entire length lacked seeds long enough for initial mapping.

To overcome the problem of missing seeds, we developed a novel read mapping tool, T-aligner, based on the Smith-Waterman algorithm and designed to mimic the 3′-5′ progression of RNA editing in kinetoplastids. Initially a fixed seed is chosen in a never-edited or universally edited 3′-terminus of the transcript (or editing domain in appropriate cases), then reads are mapped and the final edited sequence reconstructed with the help of T-aligner (see Materials and Methods). At this stage, further iterations of read mapping are possible, shifting the seed in the 5′ direction. Using T-aligner, we identified mature edited transcripts in *Perkinsela* and investigated the extent to which alternative editing occurs.

RNA editing in *Perkinsela* resembles the system described in the model *Trypanosoma* and *Leishmania* species. However, the general distribution of editing sites (Fig. 2, Table 1), namely the fact that the 3′ and 5′ regions of genes usually contain separate editing domains, more closely resembles the situation in the bodonid *Trypanoplasma borreli*, for which only a

few genes and transcripts have been sequenced (Lukeš et al. 1994). Interestingly, in the case of the *Perkinsela cox2*, we show that the 5′ domain is edited prior to the 3′ domain, despite the canonical 3′ to 5′ progression of U-indel editing inside these domains (Maslov and Simpson, 1992). Upon inspection of the longest read fraction, we observed no reads in which the 3′ domain is at least partially edited but the 5′ domain is not (Supporting).

In total, fully edited versions of six transcripts have 1,196 Us inserted and 103 Us deleted at 576 distinct edited sites in the *Perkinsela* CCAP1560/4 strain, and 1,192 Us inserted and 92 Us deleted at 568 edited sites in the GillNOR1/I strain (Table 1). Alignments of edited and pre-edited mRNAs, their translation, and trees built for predicted proteins and their kinetoplastid orthologs are shown for *cox2* in Fig. 5 (and for the other five mitochondrial genes in Supporting Fig. 6). Finding a protein with an expected length and an expected position in a phylogenetic tree constitutes strong *in silico* evidence that the predicted translation product from a reconstructed edited mRNA sequence is most probably correct. The divergence of editing patterns between the two studied isolates (Figs. 5A and 5B; Supporting Fig. 6) is similar to that observed among various species of trypanosomatids (Landweber and Gilbert 1993), and

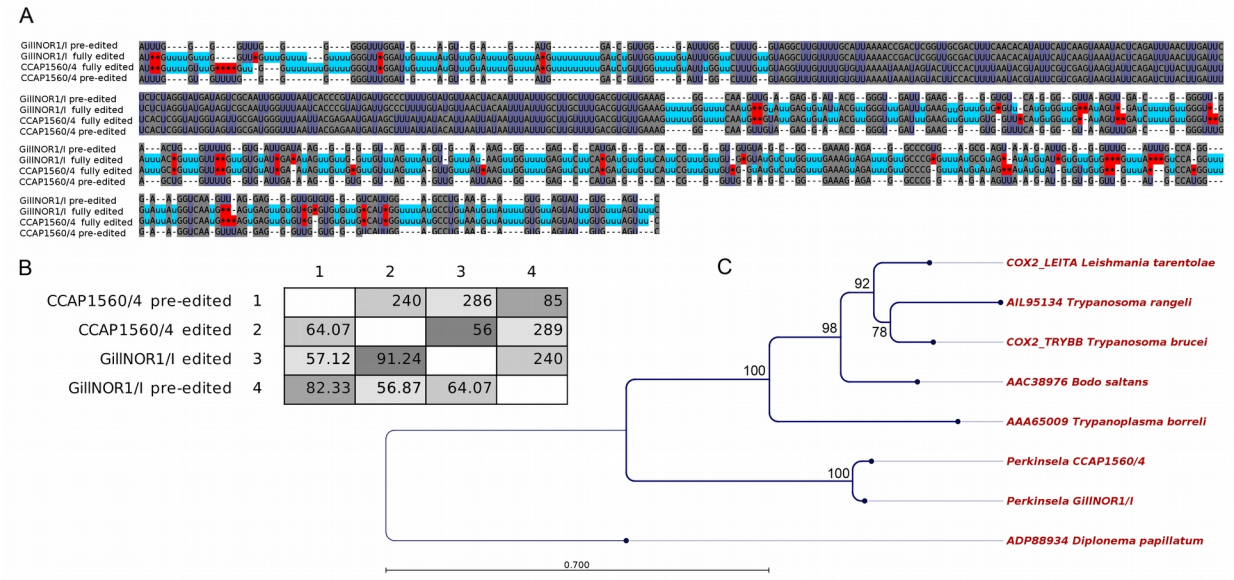


Figure 5. U-indel editing in the cox2 mRNA of Perkinsela strains CCAP1560/4 and GillNOR1/I. ***A****. Alignment of edited and pre-edited transcript sequences. U-insertions/deletions are highlighted in light blue and red, respectively.* ***B****. Pairwise percent identities (in the lower left half of the matrix) and numbers of different positions (upper right) between edited/pre-edited sequences of both strains.* ***C****. A maximum likelihood unrooted tree of COX2 proteins of Perkinsela, other kinetoplastids and Diplonema papillatum used as an outgroup. The tree was constructed with the following settings: WAG+Γ substitution model, neighbor-joining starting tree, 1000 bootstrap replicates. Branches supported by bootstrap values >70% are shown with thicker lines. Scale bar shows inferred number of amino acid substitutions per site.*

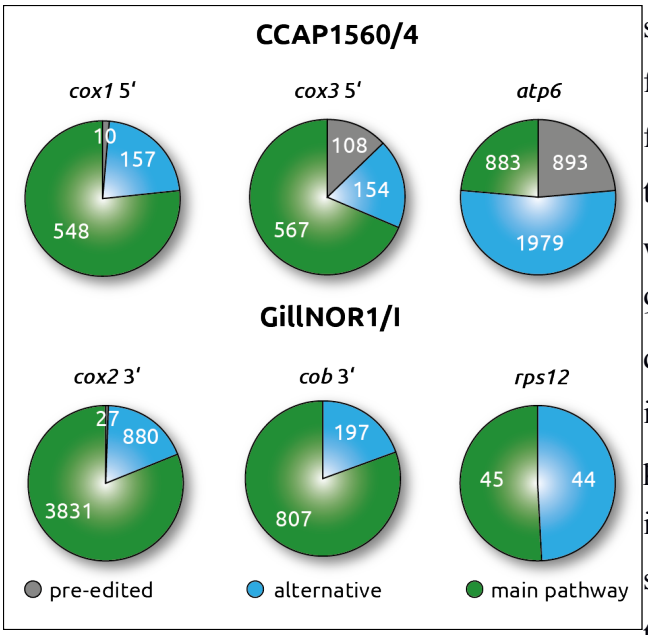


Figure 6. Pie charts illustrating counts of reads matching the main editing pathway, the pre-edited sequence, and alternatively edited reads in Perkinsela strains CCAP1560/4 and GillNOR1/I.

sequences become noticeably less divergent following RNA editing, as shown in Fig. 5B for *cox2*: pre-edited mRNA sequences of the two *Perkinsela* strains have 82% identity, while the respective edited molecules have 91% identity (85 and 56 nucleotide differences, respectively). Indeed, this effect is even more pronounced in the case of the pan-edited *atp6* transcript, with just 79% identity of pre-edited mRNAs between the strains, but with 94% identity after post-transcriptional modification (Supporting Fig. 6). These results are consistent with the notion that while protein sequences are maintained by selective forces, the sequence of a cryptic gene is able to evolve more freely, with mutations 'corrected' by RNA editing (Landweber and Gilbert 1993).

Alternative editing and 'misediting'

We observed a certain fraction of alternatively edited reads for each of the 10 edited transcript domains in the *Perkinsela* mitochondrial genome. We define 'alternative' reads as those containing at least one alternatively edited site satisfying the following conditions: (i) it was never edited in the main editing product; (ii) the U-indel was longer than in the main product; and (iii) insertion occurred instead of deletion in the main product or *vice versa*. Redundant alternatively edited reads were grouped into clusters of non-redundant (longest) editing intermediates. The fraction of alternatively edited reads (relative to all edited reads), as inferred using T-aligner, was found to vary from 19% in the investigated edited domains of *cox2, cox3, cob* to 52% in the pan-edited *atp6* transcript (Fig. 6). Absolute numbers of alternative reads in our dataset varied from 44 for *rps12* to 1,979 for *atp6*, depending on the level of coverage for a particular strain, transcript abundance and T-aligner seed selection (Fig. 6).

Importantly, we observed no cases of a clearly predominant single alternative editing

intermediate, and an overwhelming majority of alternative intermediates was represented by single reads, as illustrated in Supporting Fig. 7. We used the following edited domains as model cases: (i) *cox1*, 5′ domain, strain CCAP1560/4; (ii) *cox2*, 3′ domain, strain GillNOR1/I; (iii) *cox3*, 5′ domain, strain CCAP1560/4; (iv) *cob*, 3′ domain, strain GillNOR1/I; (v) 3′ part of the pan-edited *atp6* transcript, strain CCAP1560/4; and (vi) the pan-edited *rps12* transcript, strain GillNOR1/I (Supporting Fig. 8). The *cox2* 3′ domain (Fig. 2) was the most insightful due to its high coverage with long strand-specific reads (average length 192 nt; maximum length ~400 nt; 4,711 edited reads in total) (Supporting Fig. 8). A maximum number of 14 alternatively edited sites was observed in the reads available for the 3′ domain of *cox2*. However, just 10 out of 880 reads mapping to this domain contained 10 or more alternatively edited sites (Supporting Fig. 9). An even larger pool of 1,979 alternatively edited reads mapping to the 3′ part of *atp6* contained just 6 reads with 10 or more alternatively edited sites. However, shorter reads were available in this case (Supporting Fig. 8). Taken together, these numbers, the read length distribution (Supporting Fig. 3), and the read counts for alternative intermediates (Supporting Fig. 7) strongly indicate that alternative final transcripts, comparable with the main transcript in length and abundance, do not occur in this system. A typical selection of alternative intermediates is shown for the 3′ edited domain of *cox2* (Fig. 7).

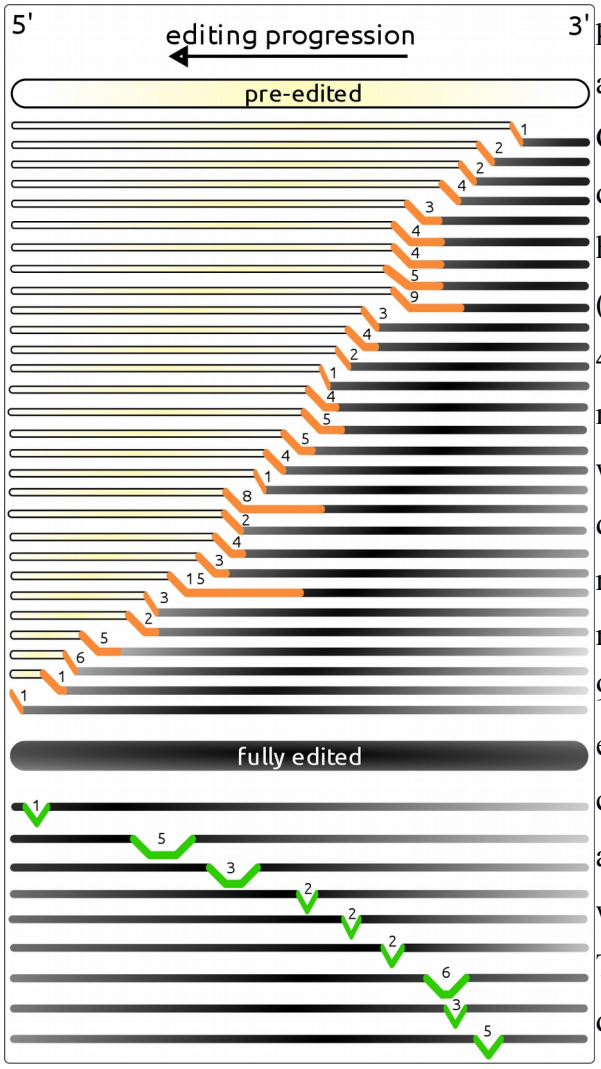


Figure 7. The most abundant alternatively edited intermediates mapped to the cox2 transcript (3′ edited domain) in Perkinsela strain GillNOR1/I. Pre-edited sequence is shown in yellow and the main edited product in black. Alternatively edited read fragments are shown in orange if they follow the sequence edited in a standard way or in green if they occur in the middle of such a sequence. The number of alternatively edited sites is shown in each case, and the length of highlighted regions correlates with the number of inserted Us.

The majority of editing intermediates contain one or several alternative sites at the end of an edited stretch of sequence, followed by a pre-edited sequence. Considering that in *T. brucei* approximately 45 nucleotides (from 24 to 61**)** are covered by an average gRNA (Koslowsky et al. 2013), the terminal stretches observed in *Perkinsela* are probably generated by one or two consecutively acting gRNAs. The paucity of longer terminal stretches (Supporting Fig. 9) suggests that we are mainly observing instances of abortive 'mis-editing' (Sturm et al. 1992). As is apparent even from a small selection of alternative intermediates (Fig. 7), editing errors occur almost everywhere along the transcript. However, a wider selection of intermediates (Supporting Fig. 8) reveals a few hotspots.

Another type of editing intermediate contains one to six alternative sites within a sequence corresponding to the main editing pathway. These 'internal' intermediates are apparently produced by a single gRNA guiding several editing sites in a non-canonical way, but still generating an anchor sequence for a subsequent gRNA in the main editing pathway. Remarkably, both types of alternative editing have been predicted in *T. brucei* by deep sequencing a gRNA library with a total of ~600 major sequence classes (Koslowsky et al. 2013): gRNAs were identified that create an alternative sequence not usable as an anchor, as were gRNAs that edit several sites in an alternative way, but create an anchor region for the next gRNA in the main editing pathway. For instance, an alternative gRNA might initiate editing at the 3′ end of *atp6* (also known as A6) in *T. brucei*, but is also able to create a normal anchor for the next gRNA. The same is true for alternative gRNA editing of the ND8 transcript (*nad8*). In *Perkinsela*, we also observed intermediates containing more than one internal alternatively edited stretch, or intermediates with a combination of terminal and internal alternatively edited stretches (Supporting Fig. 8), all of which are of low abundance.

Based on our data, the RNA editing pathway in *Perkinsela* and probably all kinetoplastids can be viewed as a 'tree' with numerous branching points, with only one path in the tree being predominant and the rest probably representing errors of the editing system. In *T. brucei*, alternative gRNAs were identified for at least five genes, with some being even more abundant than the standard gRNAs for the same site (Koslowsky et al. 2013). Given a high percentage of alternative reads accumulated for some edited domains in *Perkinsela* (e.g., 52% for *atp6*), we speculate that the mitochondrial transcription-translation system in this organism

can tolerate a large number of 'incorrect' transcripts. Moreover, all alternative reads that map to the *rps12* gene lack stop codons in at least one frame. In plant organelles, only edited translation products appear to accumulate in mitochondrial ribosomes (Phreaner et al. 1996). Whether or not the *Perkinsela* mitochondrion is able to tolerate 'incorrect' protein products, or some sort of discrimination by the translation machinery is in place, remains an open question.

