University of South Bohemia

Faculty of Science

Assembly and annotation of a mitochondrial genome of kinetoplastid protist *Perkinsela*

Bachelor thesis

Vojtěch David

Supervisor: Msc. Pavel Flegontov Ph.D

České Budějovice 2013

David, V., 2013: Assembly and annotation of mitochondrial genome of kinetoplastid protist *Perkinsela*. Bc. Thesis, in English. – 33p., Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic.

Annotation:

This thesis is based on a collaborative project between Laboratory of Molecular Biology of Protists led by Prof. Julius Lukeš and the laboratory of John Archibald (Department of Biochemistry & Molecular Biology, Dalhousie University, Canada). My contribution to the project, as the name of the thesis suggests is to process DNA and RNA sequencing data from *Neoparamoeba pemaquidensis* isolate that correspond to the mitochondrion of *Perkinsela,* also called *Ichthyobodo*-related organism (IRO). The main aim is to characterize mitochondrial gene content and to attempt to understand the mechanism of U-insertion/deletion RNA editing in *Perkinsela* and its evolution.

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.

Date ....

Signature ....

Thesis acknowledgments

Foremost, I would like to express my sincere gratitude to my supervisor Msc. Pavel Flegontov Ph.D.for the continuous support of my work, for his patience and for beneficial answers to my many questions. My sincere thanks also goes to Prof. RnDr. Julius Lukeš Csc. for offering me participation on *Perkinsela* genome project, as well as for many beneficial consultations and insightful recommendations. I would like to thank to all my labmates, for creating such a friendly and scientific environment. Lastly, I would like to thank my family and those close to me for their moral support.

Table of contents

Introduction

Mitochondria of kinetoplastids

All Kinetoplastea posses single mitochondrion. Its genome-bearing part is termed kinetoplast, which is rather special in many aspects. First, its genome, called kinetoplast DNA or kDNA, is relatively large and in many cases highly compacted. Mitochondrial DNA is missing in diskinetoplastic (without kinetoplast) species/forms *Trypanosoma equiperdum* and *T. evansi* (Schnaufer, 2010,Lun et al., 2010). In species *Trypanosoma* spp. and *Leishmania* spp., which have been for a long time model organisms for Kinetoplastea, along with *Crithidia fasciculata*, there is a disc-like network of thousands of concatenated circles where each so-called maxi-circle is homologous to the mitochondrial genome, and abundant mini-circles encode various guide (g)RNAs for editing of mitochondrial transcripts (Lukes et al., 2002). Alternatively, the closest free-living relative of trypanosomes *Bodo saltans* has its mini/maxi-circle genome less condensed and globular. (Blom et al., 2000) Also dispersed forms of kDNA occur in early branching bodonids (Lukes et al., 2002) (see Fig. 1 for

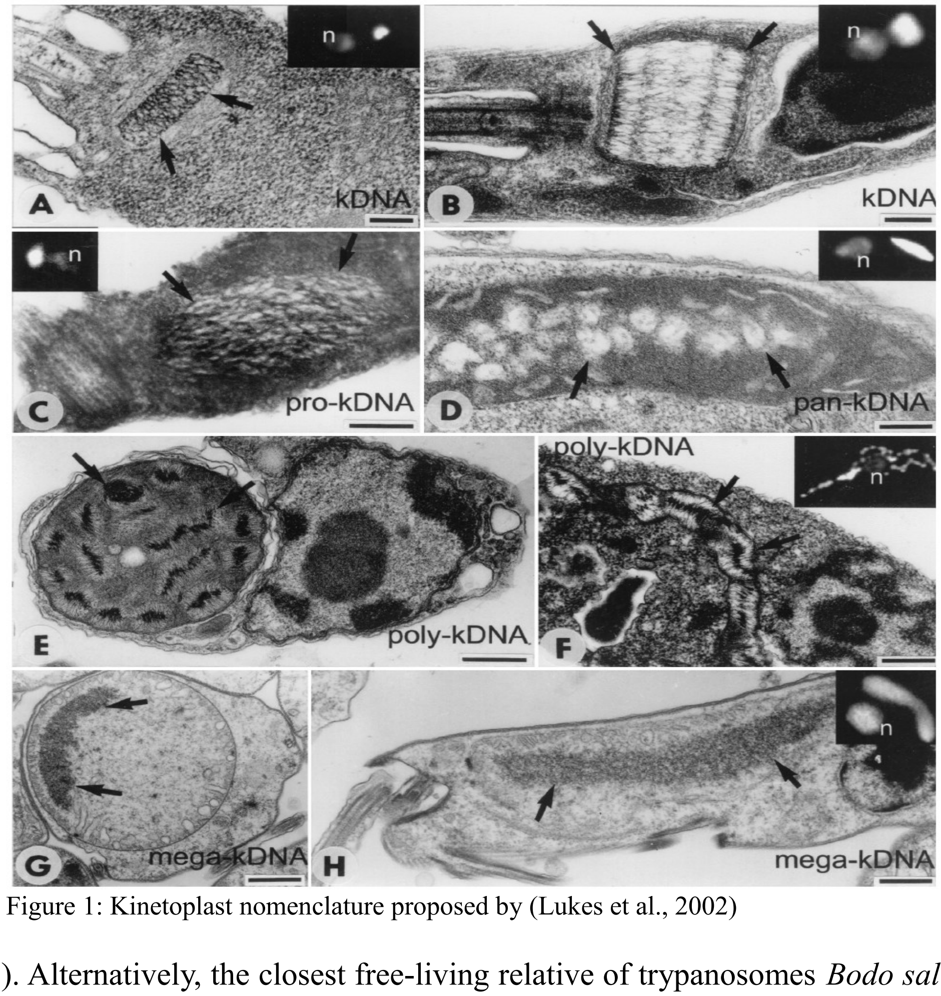


Figure 1: Kinetoplast nomenclature proposed by (Lukes et al., 2002)

overview and proposed nomenclature by (Lukes et al., 2002). In case of *Perkinsela,* pure mitochondria have never been successfully isolated (Dyková et al., 2000). However, there are several TEM pictures available (Fig. 2) (Dyková et al., 2000; 2003). Therefore estimation of kDNA morphological type in *Perkinsela* seems to be possible.