Acknowledgements

We thank Ivan Fiala, Hanka Pecková, and Eva Horáková for technical assistance, and Bruce Curtis for bioinformatics support. This work was supported by Czech Grant Agency P305/12/2261, AMVIS LH 12104 grant, RNPnet FP7 program 289007, and Praemium Academiae award to J.L.; and an operating grant awarded to J.M.A. from the Canadian Institutes of Health Research. J.M.A. also acknowledges support from the Canadian Institute of Advanced Research and Dalhousie University's Centre for Comparative Genomics and Evolutionary Bioinformatics. We acknowledge the use of research infrastructure funded from the EU 7th Framework Programme n˚316304. P.F. was supported by the Moravian-Silesian region projects MSK2013-DT1, MSK2013-DT2, and MSK2014-DT1; P.F., E.G. obtained support from the Russian Foundation for Basic Research, project 14-04-31936; E.G. from the Russian Foundation for Basic Research, project 14-04-01717; and M.D.L. was supported by project 14-50-00029 of the Russian Science Foundation.

References

Adler, B. K., Harris,M. E., Bertrand, K I. and Hajduk, S. L. (1991) Modification of *Trypanosoma brucei* mitochondrial rRNA by posttranscriptional 3′ polyuridine tail formation. *Mol. Cell. Biol.*, **11**, 5878-5884.

Ammerman, M. L., Downey, K. M., Hashimi, H., Fisk, J. C., Tomasello, D. L., Faktorová, D., Kafková, L., King, T., Lukeš, J. and Read, L. K. (2012) Architecture of the trypanosome RNA editing accessory complex, MRB1. *Nucleic Acids Res.*, **40**, 5637-5650.

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

Aphasizhev, R. and Aphasizheva, I. (2014) Mitochondrial RNA editing in trypanosomes: small RNAs in control. *Biochimie*, **100**, 125-131.

Aphasizheva, I., Maslov, D., Wang, X., Huang, L. and Aphasizhev, R. (2011) Pentatricopeptide repeat proteins stimulate mRNA adenylation/uridylation to activate mitochondrial translation in trypanosomes. *Mol. Cell*, **42**, 106-117.

Aphasizheva, I., Maslov, D. A. and Aphasizhev, R. (2013) Kinetoplast DNA-encoded ribosomal protein S12: a possible functional link between mitochondrial RNA editing and translation in *Trypanosoma brucei*. *RNA Biol.*, **10**, 1679-1688.

Arts,G.J., van der Spek,H., Speijer,D., van den Burg,J., van Steeg,H., Sloof,P. and Benne R. (1993) Implications of novel guide RNA features for the mechanism of RNA editing in *Crithidia fasciculata*. *EMBO J.*, **12**, 1523-1532.

Avesson, L. and Barry, G. (2014) The emerging role of RNA and DNA editing in cancer. *Biochim. Biophys. Acta*, **1845**, 308-316.

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

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

Carnes, J., Trotter, J. R., Peltan, A., Fleck, M. and Stuart. K. (2008) RNA editing in *Trypanosoma brucei* requires three different editosomes. *Mol. Cell. Biol.*, **28**, 122-130.

Carnes, J., Soares, C. Z., Wickham, C. and Stuart, K. (2011) Endonuclease associations with three distinct editosomes in *Trypanosoma brucei*. *J. Biol. Chem.*, **286**, 19320-19330.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahé, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., *et al*. (2015) Eukaryotic plankton diversity in the sunlit global ocean. *Science* (accepted)

Dyková, I., Figueras, A. and Peric, Z. (2000) *Neoparamoeba* Page, 1987: light and electron microscopic observations on six strains of different origin. *Dis. aquat. Org.*, **43**, 217-223.

Dyková, I., Nowak, B. F., Crosbie, P.B., Fiala, I., Pecková, H., Adams, M. B., Macháčková, B. and Dvořáková, H. (2005) *Neoparamoeba branchiphila* n. sp., and related species of the genus *Neoparamoeba* Page, 1987: morphological and molecular characterization of selected strains. *J. Fish Dis.*, **28**, 49-64.

Dyková, I., Fiala, I., Pecková H., (2008) *Neoparamoeba* spp. and their eukaryotic endosymbionts similar to *Perkinsela amoebae* (Hollande, 1980): coevolution demonstrated by SSU rRNA gene phylogenies. *Eur. J. Protistol.*, **44**, 269-277.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792-1797.

Ernst, N.L., Panicucci, B., Igo, R.P. Jr, Panigrahi, A.K., Salavati, R. and Stuart, K. (2003) TbMP57 is a 3′ terminal uridylyl transferase (TUTase) of the *Trypanosoma brucei* editosome. *Mol. Cell*, **11**, 1525-1536.

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

Farrar, M. (2007) Striped Smith-Waterman speeds database searches six times over other SIMD implementations. *Bioinformatics*, **23**, 156-161.

Feehan, C. J., Johnson-Mackinnon, J., Scheibling, R. E., Lauzon-Guay, J. S. and Simpson, A. G. (2013) Validating the identity of *Paramoeba invadens*, the causative agent of recurrent mass mortality of sea urchins in Nova Scotia, Canada. *Dis. aquat. Org.*, **103**, 209-227.

Fisk, J. C., Presnyak, V., Ammerman, M. L. and Read, L. K. (2009) Distinct and overlapping functions of MRP1/2 and RBP16 in mitochondrial RNA metabolism. *Mol. Cell. Biol.*, **29**, 5214-5225.

Golden, D. E. and Hajduk, S. L. (2005) The 3′-untranslated region of cytochrome oxidase II mRNA functions in RNA editing of African trypanosomes exclusively as a *cis* guide RNA. *RNA*, **11**, 29-37.

Gott, J.M. and Emeson, R. B. (2000) Functions and mechanisms of RNA editing. *Annu. Rev. Genet.*, **34**, 499-531.

Grams, J., Morris, J.C., Drew, M.E., Wang, Z.F., Englund, P.T. and Hajduk, S.L. (2002) A trypanosome mitochondrial RNA polymerase is required for transcription and replication. *J. Biol. Chem.*, **277**, 16952-16959.

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

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., *et al*. (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat. Prot.*, **8**, 1494-1512.

Hashimi, H., Zíková, A., Panigrahi, A.K., Stuart, K.D. and Lukeš, J. (2008) TbRGG1, an essential protein involved

in kinetoplastid RNA metabolism that is associated with a novel multiprotein complex. *RNA*, **14**, 970-980.

Hashimi, H., Zimmer, S.L., Ammerman, M.L., Read, L.K. and Lukeš, J. (2013) Dual core processing: MRB1 is an emerging kinetoplast RNA editing complex. *Trends Parasitol.*, **29**, 91-99.

Hernandez, A., Madina, B.R., Ro, K., Wohlschlegel, J.A., Willard, B., Kinter, M.T. and Cruz-Reyes, J. (2010) REH2 RNA helicase in kinetoplastid mitochondria: ribonucleoprotein complexes and essential motifs for unwinding and guide RNA (gRNA) binding. *J. Biol. Chem.*, **285**, 1220-1228.

Kafková, L., Ammerman, M.L., Faktorová, D., Fisk, J.C., Zimmer, S.L., Sobotka, R., Read, L.K., Lukeš, J. and Hashimi, H. (2012) Functional characterization of two paralogs that are novel RNA binding proteins influencing mitochondrial transcripts of *Trypanosoma brucei*. *RNA*, **18**, 1846-1861.

Koslowsky, D., Sun, Y., Hindenach, J., Theisen, T. and Lucas, J. (2013) The insect-phase gRNA transcriptome in *Trypanosoma brucei*. *Nucleic Acids Res.*, **42**, 1873-1886.

Kudryavtsev, A., Pawlowski, J. and Hausmann, K. (2011) Description of *Paramoeba atlantica* n. sp. (Amoebozoa, Dactylopodida) - a marine amoeba from the eastern Atlantic, with emendation of the dactylopodid families. *Acta Protozool.*, **50**, 239-253.

Landweber, L.F. and Gilbert, W. (1993) RNA editing as a source of genetic variation. *Nature*, **363**, 179-182.

Landweber, L.F., Fiks, A.G. and Gilbert, W. (1993) The boundaries of partially edited transcripts are not conserved in kinetoplastids: implications for the guide RNA model of editing. *Proc. Natl. Acad. Sci. USA*, **90**, 9242-9246.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357-359.

Li, J.B. and Church, G.M. (2013) Deciphering the functions and regulation of brain-enriched A-to-I RNA editing. *Nat. Neurosci.*, **16**, 1518-1522.

Li, F., Herrera, J., Zhou, S., Maslov, D.A, and Simpson L. (2011) Trypanosome REH1 is an RNA helicase involved with the 3′-5′ polarity of multiple gRNA-guided uridine insertion/deletion RNA editing. *Proc Natl Acad Sci USA.*, **108**, 3542-3547.

Lukeš, J., Arts, G.J., van den Burg, J., de Haan, A., Opperdoes, F., Sloof, P. and Benne R. (1994) Novel pattern of editing regions in mitochondrial transcripts of the cryptobid *Trypanoplasma borreli*. *EMBO J.*, **13**, 5086-5098.

Lukeš, J., Archibald, J.M., Keeling, P.J., Doolittle, W.F. and Gray, M.W. (2011) How a neutral evolutionary ratchet can build cellular complexity. *IUBMB Life*, **63**, 528-537.

Lukeš, J., Skalický, T., Týč, J., Votýpka, J. and Yurchenko, V. (2014) Evolution of parasitism in kinetoplastid flagellates. *Mol. Biochem. Parasitol.*, **195**, 115-122.

Maas, S. (2012) Posttranscriptional recoding by RNA editing. *Adv. Protein Chem. Struct. Biol.*, **86**, 193-224.

Madina, B.R., Kuppan, G., Vashisht, A.A., Liang, Y.H., Downey, K.M., Wohlschlegel, J.A., Ji, X., Sze, S.H., Sacchettini, J.C., Read, L.K. and Cruz-Reyes, J. (2011) Guide RNA biogenesis involves a novel RNase III family endoribonuclease in *Trypanosoma brucei*. *RNA*, **17**, 1821-1830.

Madina, B.R., Kumar, V., Metz, R., Mooers, B.H., Bundschuh, R. and Cruz-Reyes, J. (2014) Native mitochondrial RNA-binding complexes in kinetoplastid RNA editing differ in guide RNA composition *RNA*, **20**, 1142-1152.

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

Maslov, D.A. and Simpson, L. (1992) The polarity of editing within a multiple gRNA-mediated domain is due to formation of anchors for upstream gRNAs by downstream editing. *Cell*, **70**, 459-467.

Maslov, D.A., Thiemann, O. and Simpson, L. (1994) Editing and misediting of transcripts of the kinetoplast maxicircle G5 (ND3) cryptogene in an old laboratory strain of *Leishmania tarentolae*. *Mol. Biochem. Parasitol.*, **68**, 155-159.

Maslov, D.A., Spremulli, L.L., Sharma, M.R., Bhargava, K., Grasso, D., Falick, A.M., Agrawal, R.K., Parker, C.E. and Simpson, L. (2007) Proteomics and electron microscopic characterization of the unusual mitochondrial ribosome-related 45S complex in *Leishmania tarentolae*. *Mol. Biochem. Parasitol.*, **152**, 203-212.

Maslov, D.A., Votýpka, J., Yurchenko, V. and Lukeš, J. (2013) Diversity and phylogeny of insect trypanosomatids: all that is hidden shall be revealed. *Trends Parasitol.*, **29**, 43-52.

McAdams, N.M., Ammerman, M.L., Nanduri, J., Lott, K., Fisk, J.C. and Read, L.K. (2015) An arginine-glycine-rich RNA binding protein impacts the abundance of specific mRNAs in the mitochondria of *Trypanosoma brucei*. *Eukaryot. Cell*, **14**, 149-157.

Mitchell, S.O., Rodger, H.D. (2011) A review of infectious gill disease in marine salmonid fish. *J. Fish Dis.*, **34**, 411-432.

Moreira, D., López-García, P. and Vickerman K. (2004) An updated view of kinetoplastid phylogeny using environmental sequences and a closer outgroup: proposal for a new classification of the class Kinetoplastea. *Int. J. Syst. Evol. Microbiol.*, **54**, 1861-1875.

Ochsenreiter, T., Hajduk, S.L. (2006) Alternative editing of cytochrome c oxidase III mRNA in trypanosome mitochondria generates protein diversity. *EMBO Rep.*, **7**, 1128-1133.

Page, F.C. (1973) *Paramoeba* - Common marine genus. *Hydrobiologia*, **4**, 183-188.

Panigrahi, A.K., Zíková, A., Halley, r.A., Acestor, N., Ogata, Y., Myler, P.J. and Stuart, K. (2008) Mitochondrial complexes in *Trypanosoma brucei*: a novel complex and a unique oxidoreductase complex. *Mol. Cell. Proteomics*, **7**, 534-545.

Ridlon, L., Škodová, I., Pan, S., Lukeš, J. and Maslov, D.A. (2013) The importance of the 45 S ribosomal small subunit-related complex for mitochondrial translation in *Trypanosoma brucei*. *J. Biol. Chem.*, **288**, 32963-32978.

Schumacher, M.A., Karamooz, E., Zíková, A., Trantírek, L. and Lukeš, J. (2006) Crystal structures of *T. brucei* MRP1/MRP2 guide-RNA binding complex reveal RNA matchmaking mechanism. *Cell*, **126**, 701-711.

Sharma, M.R., Booth, T.M., Simpson, L., Maslov, D.A. and Agrawal, R.K. (2009) Structure of a mitochondrial ribosome with minimal RNA. *Proc. Natl. Acad. Sci. USA*, **106**, 9637-9642.

Sloof, P., Van den Burg, J., Voogd, A., Benne, R., Agostinelli, M., Borst, P., Gutell, R. and Noller, H. (1985) 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.*, **13**, 4171-4190.

Sturm, N.R., Maslov, D.A., Blum, B. and Simpson, L. (1992) Generation of unexpected editing patterns in *Leishmania tarentolae* mitochondrial mRNAs: misediting produced by misguiding. *Cell*, **70**, 469-476.

Surve, S., Heestand, M., Panicucci, B., Schnaufer, A. and Parsons, M. (2012) Enigmatic presence of mitochondrial complex I in *Trypanosoma brucei* bloodstream forms. *Eukaryot. Cell*, **11**, 183-193.

Takenaka, M., Zehrmann, A., Verbitskiy, D., Härtel, B. and Brennicke A. (2013) RNA editing in plants and its evolution. *Annu. Rev. Genet.*, **47**, 335-352.

Tanifuji, G., Kim, E., Onodera, N.T., Gibeault, R., Dlutek, M., Cawthorn, R.J., Fiala, I., Lukeš, J., Greenwood, S.J. and Archibald, J.M. (2011) Genomic characterization of *Neoparamoeba pemaquidensis* (Amoebozoa) and its kinetoplastid endosymbiont. *Eukaryot. Cell*, **10**, 1143-1146.

Valach, M., Moreira, S., Kiethega, G.N. and Burger, G. (2013) Trans-splicing and RNA editing of LSU rRNA in *Diplonema* mitochondria. *Nucleic Acids Res.*, **42**, 2660-2672.

Verner, Z., Čermáková, P., Škodová, I., Kriegová, E., Horváth, A. and Lukeš, J. (2011) Complex I (NADH:ubiquinone oxidoreductase) is active in but non-essential for procyclic *Trypanosoma brucei*. *Mol. Biochem. Parasitol.*, **175**, 196-200.

Verner, Z., Basu, S., Benz, C., Dixit, S., Dobáková, E., Faktorová, D., Hashimi, H., Horáková, E., Huang, Z., Paris, Z., Pena-Diaz, P., Ridlon, L., Týč, J., Wildridge, D., Zíková, A. and Lukeš,J. (2015) The malleable mitochondrion of *Trypanosoma brucei*. *Int. Rev. Cell. Mol. Biol.* **315**, 73-152.

Votýpka, J., Kostygov, A.Y., Kraeva, N., Grybchuk-Ieremenko, A., Tesařová, M., Grybchuk, D., Lukeš, J. and Yurchenko, V. (2014) *Kentomonas* gen. n., a new genus of endosymbiont-containing trypanosomatids of Strigomonadinae subfam. n. *Protist*, **165**, 825-838.

Weng, J., Aphasizheva, I., Etheridge, R.D., Huang, L., Wang, X., Falick, A.M. and Aphasizhev, R. (2008) Guide RNA-binding complex from mitochondria of trypanosomatids. *Mol. Cell*, **32**, 198-209.

Young, N.D., Dyková, I., Nowak, B.F. and Morrison, R.N. (2008) Development of a diagnostic PCR to detect *Neoparamoeba perurans*, agent of amoebic gill disease. *J. Fish Dis.*, **29**, 1-11.

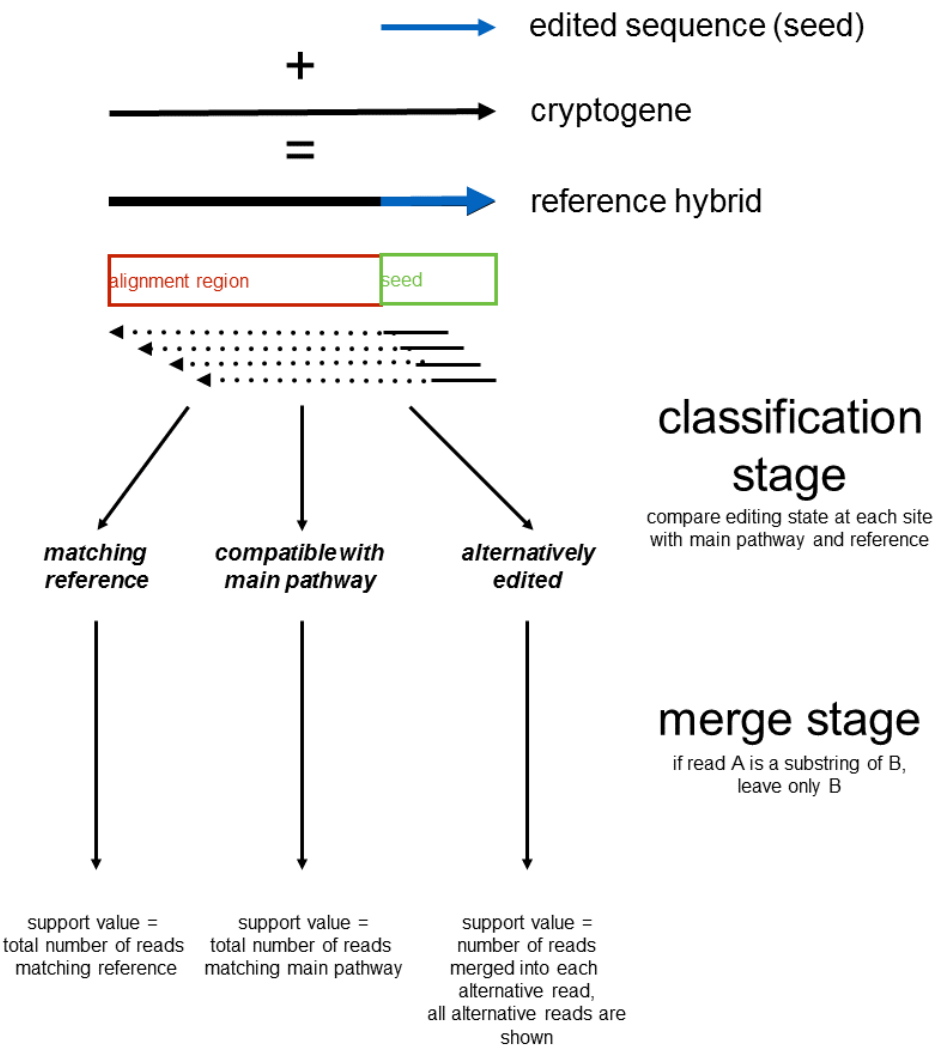
Young, N.D., Dyková, I., Crosbie, P.B., Wolf, M., Morrison, R.N., Bridle, A.R. and Nowak, B.F. (2014) Support for the coevolution of *Neoparamoeba* and their endosymbionts, *Perkinsela amoebae*-like organisms. *Eur. J. Protistol.*, **50**, 509-523.

Zíková, A., Panigrahi, A.K., Dalley, R.A., Acestor, N., Anupama, A., Ogata, Y., Myler ,P.J. and Stuart,K. (2008a) *Trypanosoma brucei* mitochondrial ribosomes: affinity purification and component identification by mass spectrometry. *Mol. Cell. Proteomics*, **7**, 1286-1296.

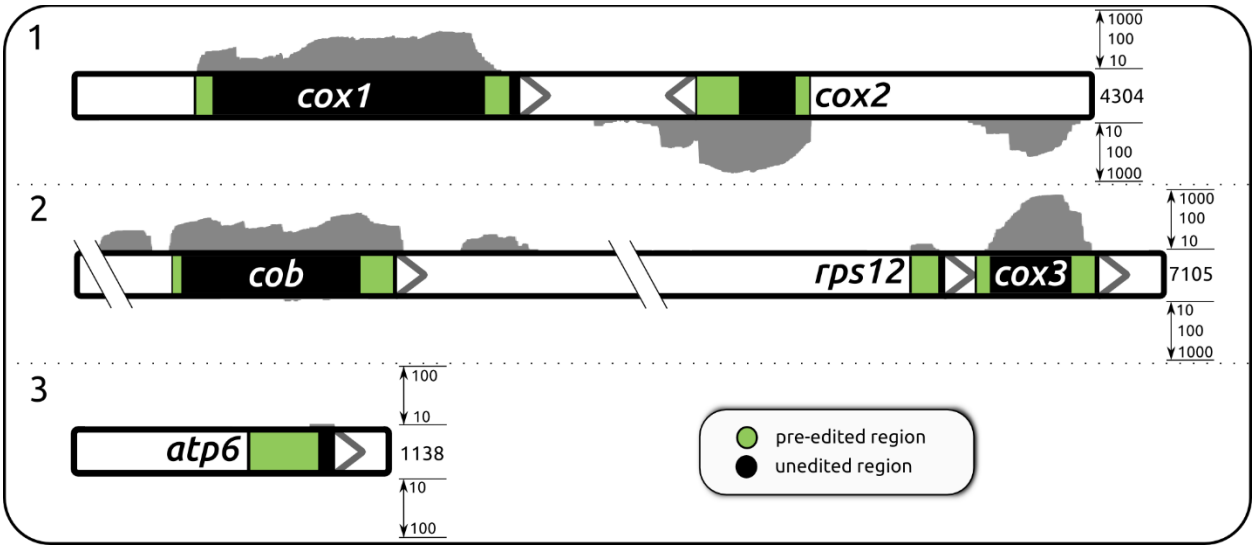
Zíková, A., Kopečná, J., Schumacher, M.A., Stuart, K., Trantírek, L. and Lukeš, J. (2008b) Structure and function of the native and recombinant mitochondrial MRP1/MRP2 complex from *Trypanosoma brucei*. *Int. J. Parasitol.*, **38**, 901-912.

Supplementary data

Supporting Fig. 1. Workflow of T-aligner.

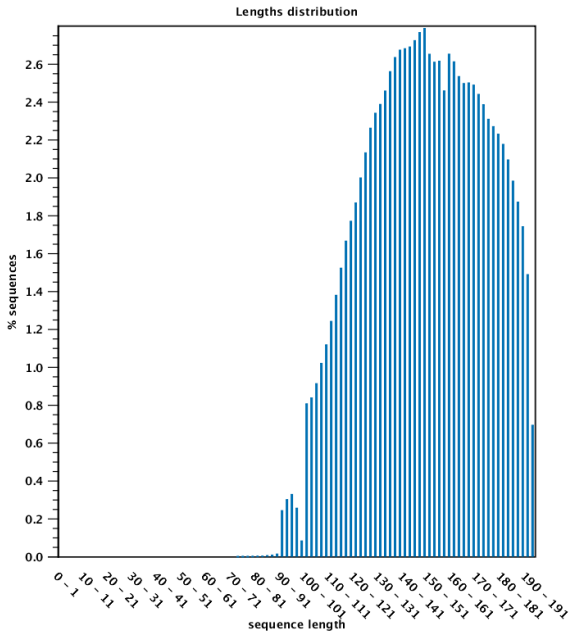


Supporting Fig. 2. *Perkinsela* strain GillNOR1/I mitochondrial scaffolds with both sense and anti-sense transcriptomic reads mapped. Almost no antisense transcription is visible, which is supported by the Northern blot in Fig. 3. The sense transcription profile, showing very low coverage for pan-edited genes *rps12* and *atp6*, is different from that shown in Fig. 2 since 'U-indel optimized' settings in Bowtie were not used here. 'U-indel optimized' settings may produce strand biases, e.g., favoring U-indels on the forward strand, but not A-indels on the reverse strand. Therefore they were not used for the purpose of inter-strand comparison of transcription profiles. However, regular Bowtie 'very sensitive' settings produce especially poor coverage in the case of pan-edited transcripts.

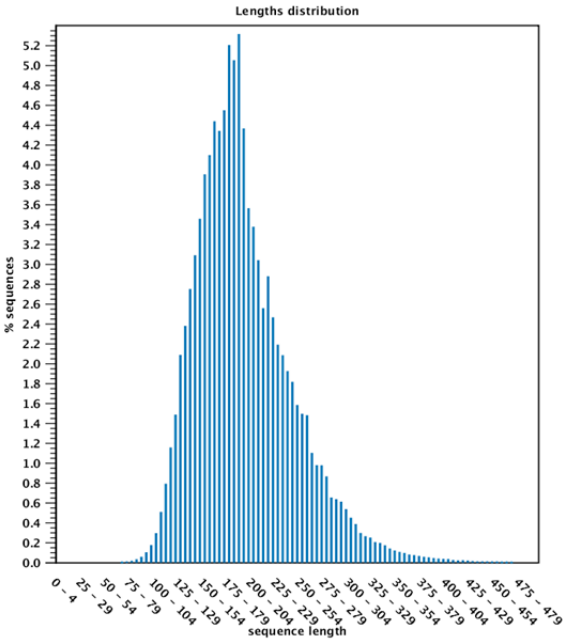


Supporting Fig. 3. Length distribution of transcriptomic reads of the *Perkinsela* CCAP1560/4 (**A**) and GillNOR1/I strains (**B**). Paired reads were merged with the CLC Genomics Workbench v.6.5 prior to mapping, which explains abrupt edges of the distribution in panel A (100 bp and shorter trimmed reads produce merged reads of 190 bp or shorter, if a minimum overlap of 10 bp is required).

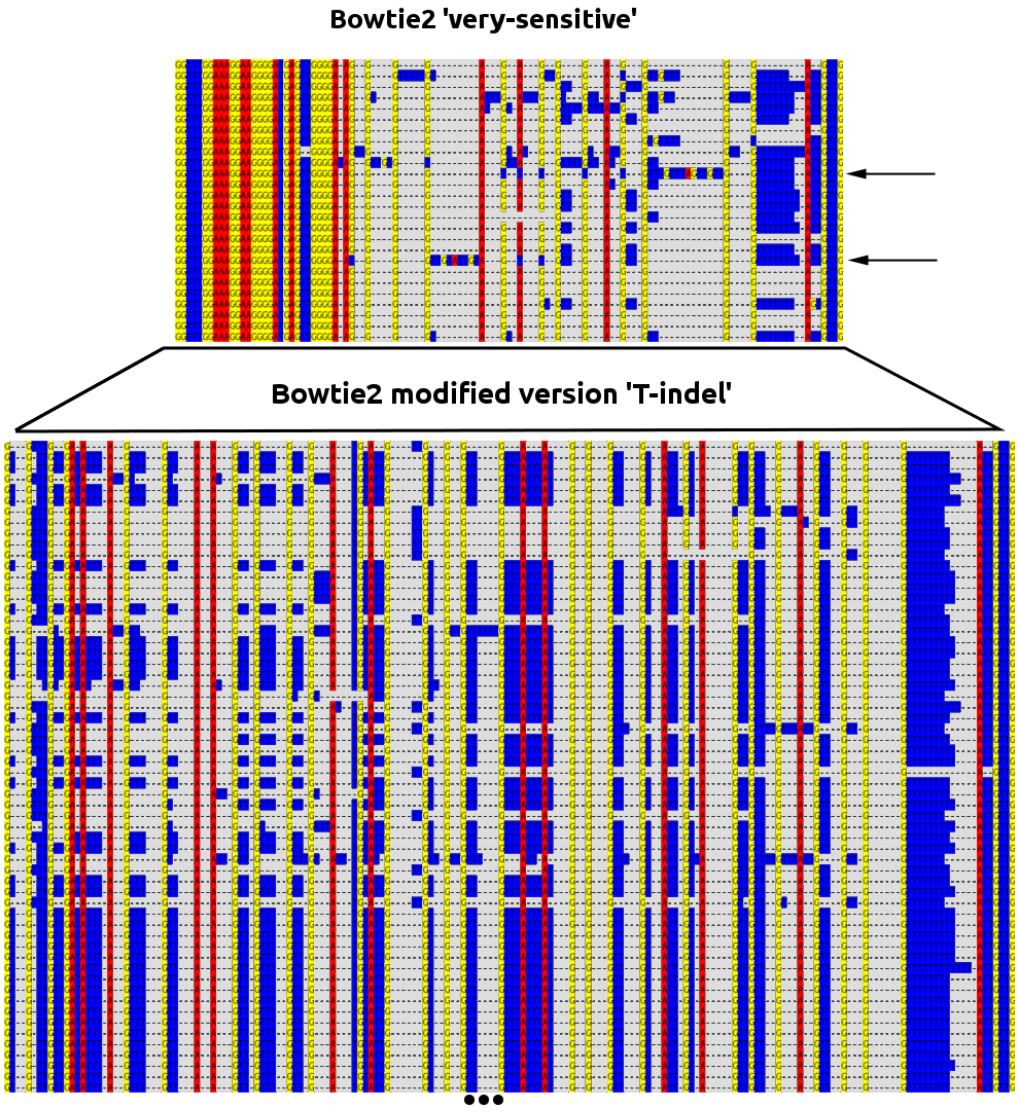
A



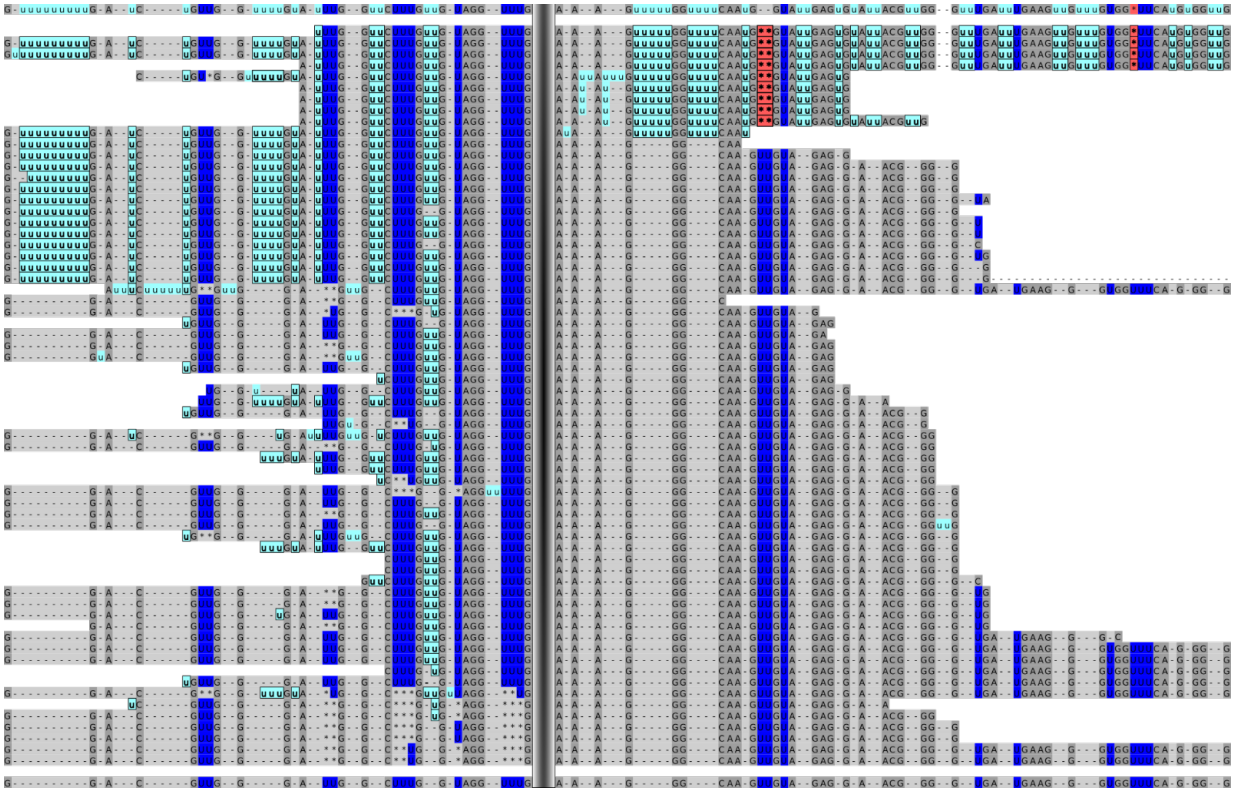
B



Supporting Fig. 4. Mapping of edited reads with Bowtie2 v.2.0.2 and its modified version. An alignment window shown here covers the 3′ editing region of cox1 in *Perkinsela* strain GillNOR1/I. Just a few reads are mapped by the standard Bowtie2 algorithm using the 'very-sensitive' setting. In contrast, modified Bowtie2 with T-indel-sensitive settings results in 12-fold increase of mapped read count (not all reads are shown in the figure). Moreover, misalignments such as those shown with arrows are missing because gaps containing ACG are penalized.

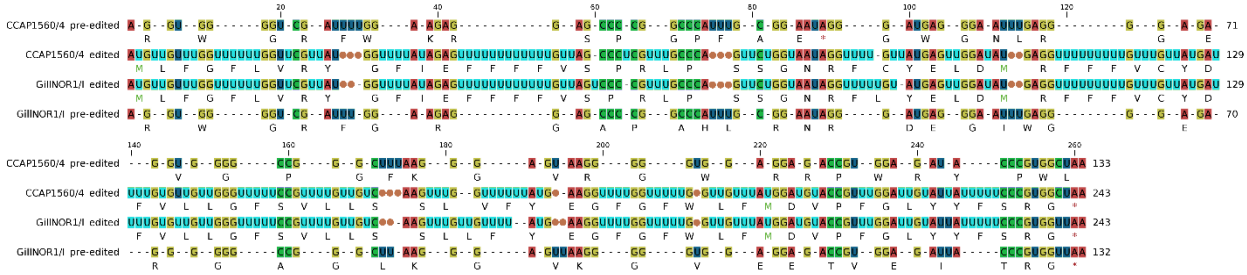


Supporting Fig. 5. All reads spanning both 5′ and 3′ edited domains of *cox2* in *Perkinsela* strain GillNOR1/I. Only parts of the edited domains adjoining the central non-edited region are shown, and the non-edited region itself is omitted (represented by the black bar in the center of the picture). The pre-edited sequence is shown at the bottom and the final edited sequence is on top. Insertions are shown in light blue, deletions in red and edits corresponding to the main edited product are boxed (alternative edits are not boxed). We found virtually no reads edited in the 3′ domain but not edited in the 5′ domain, but many examples of the opposite arrangement. Only a single read carries one alternative edit in the 3’ domain and no other edits.

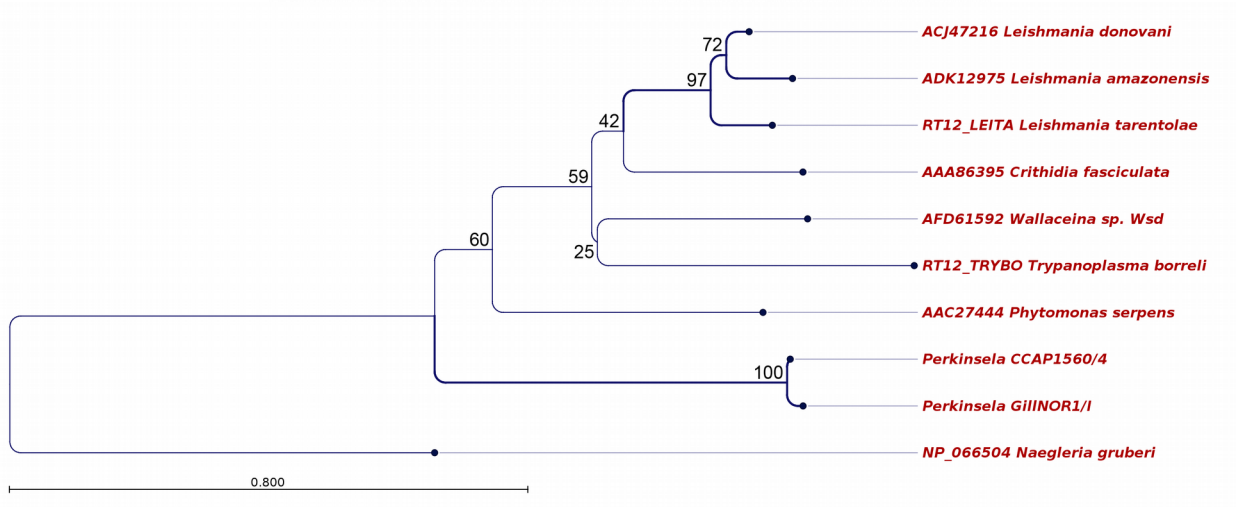


Supporting Fig. 6. Aligned edited/pre-edited transcripts and trees for the final protein sequences of 6 mitochondrial genes in *Perkinsela* strains CCAP1560/4 and GillNOR1/I. For each gene (cox1, cox2, cox3, cob, rps12, and atp6) the following information is shown: (i) edited and pre-edited transcript sequences with corresponding translations (U -insertions are denoted in blue and U deletions are marked with orange circles); (ii) pairwise percent identities (in the lower left part of the matrix) and numbers of different positions (in the upper right part) between edited/pre-edited sequences of both strains; (iii) a maximum likelihood unrooted tree for protein sequences of *Perkinsela*, other kinetoplastids and an outgroup (*Diplonema papillatum*, *Euglena gracilis*, or *Naegleria gruberi*, depending on sequence availability). The trees were constructed using the following settings: WAG+Γ substitution model, neighbor-joining starting tree, 1000 bootstrap replicates. Branches supported by bootstrap values >70% are shown with thicker lines. Scale bars show inferred number of amino acid substitutions per site.

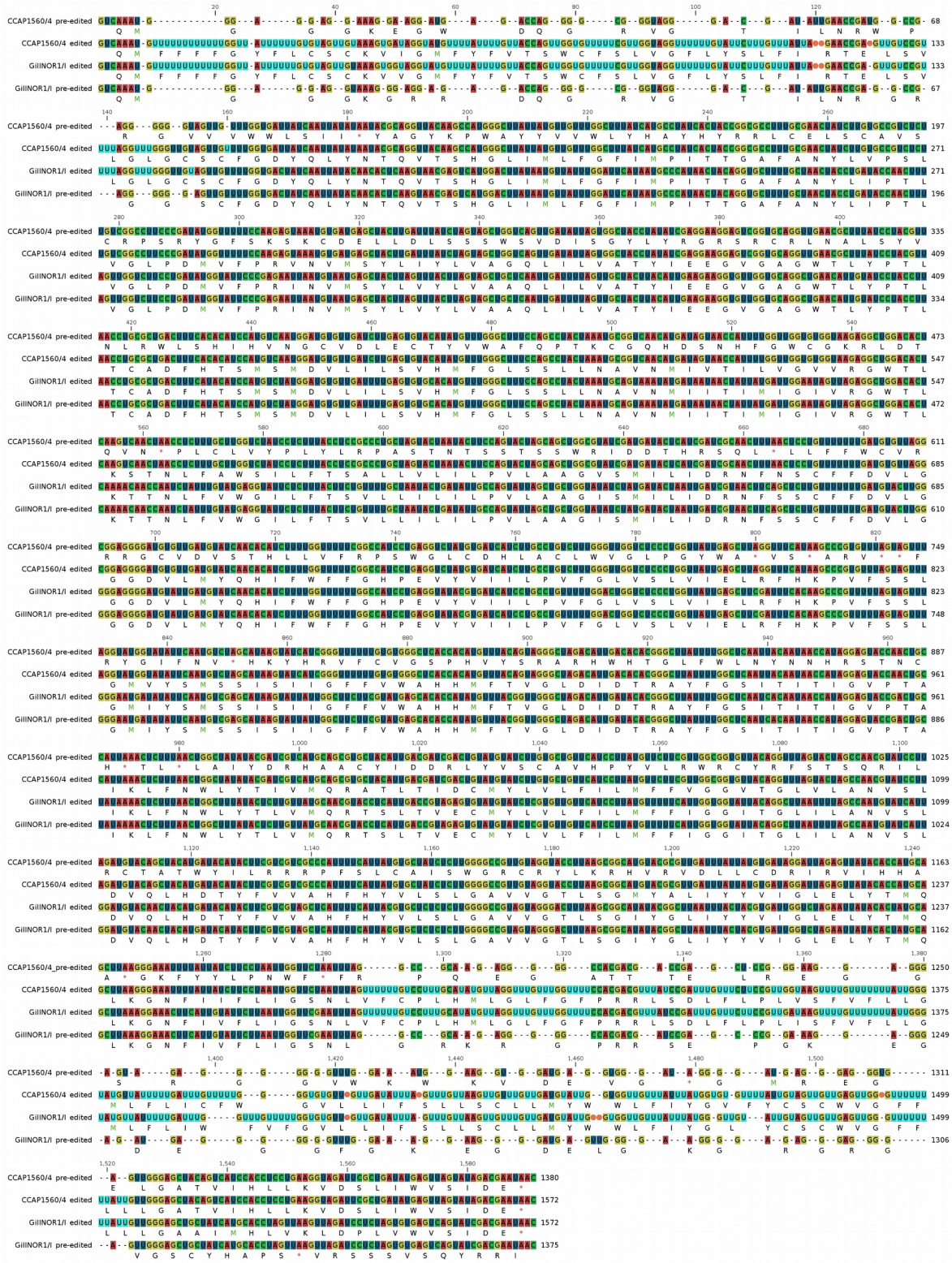
rps12



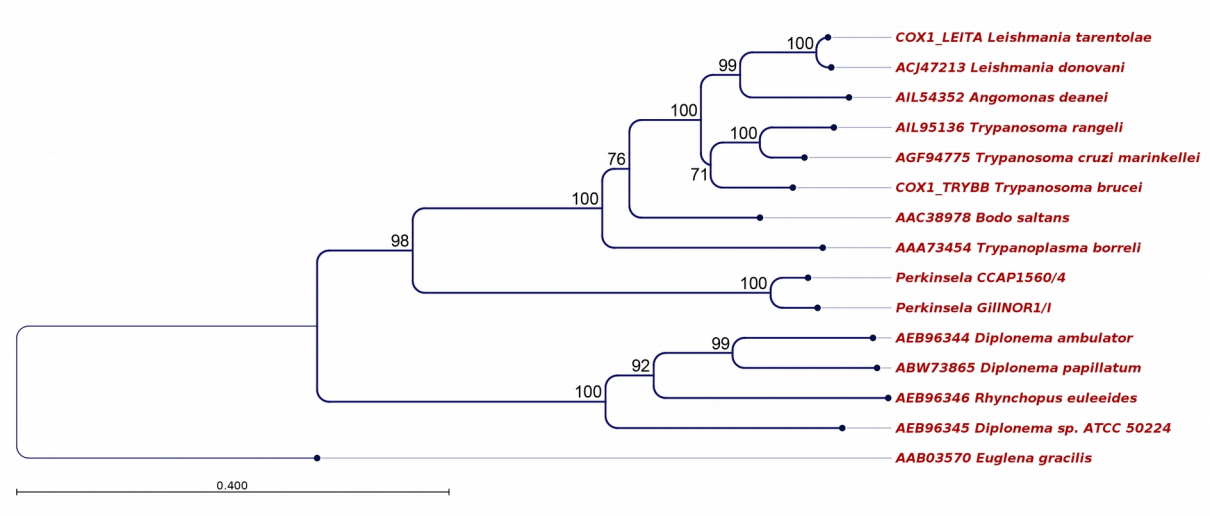
	1	2	3	4
CCAP1560/4 pre-edited	1	136	137	6
CCAP1560/4 edited	46.88	2	9	136
GillNOR1/I edited	46.48	96.36	3	135
GillNOR1/I pre-edited	95.56	46.67	47.06	4



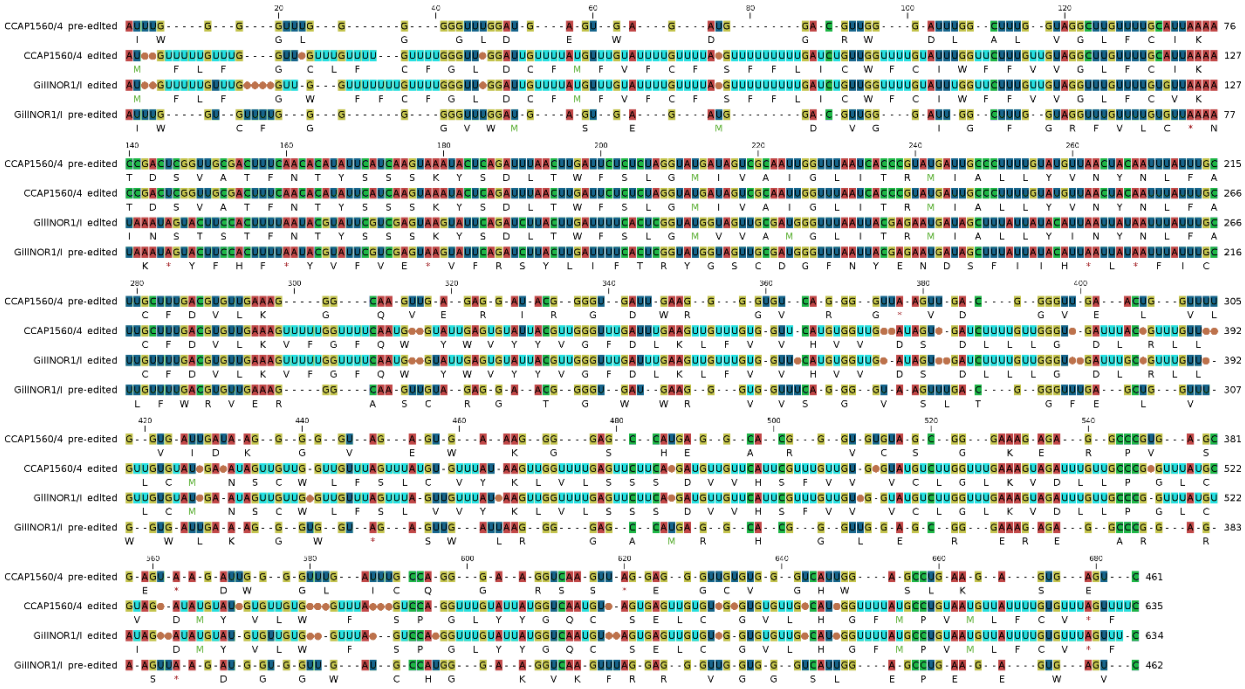
cox1


CCAP1560/4 pre-edited
CCAP1560/4 edited
GilINOR1/I edited
GilINOR1/I pre-edited
CCAP1560/4_pre-edited

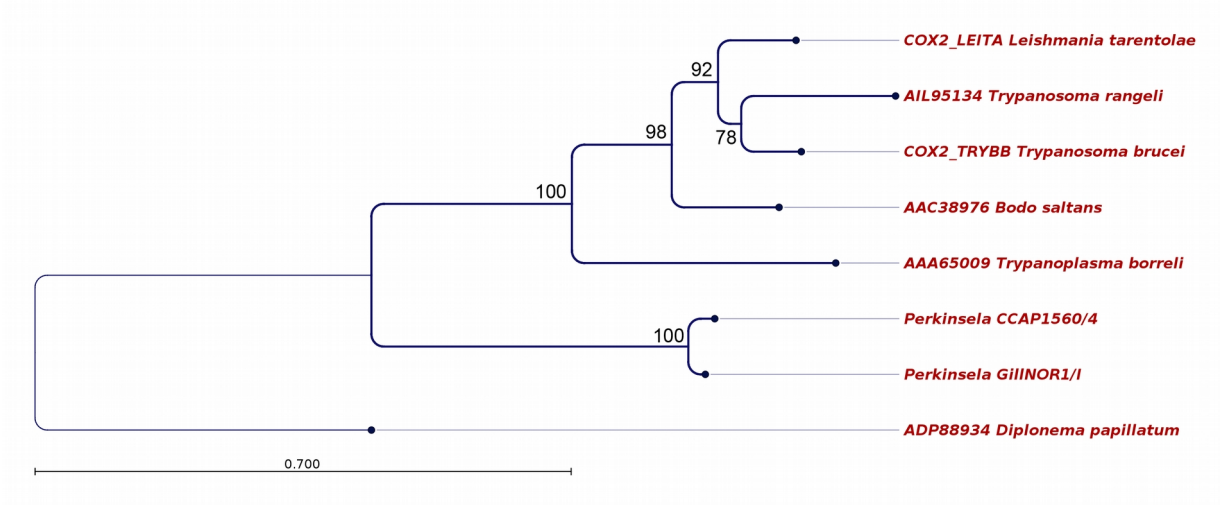
		1	2	3	4
CCAP1560/4 pre-edited	1		204	412	225
CCAP1560/4 edited	2	87.07		226	415
GillNOR1/I edited	3	73.89	85.71		207
GillNOR1/I pre-edited	4	83.77	73.68	86.87	



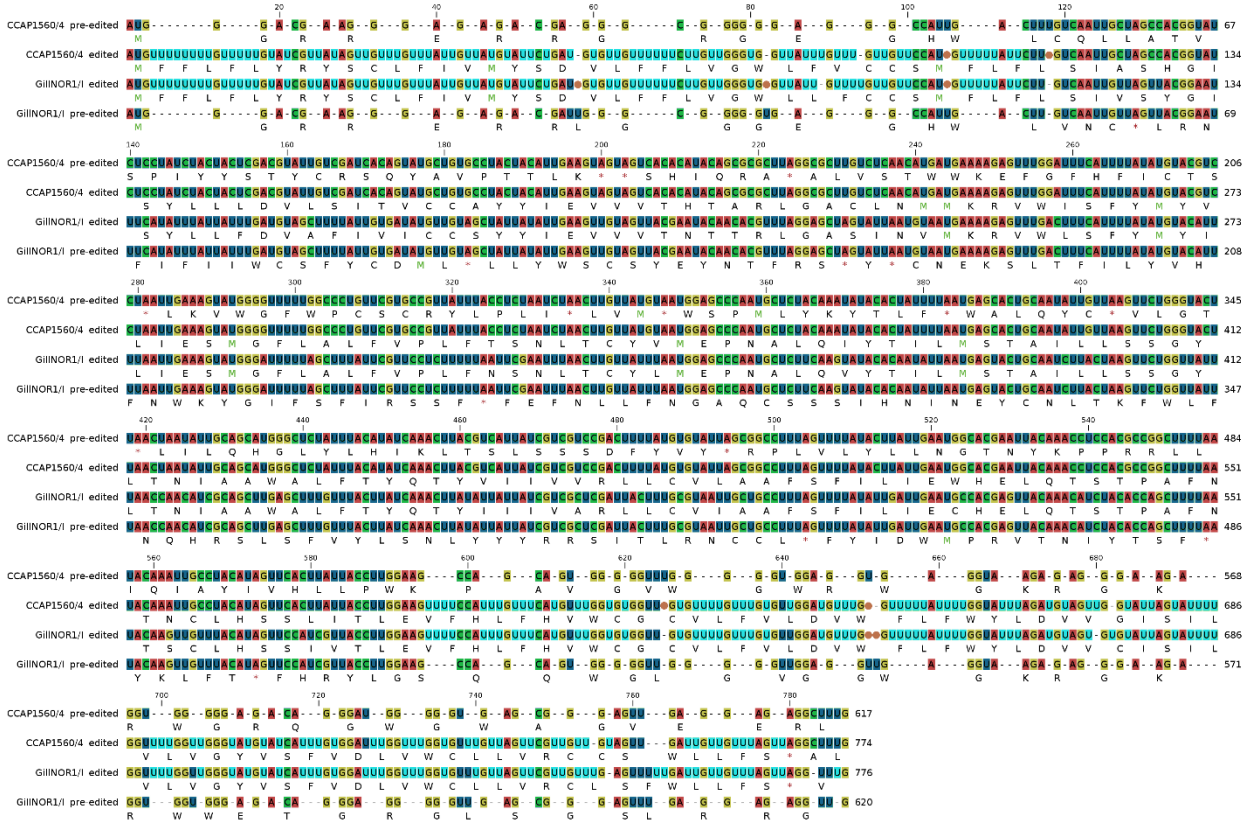
cox2



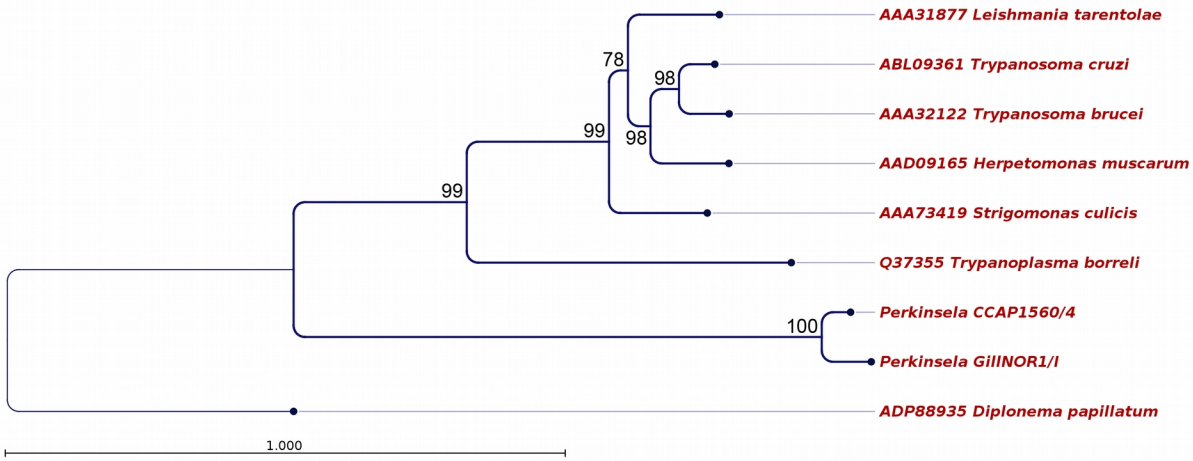
		1	2	3	4
CCAP1560/4 pre-edited	1		240	286	85
CCAP1560/4 edited	2	64.07		56	289
GillNOR1/I edited	3	57.12	91.24		240
GillNOR1/I pre-edited	4	82.33	56.87	64.07	



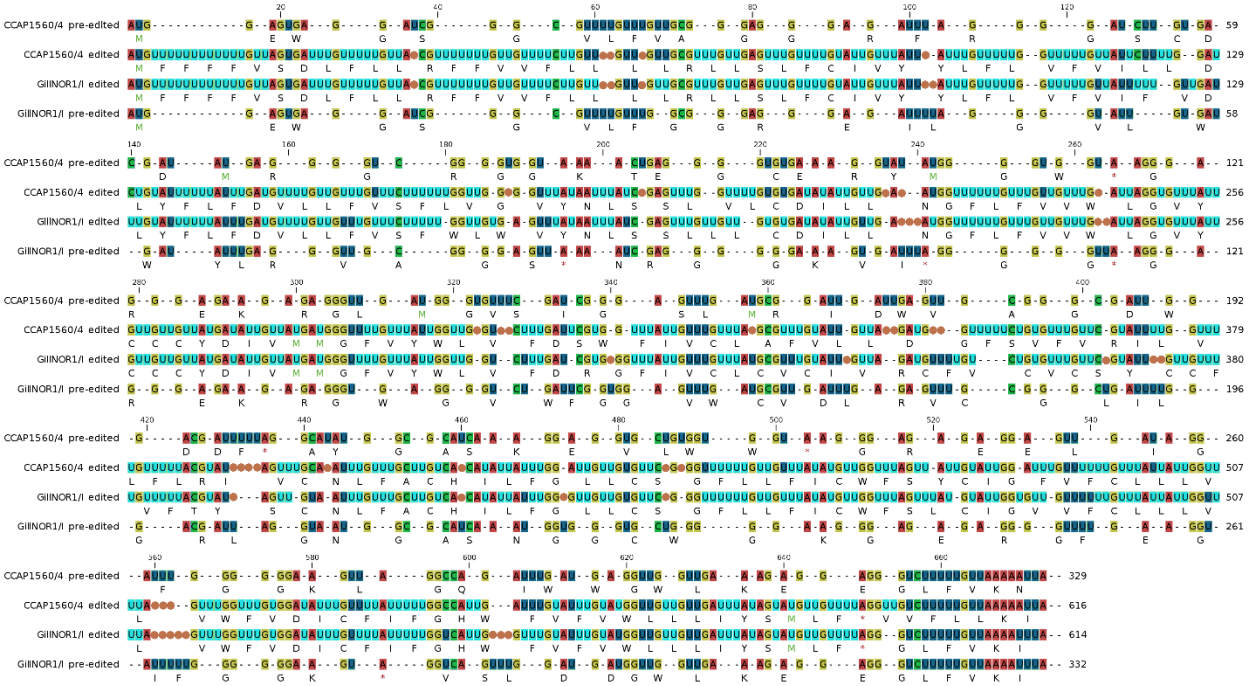
cox3



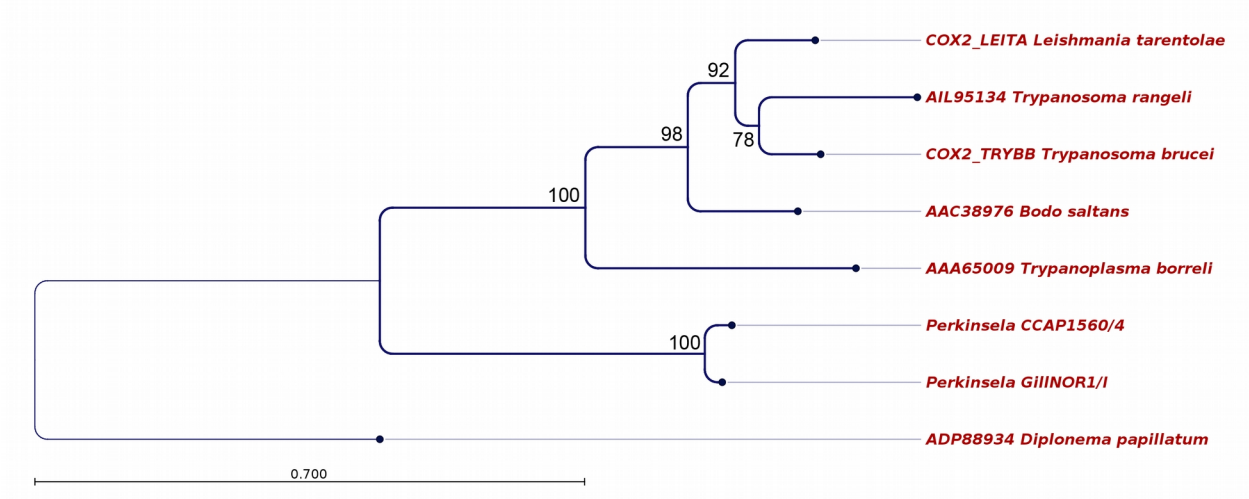
	1	2	3	4
CCAP1560/4 pre-edited	1		165	274	118
CCAP1560/4 edited	2	78.79		115	271
GillNOR1/I edited	3	64.92	85.26		166
GillNOR1/I pre-edited	4	81.12	65.26	78.75	



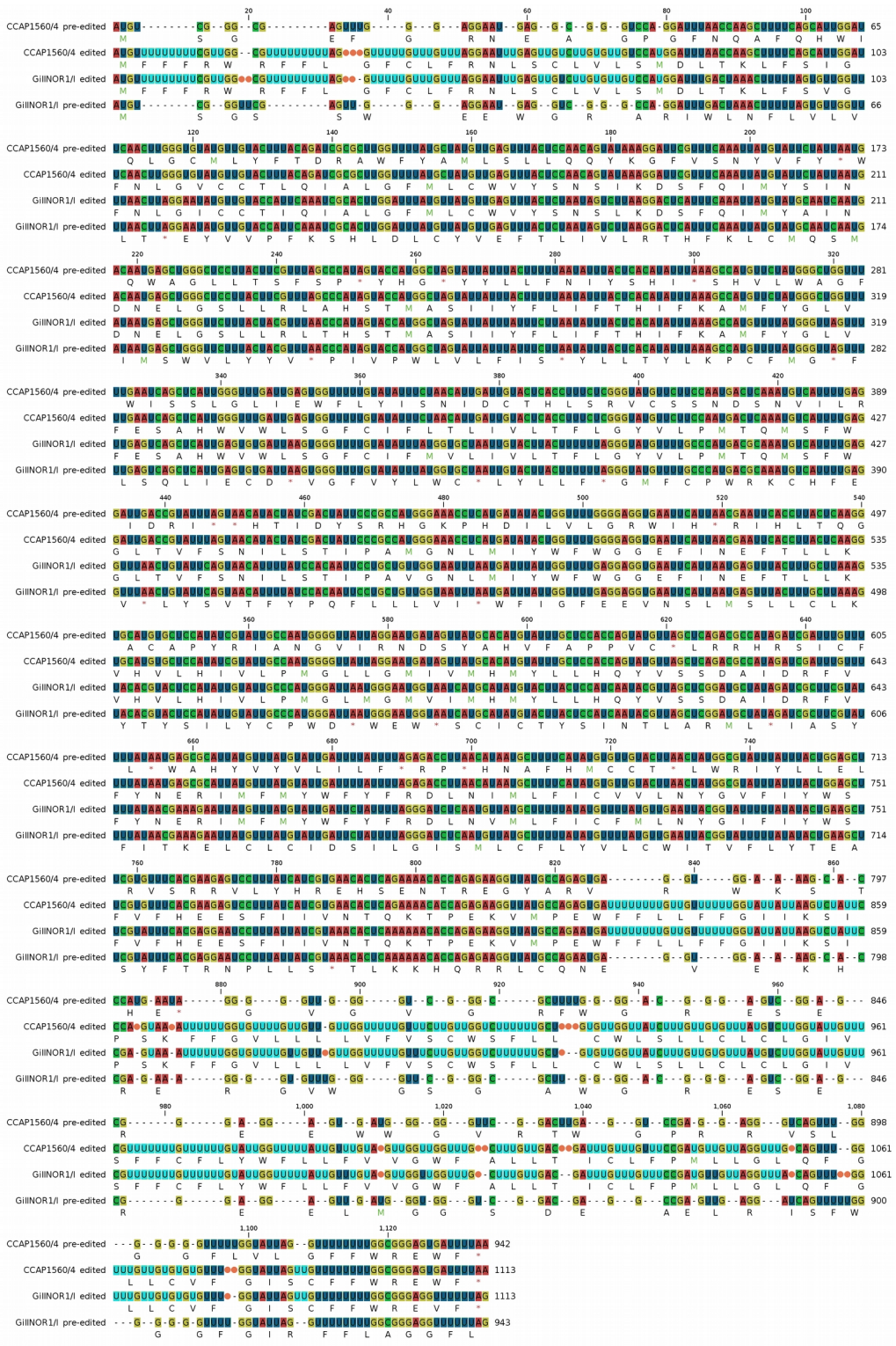
atp6



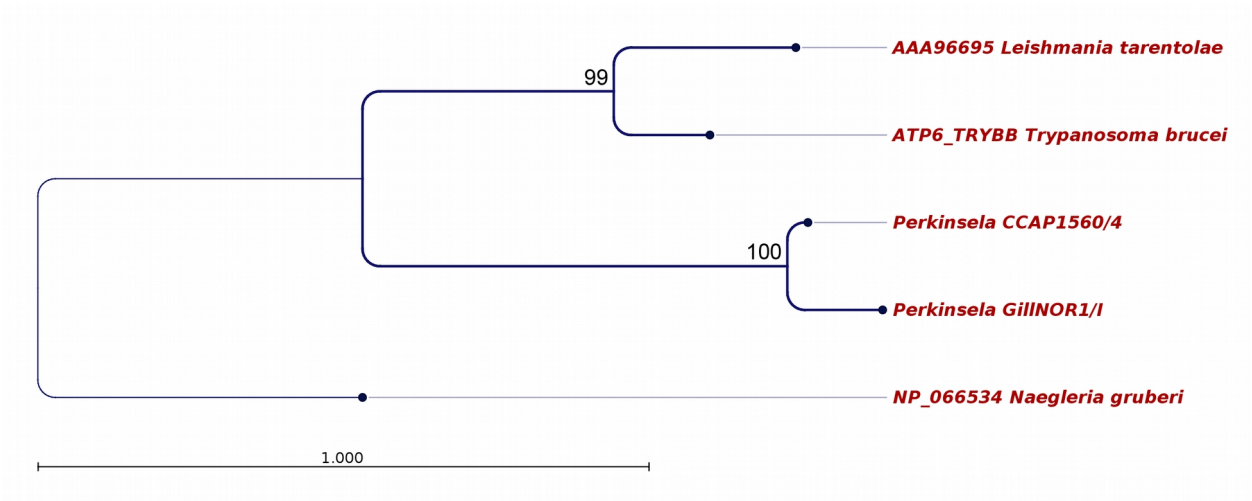
	1	2	3	4
CCAP1560/4 pre-edited	1		240	286	85
CCAP1560/4 edited	2	64.07		56	289
GillNOR1/I edited	3	57.12	91.24		240
GillNOR1/I pre-edited	4	82.33	56.87	64.07	



cob


CCAP1560/4 pre-edited
CCAP1560/4 edited
GilINOR1/I edited
GilINOR1/I pre-edited

		1	2	3	4
CCAP1560/4 pre-edited	1		347	346	76
CCAP1560/4 edited	2	46.28		41	355
GillNOR1/I edited	3	45.94	93.50		340
GillNOR1/I pre-edited	4	79.18	45.22	47.12	



#	*T.brucei* gene ID	gene name (alternative name)	product description	protein lengt	complex or function	status	E-value (HMMER3)	bit score (HMMER3)	bit score/query length	bias (HMMER3)	orthologous group ID	ortholog count	paralog count	reciprocal BLASTp hit in T.brucei
1	Tb927.10.13620	NDUFA9	NADH-ubiquinone oxidoreductase subunit	485	OxPhos complex I	missing	no hit				OG5_147054	20	0	N/A
2	Tb927.11.1320	NDUFS7 (NDHK)	NADH-ubiquinone oxidoreductase 20 kDa subunit	202	OxPhos complex I	missing	no hit				OG5_127327	19	0	N/A
3	Tb927.11.15810	NI2M	NADH-ubiquinone oxidoreductase subunit	304	OxPhos complex I	missing	no hit				OG5_157993	19	0	N/A
4	Tb927.11.16870	NI8M	NADH-ubiquinone oxidoreductase subunit	165	OxPhos complex I	missing	no hit				OG5_154698	19	0	N/A
5	Tb927.11.8910	NB6M	NADH-ubiquinone oxidoreductase subunit	173	OxPhos complex I	missing	no hit				OG5_146017	19	0	N/A
6	Tb927.11.9930		NADH-ubiquinone oxidoreductase subunit	256	OxPhos complex I	missing	no hit				OG5_151879	20	0	N/A
7	Tb927.5.450	NUBM	NADH-ubiquinone oxidoreductase subunit	496	OxPhos complex I	missing	no hit				OG5_127601	20	1	N/A
8	Tb927.7.6350		NADH-ubiquinone oxidoreductase subunit	273	OxPhos complex I	missing	no hit				OG5_127830	20	0	N/A
9	Tb927.9.15380	NDUFA9	NADH-ubiquinone oxidoreductase subunit	373	OxPhos complex I	missing	no hit				OG5_128491	19	1	N/A
10	Tb927.10.9440	NDH2	54 kDa alternative NADH dehydrogenase	491	alternative NADH dehydrogenase	missing	no hit				OG5_126960	19	0	N/A
11	Tb927.7.3590	SDH	hypothetical protein, conserved	151	OxPhos complex II	missing	no hit				OG5_151497	20	0	N/A
12	Tb927.8.6580	SDH	succinate dehydrogenase flavoprotein	609	OxPhos complex II	present	3E-299	993	1.63	0.1	OG5_126927	20	0	YES
13	Tb927.6.2490	SDH	hypothetical protein, conserved	240	OxPhos complex II	present	1E-037	128	0.53	2.8	OG5_146027	19	0	YES
14	Tb927.10.2680	SDH10	hypothetical protein, conserved	575	OxPhos complex II	missing	no hit				OG5_153698	20	0	N/A
15	Tb927.8.6890	SDH11	hypothetical protein, conserved	88	OxPhos complex II	missing	no hit				OG5_151655	16	0	N/A
16	Tb927.9.5960	SDH2C	succinate dehydrogenase	188	OxPhos complex II	missing	no hit				OG5_126893	37	1	N/A
17	Tb927.8.3380	SDH2N	electron transfer protein	242	OxPhos complex II	missing	no hit				OG5_126893	37	1	N/A
18	Tb927.6.4130	SDH3	hypothetical protein, conserved	104	OxPhos complex II	missing	no hit				OG5_148872	18	0	N/A
19	Tb927.10.11770	SDH4	hypothetical protein, conserved	129	OxPhos complex II	missing	no hit				OG5_151805	16	0	N/A
20	Tb927.3.3460	SDH5	hypothetical protein, conserved	483	OxPhos complex II	present	2E-016	57.3	0.12	5.6	OG5_154613	20	0	YES
21	Tb927.8.5640	SDH6	hypothetical protein, conserved	333	OxPhos complex II	present	9E-043	145	0.43	0.1	OG5_143915	22	0	YES
22	Tb927.2.4700	SDH8	hypothetical protein, conserved	151	OxPhos complex II	present	5E-028	95.8	0.63	0.3	OG5_151721	16	0	YES
23	Tb927.10.3040	SDH9	hypothetical protein, conserved	135	OxPhos complex II	present	3E-024	83.4	0.62	0.1	OG5_161906	16	0	YES
24	Tb927.8.7430		ubiquinol-cytochrome c reductase	71	OxPhos complex III	missing	5E-019	66.8	0.94	7.8	OG5_128457	17	1	NO
25	Tb927.8.1890	cytC1	cytochrome c1, heme protein, mitochondrial precurs	258	OxPhos complex III	present	3E-085	284	1.10	0	OG5_128006	19	0	YES
26	Tb927.9.14160	RISP	Rieske iron-sulfur protein	297	OxPhos complex III	present	2E-069	233	0.79	0	OG5_127574	20	0	YES
27	Tb927.8.5120	cytC	cytochrome *C*	114	cytochrome *C*	present	2E-050	168	1.48	0.1	OG5_127365	26	0	YES
28	Tb927.1.4100	COXIV	cytochrome oxidase subunit IV	353	OxPhos complex IV	present	8E-140	464	1.31	8.9	OG5_148365	21	0	YES
29	Tb927.9.3170	COXV	cytochrome oxidase subunit V	196	OxPhos complex IV	missing	no hit				OG5_148654	16	0	N/A
30	Tb927.10.280	COXVI	cytochrome oxidase subunit VI	158	OxPhos complex IV	present	5E-058	194	1.22	6.3	OG5_148032	18	0	YES
31	Tb927.3.1410	COXVII	cytochrome oxidase subunit VII	165	OxPhos complex IV	present	8E-052	174	1.05	5.4	OG5_140920	21	0	YES
32	Tb927.4.4620	COXVIII	cytochrome oxidase subunit VIII	157	OxPhos complex IV	present	1E-031	108	0.69	2.6	OG5_146050	16	0	YES
33	Tb927.10.8320	COXIX	cytochrome oxidase subunit IX	124	OxPhos complex IV	missing	no hit				OG5_149149	17	0	N/A
34	Tb927.11.13140	COXX	cytochrome oxidase subunit X	116	OxPhos complex IV	present	1E-040	136	1.17	0.4	OG5_151630	19	0	YES

35	Tb927.10.3120		cytochrome c oxidase assembly protein	232	OxPhos complex IV	present	3E-065	219	0.94	1.3	OG5_128258	19	0 YES
36	Tb927.7.7420	F1α	ATP synthase alpha chain, mitochondrial precursor	584	OxPhos complex V	present	3E-195	648	1.11	0.3	OG5_127165	32	1 YES
37	Tb927.3.1380	F1β	ATP synthase beta chain, mitochondrial precursor	519	OxPhos complex V	present	4E-255	846	1.63	5.2	OG5_127099	26	0 YES
38	Tb927.10.180	F1γ	ATP synthase gamma chain	305	OxPhos complex V	present	4E-052	175	0.57	3.8	OG5_127077	21	0 YES
39	Tb927.6.4990	F1δ	ATP synthase delta chain	182	OxPhos complex V	present	5E-036	123	0.67	0	OG5_127404	16	0 YES
40	Tb927.10.5050	F1ε	ATP synthase epsilon chain	75	OxPhos complex V	missing	no hit				OG5_151854	17	0 N/A
41	Tb927.11.5280		ATPase subunit 9	118	OxPhos complex V	missing	3E-032	110	0.94	5.2	OG5_126818	54	2 NO
42	Tb927.10.8030	OSCP	hypothetical protein, conserved	255	OxPhos complex V	present	2E-027	93.8	0.37	0.8	OG5_150350	19	0 YES
43	Tb927.4.570		hypothetical protein, conserved	598	OxPhos complex V	missing	4E-019	66.1	0.11	0	OG5_151582	32	1 NO
44	Tb927.11.5780	MTRNAP	mitochondrial DNA-directed RNA polymerase	1274	RNA polymerase	present	###	620	0.49	0.2	OG5_127975	23	0 YES
45	Tb927.11.7900	RBP16	mitochondrial RNA binding protein 16	141	gRNA binding/processing	present	###	71.7	0.51	1.6	OG5_126866	15	0 YES
46	Tb927.11.8400	mRPN1	mitochondrial RNA processing endonuclease 1	486	gRNA binding/processing	missing	no hit				OG5_162282	11	0 N/A
47	Tb927.11.1710	MRP1 (gBP21)	mitochondrial RNA binding protein 1	206	MRP1/MRP2 (gRNA annealing)	present	4E-025	86.6	0.42	2.1	OG5_148774	19	0 YES
48	Tb927.11.13280	MRP2 (gBP25)	mitochondrial RNA binding protein 2	224	MRP1/MRP2 (gRNA annealing)	present	6E-017	60.1	0.27	0	OG5_148320	19	0 YES
49	Tb927.9.4360	KREL1 (REL1)	RNA ligase (U-deletion)	469	RECC (20S editosome)	missing	4E-068	228	0.49	0	OG5_145811	20	0 NO
50	Tb927.1.3030	KREL2 (REL2)	RNA ligase (U-insertion)	416	RECC (20S editosome)	present	2E-083	278	0.67	0	OG5_151366	20	0 YES
51	Tb927.1.1690	KREN1 (REN1)	insertion site specific endonuclease	817	RECC (20S editosome)	missing	no hit				OG5_148564	20	0 N/A
52	Tb927.10.5440	KREN2 (REN2)	deletion site specific endonuclease	538	RECC (20S editosome)	present	2E-020	71.6	0.13	0.8	OG5_149060	21	0 YES
53	Tb927.10.5320	KREN3 (REN3)	*cis*-editing site specific endonuclease	596	RECC (20S editosome)	missing	no hit				OG5_151856	19	1 N/A
54	Tb927.2.2470	KREPA1 (MP81)	structural, U-insertion subdomain organizer	762	RECC (20S editosome)	missing	no hit				OG5_148235	20	0 N/A
55	Tb927.10.8210	KREPA2 (MP63)	structural, U-deletion subdomain organizer	587	RECC (20S editosome)	missing	6E-012	42.8	0.07	15	OG5_147498	20	0 NO
56	Tb927.8.620	KREPA3 (MP42)	structural, U-specific exonuclease	393	RECC (20S editosome)	present	6E-031	106	0.27	7	OG5_143485	19	0 YES
57	Tb927.10.5110	KREPA4 (MP24)	structural, RNA binding	218	RECC (20S editosome)	missing	no hit				OG5_162254	10	0 N/A
58	Tb927.8.680	KREPA5 (MP19)	structural	169	RECC (20S editosome)	missing	no hit				OG5_162359	11	0 N/A
59	Tb927.10.5120	KREPA6 (MP18)	structural, RNA binding	164	RECC (20S editosome)	present	1E-052	176	1.07	0.3	OG5_149055	17	0 YES
60	Tb927.11.2990	KREPB4 (MP46)	structural, heterodimer with endonuclease	414	RECC (20S editosome)	missing	2E-025	87.4	0.21	18	OG5_148952	21	0 NO
61	Tb927.11.940	KREPB5 (MP44)	structural, endonuclease	382	RECC (20S editosome)	present	7E-042	142	0.37	1.2	OG5_151705	18	0 YES
62	Tb927.3.3990	KREPB6 (MP49)	structural, part of KREN3 module	438	RECC (20S editosome)	present	1E-017	62	0.14	0.3	OG5_154623	20	0 YES
63	Tb927.9.5630	KREPB7 (MP47)	structural, part of KREN2 module	411	RECC (20S editosome)	missing	3E-024	83.9	0.20	0.4	OG5_148422	20	0 NO
64	Tb927.8.5690	KREPB8 (MP41)	structural, part of KREN1 module	368	RECC (20S editosome)	missing	no hit				OG5_148433	19	0 N/A
65	Tb927.7.3950	KRET1 (RET1)	gRNA 3'-terminal uridylyl transferase (TUTase)	975	RECC (20S editosome)	present	6E-121	403	0.41	0.3	OG5_151380	21	0 YES
66	Tb927.7.1550	KRET2 (RET2)	RNA editing 3'-terminal uridylyl transferase (TUTase	487	RECC (20S editosome)	present	4E-132	440	0.90	0	OG5_148715	20	0 YES
67	Tb927.7.1070	KREX1 (MP100)	U-specific exonuclease, 3′ nucleotidyl phosphatase	894	RECC (20S editosome)	present	5E-132	440	0.49	0	OG5_148723	20	0 YES
68	Tb927.10.3570	KREX2 (MP99)	U-specific exonuclease, 3′ nucleotidyl phosphatase (	907	RECC (20S editosome)	present	2E-099	332	0.37	0	OG5_152728	20	1 YES
69	Tb927.1.1330	MEAT1	Mitochondrial Editosome-like Complex TUTase	406	mitochondrial editosome-like con	missing	no hit				OG5_141003	21	1 N/A
70	Tb927.11.8870	KREH1 (REH1)	RNA editing associated helicase 1, RECC subunit	546	helicase	present	9E-105	349	0.64	0	OG5_148828	20	0 YES
71	Tb927.4.1500	KREH2 (REH2)	RNA editing associated helicase 2, MRB1 subunit	2167	helicase	present	0	###	0.60	0	OG5_136717	40	3 YES
72	Tb927.2.3800	GAP1 (GRBC2)	MRB1 core subunit, gRNA-binding	492	MRB1 (GRBC)	present*	4E-117	390	0.79	0.1	OG5_148963	19	0 YES*
73	Tb927.7.2570	GAP2 (GRBC1)	MRB1 core subunit, gRNA-binding	473	MRB1 (GRBC)	present	2E-061	206	0.44	0	OG5_145963	21	1 YES
74	Tb927.10.11870	MRB11870	MRB1 core subunit	310	MRB1 (GRBC)	present	9E-123	408	1.31	0.1	OG5_148940	19	0 YES
75	Tb927.5.3010	MRB3010	MRB1 core subunit	516	MRB1 (GRBC)	present	2E-058	197	0.38	6.7	OG5_139425	21	1 YES
76	Tb11.02.5390	MRB5390	MRB1 core subunit	1087	MRB1 (GRBC)	missing	9E-170	565	0.52	0.1	OG5_127058	20	1 NO
77	Tb927.11.16860	MRB8620	MRB1 core subunit	482	MRB1 (GRBC)	missing	no hit				OG5_154699	20	0 N/A
78	Tb927.11.9140	MRB0880	MRB1 subunit	174	MRB1 (GRBC)	missing	no hit				OG5_127057	0	0 N/A
79	Tb927.10.10130	MRB10130	MRB1 subunit	545	MRB1 (GRBC)	missing	no hit				OG5_148079	20	0 N/A
80	Tb927.3.1590	MRB1590	MRB1 subunit	668	MRB1 (GRBC)	present	5E-156	519	0.78	0	OG5_135848	20	0 YES

81	Tb927.6.1680	MRB1680	MRB1 subunit	524	MRB1 (GRBC)	missing	no hit				OG5_135561	20	0	N/A
82	Tb927.2.1860	MRB1860	MRB1 subunit	872	MRB1 (GRBC)	missing	no hit				OG5_146066	21	0	N/A
83	Tb927.4.4150	MRB4150/MRB8180	MRB1 subunit	934	MRB1 (GRBC)	missing	no hit				OG5_142243	21	1	N/A
84	Tb927.4.4160	MRB4160	MRB1 subunit	915	MRB1 (GRBC)	missing	no hit				OG5_142244	23	1	N/A
85	Tb927.8.8170	MRB8170	MRB1 subunit	905	MRB1 (GRBC)	missing	no hit				OG5_142244	23	1	N/A
86	Tb927.2.6070	MRB6070	MRB1 subunit	285	MRB1 (GRBC)	present	1E-032	111	0.39	178	OG5_152437	15	0	YES
87	Tb927.7.800	MRB800	MRB1 subunit	543	MRB1 (GRBC)	present	6E-083	277	0.51	8.9	OG5_154591	19	0	YES
88	Tb927.10.380	PPR5 (KRIPP5)	MRB1 subunit	342	MRB1 (GRBC) / LSU	present	3E-100	334	0.98	7	OG5_148611	19	1	YES
89	Tb927.6.2230	TbRGG1	RGG-containing protein 1, MRB1 subunit	775	MRB1 (GRBC)	missing	no hit				OG5_142377	21	0	N/A
90	Tb927.10.10830	TbRGG2 (RGGm)	RGG-containing protein 2, MRB1 subunit	320	MRB1 (GRBC)	missing	no hit				OG5_137263	20	0	N/A
91	Tb927.3.1820	TbRGG3 (MRB1820)	RGG-containing protein 3, MRB1 subunit	245	MRB1 (GRBC)	present	3E-013	47.5	0.19	33	OG5_140599	18	0	YES
92	Tb927.11.15850	KPAP1	kinteoplast poly(A) polymerase / SSU	754	mRNA polyadenylation/uridylatio	present	1E-122	409	0.54	22	OG5_154687	21	0	YES
93	Tb927.2.3180	KPAF1 (PPR1)	kinetoplast polyadenylation/uridylation factor 1	1003	mRNA polyadenylation/uridylatio	present	1E-257	855	0.85	17	OG5_137763	22	1	YES
94	Tb927.11.14380	KPAF2	kinetoplast polyadenylation/uridylation factor 2	643	mRNA polyadenylation/uridylatio	missing	no hit				OG5_154655	19	0	N/A
95	Tb927.10.6850	mtRPS18	mitochondrial edited mRNA stability factor 1 subuni	320	SSU*	present	4E-074	247	0.77	0.8	OG5_151867	20	0	YES
96	Tb927.6.4930	Rhod	rhodanese domain protein; thiosulfate sulfurtransfe	247	SSU*	present	9E-125	413	1.67	6.7	OG5_150803	22	0	YES
97	Tb927.11.13890		hypothetical protein, conserved	268	SSU*	present	3E-015	53.7	0.20	0.1	OG5_133322	19	0	YES
98	Tb927.5.3360	TbMRPL2	50S ribosomal protein L2	411	LSU conserved	present	7E-127	422	1.03	3.7	OG5_148095	18	0	YES
99	Tb927.3.5610	TbMRPL3	ribosomal protein L3 mitochondrial	473	LSU conserved	present	3E-096	321	0.68	0.3	OG5_127133	21	0	YES
100	Tb927.11.6000	TbMRPL4	hypothetical protein, conserved	351	LSU conserved	present	1E-156	520	1.48	1.2	OG5_150007	19	0	YES
101	Tb927.7.4550	TbMRPL7/12	60S ribosomal protein-like	183	LSU conserved	present	7E-025	86	0.47	4.1	OG5_126884	39	2	YES
102	Tb927.5.3410	TbMRPL9	hypothetical protein, conserved	263	LSU conserved	present	9E-067	223	0.85	2.1	OG5_148303	20	0	YES
103	Tb927.2.4890	TbMRPL11	ribosomal protein L11	342	LSU conserved	missing	4E-039	133	0.39	0	OG5_127103	24	1	NO
104	Tb927.4.1070	TbMRPL13	50S ribosomal protein L13	202	LSU conserved	present	1E-027	94.8	0.47	0.3	OG5_127268	19	0	YES
105	Tb927.4.930	TbMRPL14	50S ribosomal protein L14	189	LSU conserved	present	5E-034	116	0.61	0.8	OG5_146650	17	0	YES
106	Tb927.5.3980	TbMRPL15	hypothetical protein, conserved	374	LSU conserved	present	8E-119	395	1.06	1.6	OG5_146726	20	0	YES
107	Tb927.7.3960	TbMRPL16	50S ribosomal protein L16	167	LSU conserved	present	4E-064	214	1.28	0	OG5_149623	16	0	YES
108	Tb927.8.5860	TbMRPL17	50S ribosomal protein L17	301	LSU conserved	present	5E-077	257	0.85	0.9	OG5_146611	19	0	YES
109	Tb927.11.10170	TbMRPL20	hypothetical protein, conserved	213	LSU conserved	present	7E-022	75.9	0.36	0.2	OG5_145747	16	0	YES
110	Tb927.7.4140	TbMRPL21	hypothetical protein, conserved	188	LSU conserved	present	4E-025	87	0.46	0.3	OG5_127445	18	0	YES
111	Tb927.7.2760	TbMRPL22	hypothetical protein, conserved	278	LSU conserved	present	1E-076	256	0.92	0.1	OG5_141619	21	0	YES
112	Tb927.11.870	TbMRPL23	hypothetical protein, conserved	246	LSU conserved	present	5E-040	135	0.55	0.8	OG5_154063	18	0	YES
113	Tb927.3.1710	TbMRPL24	hypothetical protein, conserved	378	LSU conserved	present	3E-053	179	0.47	1.1	OG5_127372	20	0	YES
114	Tb927.11.3640	TbMRPL27	hypothetical protein, conserved	185	LSU conserved	present	1E-023	81.9	0.44	4.6	OG5_151490	16	0	YES
115	Tb927.6.4040	TbMRPL28	hypothetical protein, conserved	241	LSU conserved	present	3E-055	185	0.77	0	OG5_144772	21	0	YES
116	Tb927.10.600	TbMRPL29	hypothetical protein, conserved	541	LSU conserved	present	###	277	0.51	4.6	OG5_151613	18	0	YES
117	Tb927.9.8290	TbMRPL30	hypothetical protein, conserved	218	LSU conserved	present	3E-041	140	0.64	2.6	OG5_153700	17	0	YES
118	Tb927.4.1810	TbMRPL33	hypothetical protein, conserved	114	LSU conserved	present	3E-029	99.4	0.87	0	OG5_148537	18	0	YES
119	Tb927.11.14980	TbMRPL38	hypothetical protein, conserved	507	LSU conserved	present	1E-056	190	0.38	11	OG5_154671	20	0	YES
120	Tb927.4.4600	TbMRPL43	hypothetical protein, conserved	260	LSU conserved	present	2E-065	219	0.84	4.3	OG5_144100	20	0	YES
121	Tb927.7.4710	TbMRPL46	hypothetical protein, conserved	296	LSU conserved	present	7E-014	49.6	0.17	0.2	OG5_129089	22	0	YES
122	Tb927.9.7170	TbMRPL47	hypothetical protein, conserved	471	LSU conserved	present	5E-106	353	0.75	5	OG5_149408	17	0	YES
123	Tb927.5.3110	TbMRPL49	hypothetical protein, conserved	218	LSU conserved	present	8E-034	115	0.53	0.7	OG5_148200	20	0	YES
124	Tb927.11.4650	TbMRPL52	hypothetical protein, conserved	1522	LSU conserved	missing	no hit				OG5_140913	29	0	N/A
125	Tb927.1.1160	KRIPP3	kinetoplast ribosomal PPR-repeat containing protei	531	LSU recognized domains	missing	2E-052	176	0.33	0	OG5_145939	19	1	NO
126	Tb927.11.9450	PPIase	cyclophilin type peptidyl-prolyl cis-trans isomerase (	190	LSU recognized domains	present	1E-043	147	0.77	0.1	OG5_140933	22	0	YES

127	Tb927.7.3430	PPIase	cyclophilin-type peptidyl-prolyl cis-trans isomerase (	231	LSU recognized domains	present	2E-073	245	1.06	0	OG5_141132	20	0	YES
128	Tb927.4.2720	RH	ATP dependent DEAD-box helicase (RH)	739	LSU recognized domains	present	9E-060	201	0.27	0	OG5_143922	21	1	YES
129	Tb927.7.1640	TbEAR	ras-like small GTPase (TbEAR)	576	LSU recognized domains	present	1E-108	362	0.63	0	OG5_128684	20	0	YES
130	Tb927.10.12050		hypothetical protein, conserved	289	LSU recognized domains	present	1E-048	164	0.57	0.5	OG5_148943	21	0	YES
131	Tb927.10.6090		tRNA pseudouridine synthase A	688	LSU recognized domains	present	2E-045	154	0.22	0	OG5_128305	20	1	YES
132	Tb927.11.15500		hypothetical protein, conserved	283	LSU recognized domains	present	5E-023	80	0.28	0.2	OG5_154679	21	0	YES
133	Tb927.11.16990		hypothetical protein, conserved	655	LSU recognized domains	present	2E-043	147	0.22	0	OG5_152569	21	1	YES
134	Tb927.11.5880		hypothetical protein, conserved	557	LSU recognized domains	missing	no hit				OG5_145985	21	0	N/A
135	Tb927.11.5990		hypothetical protein, conserved	616	LSU recognized domains	present	4E-056	188	0.31	0	OG5_151483	19	0	YES
136	Tb927.6.2480		chaperone protein DNAj	345	LSU recognized domains	missing	6E-033	112	0.33	1.2	OG5_142862	18	0	NO
137	Tb927.6.3600		hypothetical protein, conserved	439	LSU recognized domains	missing	no hit				OG5_148049	21	0	N/A
138	Tb927.6.3930		hypothetical protein, conserved	426	LSU recognized domains	present	4E-051	172	0.40	9.2	OG5_151778	20	0	YES
139	Tb927.6.4200		hypothetical protein, conserved	444	LSU recognized domains	present	2E-017	61.3	0.14	0.2	OG5_146041	20	0	YES
140	Tb927.7.2630		hypothetical protein, conserved	900	LSU recognized domains	present	6E-055	185	0.21	0.2	OG5_127209	23	0	YES
141	Tb927.7.3460		hypothetical protein, conserved	449	LSU recognized domains	present	2E-020	70.9	0.16	9.1	OG5_148628	21	0	YES
142	Tb927.7.6800		hypothetical protein, conserved	378	LSU recognized domains	missing	no hit				OG5_148463	20	0	N/A
143	Tb927.8.2760		hypothetical protein, conserved	477	LSU recognized domains	present	3E-067	225	0.47	0.3	OG5_146653	20	0	YES
144	Tb927.8.3170		hypothetical protein, conserved	796	LSU recognized domains	present	3E-077	259	0.33	4.8	OG5_148650	21	0	YES
145	Tb927.9.12850		hypothetical protein, conserved	586	LSU recognized domains	present	3E-050	169	0.29	0	OG5_146898	20	0	YES
146	Tb927.9.14050		hypothetical protein, conserved	524	LSU recognized domains	present	5E-047	159	0.30	1.2	OG5_151819	19	0	YES
147	Tb927.9.3350		pseudouridylate synthase	406	LSU recognized domains	present	6E-067	225	0.55	0	OG5_129784	21	0	YES
148	Tb927.9.9150		GTP-binding protein	451	LSU recognized domains	present	6E-063	211	0.47	0	OG5_128449	19	0	YES
149	Tb927.10.11050		hypothetical protein, conserved	312	LSU kinetoplastid-specific	missing	no hit				OG5_148914	21	0	N/A
150	Tb927.10.11350		hypothetical protein, conserved	133	LSU kinetoplastid-specific	missing	9E-030	102	0.76	2.1	OG5_148925	17	0	NO
151	Tb927.10.1870		hypothetical protein, conserved	181	LSU kinetoplastid-specific	present	4E-016	57.3	0.32	1	Tb927.10.1870	0	0	YES
152	Tb927.10.7380		hypothetical protein, conserved	349	LSU kinetoplastid-specific	present	1E-078	263	0.75	1.2	OG5_151873	20	0	YES
153	Tb927.11.10050		hypothetical protein, conserved	102	LSU kinetoplastid-specific	missing	no hit				OG5_151874	16	0	N/A
154	Tb927.11.10080		hypothetical protein, conserved	189	LSU kinetoplastid-specific	present	4E-027	93.2	0.49	0	OG5_146106	17	0	YES
155	Tb927.11.10570		hypothetical protein, conserved	333	LSU kinetoplastid-specific	present	1E-078	262	0.79	5.1	OG5_146122	20	0	YES
156	Tb927.11.11630		hypothetical protein, conserved	242	LSU kinetoplastid-specific	present	8E-086	285	1.18	15	OG5_148831	20	0	YES
157	Tb927.11.1630		hypothetical protein, conserved	831	LSU kinetoplastid-specific	missing	no hit				OG5_148770	20	0	N/A
158	Tb927.11.5530		hypothetical protein, conserved	262	LSU kinetoplastid-specific	missing	no hit				OG5_151593	20	0	N/A
159	Tb927.11.8040		hypothetical protein, conserved	185	LSU kinetoplastid-specific	present	2E-036	124	0.67	1.7	OG5_148814	17	0	YES
160	Tb927.11.9830		hypothetical protein, conserved	197	LSU kinetoplastid-specific	present	1E-030	104	0.53	2.9	OG5_149091	17	0	YES
161	Tb927.3.820		hypothetical protein, conserved	188	LSU kinetoplastid-specific	present	3E-050	168	0.90	0.1	OG5_154592	15	0	YES
162	Tb927.4.4610		hypothetical protein, conserved	319	LSU kinetoplastid-specific	missing	no hit				OG5_143940	21	0	N/A
163	Tb927.5.2070		hypothetical protein, conserved	634	LSU kinetoplastid-specific	missing	no hit				OG5_148465	19	0	N/A
164	Tb927.5.3870		hypothetical protein, conserved	731	LSU kinetoplastid-specific	missing	no hit				OG5_148430	20	0	N/A
165	Tb927.5.4120		hypothetical protein, conserved	191	LSU kinetoplastid-specific	missing	no hit				OG5_148906	17	0	N/A
166	Tb927.6.1440		hypothetical protein, conserved	258	LSU kinetoplastid-specific	present	5E-048	162	0.63	0.1	OG5_148844	20	0	YES
167	Tb927.7.2990		hypothetical protein, conserved	309	LSU kinetoplastid-specific	present	8E-071	236	0.77	0.2	OG5_143899	23	1	YES
168	Tb927.7.3510		hypothetical protein, conserved	482	LSU kinetoplastid-specific	present	2E-037	127	0.26	11	OG5_151499	20	0	YES
169	Tb927.7.7010		hypothetical protein, conserved	154	LSU kinetoplastid-specific	present	1E-017	62.5	0.41	0.3	OG5_148265	17	0	YES
170	Tb927.8.1880		hypothetical protein, conserved	190	LSU kinetoplastid-specific	present	6E-027	92.4	0.49	0.4	Tb927.8.1880	0	0	YES
171	Tb927.8.3300		hypothetical protein, conserved	691	LSU kinetoplastid-specific	present	1E-149	498	0.72	0.6	OG5_151642	18	0	YES
172	Tb927.9.3640		hypothetical protein, conserved	198	LSU kinetoplastid-specific	missing	no hit				OG5_148224	17	0	N/A

173	Tb927.6.4080		hypothetical protein, conserved	205	LSU/SSU kinetoplastid-specific	present	2E-101	336	1.64	1.1	OG5_148871	20	0 YES
174	Tb927.10.6300	TbMRPS5	hypothetical protein, conserved	435	SSU conserved/SSU*	present	7E-172	570	1.31	1.1	OG5_151368	18	1 YES
175	Tb927.10.2800	TbMRPS6	hypothetical protein, conserved	160	SSU conserved	present	8E-041	138	0.86	0.1	OG5_154558	15	0 YES
176	Tb927.10.13300	TbMRPS8	30S ribosomal protein S8	282	SSU conserved/SSU*	present	6E-042	142	0.50	0.3	OG5_150451	22	0 YES
177	Tb927.8.3110	TbMRPS9	hypothetical protein, conserved	443	SSU conserved/SSU*	missing	2E-017	60.8	0.14	0	OG5_151640	22	1 NO
178	Tb927.10.10400	TbMRPS11	hypothetical protein, conserved	326	SSU conserved/SSU*	present	4E-071	237	0.73	0.1	OG5_148256	21	0 YES
179	Tb927.1.1200	TbMRPS15	SSU ribosomal protein, mitochondrial (MRPS15)	429	SSU conserved	present	2E-106	354	0.83	3.4	OG5_150564	19	0 YES
180	Tb927.11.7790	TbMRPS16	hypothetical protein, conserved	188	SSU conserved	present	1E-046	157	0.83	0.8	OG5_151731	16	0 YES
181	Tb927.9.11280	TbMRPS17	unspecified product	307	SSU conserved/SSU*	present	6E-055	185	0.60	0	OG5_128969	16	0 YES
182	Tb927.6.1250	TbMRPS29	hypothetical protein, conserved	498	SSU conserved	present	1E-127	425	0.85	0	OG5_139168	21	1 YES
183	Tb927.8.5280	TbMRPS34	hypothetical protein, conserved	257	SSU conserved/SSU*	present	5E-084	280	1.09	0.1	OG5_148747	28	0 YES
184	Tb927.11.5500	KRIPP1	kinetoplast ribosomal PPR-repeat containing protei	812	SSU recognized domains/SSU*	present	5E-213	708	0.87	###	OG5_148670	21	0 YES
185	Tb927.1.2990		hypothetical protein, conserved	1024	SSU recognized domains	present	4E-158	526	0.51	###	OG5_148576	20	0 YES
186	Tb927.10.11820		hypothetical protein, conserved	334	SSU recognized domains	present	1E-095	319	0.95	###	OG5_148939	19	0 YES
187	Tb927.10.15650		tRNA pseudouridine synthase A-like protein	579	SSU recognized domains	present	4E-079	265	0.46	###	OG5_139963	19	1 YES
188	Tb927.11.10150		hypothetical protein, conserved	1211	SSU recognized domains	present	2E-236	785	0.65	###	OG5_146103	21	0 YES
189	Tb927.11.11870		hypothetical protein, conserved	349	SSU recognized domains	missing	no hit				OG5_151746	19	0 N/A
190	Tb927.11.5060		hypothetical protein, conserved	1041	SSU recognized domains	present	1E-173	578	0.55	###	OG5_142397	19	0 YES
191	Tb927.3.2260		hypothetical protein, conserved	228	SSU recognized domains	present	2E-034	117	0.51	###	OG5_157943	18	0 YES
192	Tb927.3.5240	KRIPP8	hypothetical protein, conserved	581	SSU recognized domains/SSU*	present	2E-066	223	0.38	###	OG5_154606	21	0 YES
193	Tb927.3.970		hypothetical protein, conserved	370	SSU recognized domains	present	6E-074	247	0.67	###	OG5_144218	19	0 YES
194	Tb927.4.3690		hypothetical protein, conserved	439	SSU recognized domains	missing	no hit				OG5_148518	21	0 N/A
195	Tb927.7.2620		hypothetical protein, conserved	294	SSU recognized domains	present	1E-095	319	1.08	###	OG5_145964	20	0 YES
196	Tb927.8.4860		hypothetical protein, conserved	679	SSU recognized domains	present	4E-048	162	0.24	###	OG5_151514	19	0 YES
197	Tb927.10.11260		hypothetical protein, conserved	187	SSU kinetoplastid-specific	present	4E-036	122	0.65	###	OG5_148921	20	0 YES
198	Tb927.10.13820		hypothetical protein, conserved	261	SSU kinetoplastid-specific	missing	no hit				OG5_148480	19	0 N/A
199	Tb927.10.16090		hypothetical protein, conserved	803	SSU kinetoplastid-specific	present	6E-054	182	0.23	###	OG5_148515	21	1 YES
200	Tb927.10.3250		hypothetical protein, conserved	307	SSU kinetoplastid-specific/SSU*	present	1E-090	302	0.98	###	OG5_157912	19	1 YES
201	Tb927.10.3580		hypothetical protein, conserved	324	SSU kinetoplastid-specific/SSU*	missing	no hit				OG5_154514	20	0 N/A
202	Tb927.11.10400		hypothetical protein, conserved	179	SSU kinetoplastid-specific	present	2E-032	110	0.61	###	OG5_149100	15	0 YES
203	Tb927.11.11470	KRIPP14	hypothetical protein, conserved	282	SSU kinetoplastid-specific/SSU*	present	8E-042	141	0.50	###	OG5_151742	18	0 YES
204	Tb927.11.1250		mitochondrial edited mRNA stability factor 1 subuni	874	SSU kinetoplastid-specific/SSU*	present	4E-028	95.9	0.11	###	OG5_148764	20	0 YES
205	Tb927.11.2530		mitochondrial RNA binding complex 1 subunit, kinte	747	SSU kinetoplastid-specific/SSU*	present	6E-020	68.8	0.09	###	OG5_148961	22	0 YES
206	Tb927.2.4400		hypothetical protein, conserved	1181	SSU kinetoplastid-specific/SSU*	present	5E-148	493	0.42	###	OG5_148788	21	0 YES
207	Tb927.3.770		hypothetical protein, conserved	181	SSU kinetoplastid-specific	present	1E-037	128	0.70	###	OG5_148757	19	0 YES
208	Tb927.5.1510		hypothetical protein, conserved	312	SSU kinetoplastid-specific/SSU*	present	3E-062	209	0.67	###	OG5_145904	20	0 YES
209	Tb927.5.1790	PPR29	hypothetical protein, conserved	631	SSU kinetoplastid-specific/SSU*	present	7E-068	228	0.36	###	OG5_145907	26	0 YES
210	Tb927.5.3640		hypothetical protein, conserved	270	SSU kinetoplastid-specific	present	4E-034	116	0.43	###	OG5_145864	21	0 YES
211	Tb927.5.4040	coiled coil	hypothetical protein, conserved	817	SSU kinetoplastid-specific/SSU*	present	2E-110	368	0.45	###	OG5_143946	24	0 YES
212	Tb927.6.2080	KRIPP22	hypothetical protein, conserved	396	SSU kinetoplastid-specific/SSU*	present	4E-048	162	0.41	###	OG5_151764	20	0 YES
213	Tb927.6.2180		hypothetical protein, conserved	172	SSU kinetoplastid-specific/SSU*	present	8E-030	102	0.59	###	OG5_151767	15	0 YES
214	Tb927.6.4560		hypothetical protein, conserved	407	SSU kinetoplastid-specific/SSU*	present	3E-063	212	0.52	###	OG5_148881	21	0 YES
215	Tb927.6.4580		hypothetical protein, conserved	94	SSU kinetoplastid-specific	present	1E-040	138	1.47	###	OG5_151782	18	0 YES
216	Tb927.7.3050		hypothetical protein, conserved	1165	SSU kinetoplastid-specific/SSU*	present	6E-123	410	0.35	###	OG5_148623	20	0 YES
217	Tb927.7.3240		hypothetical protein, conserved	163	SSU kinetoplastid-specific/SSU*	present	5E-045	151	0.93	###	OG5_151629	16	0 YES
218	Tb927.8.1430		hypothetical protein, conserved	166	SSU kinetoplastid-specific	present	5E-038	129	0.78	###	OG5_151458	15	0 YES

219	Tb927.8.4550		hypothetical protein, conserved	183	SSU kinetoplastid-specific	present	3E-018	64.3	0.35###	OG5_148349	17	0	YES
220	Tb927.8.5200	coiled coil	hypothetical protein, conserved	1788	SSU kinetoplastid-specific/SSU*	present	1E-202	674	0.38###	OG5_145283	22	0	YES
221	Tb927.9.11120		hypothetical protein, conserved	602	SSU kinetoplastid-specific/SSU*	present	1E-259	860	1.43###	OG5_151835	19	0	YES
222	Tb927.9.11880		hypothetical protein, conserved	440	SSU kinetoplastid-specific	present	6E-056	188	0.43###	OG5_149015	19	0	YES
223	Tb927.9.13780		hypothetical protein, conserved	293	SSU kinetoplastid-specific	present	3E-075	251	0.86###	OG5_151820	19	0	YES
224	Tb927.9.5280		unspecified product	274	SSU kinetoplastid-specific	present	4E-070	234	0.86###	OG5_148425	19	0	YES
225	Tb927.9.6510		hypothetical protein, conserved	666	SSU kinetoplastid-specific/SSU*	present	2E-069	233	0.35###	OG5_148420	21	0	YES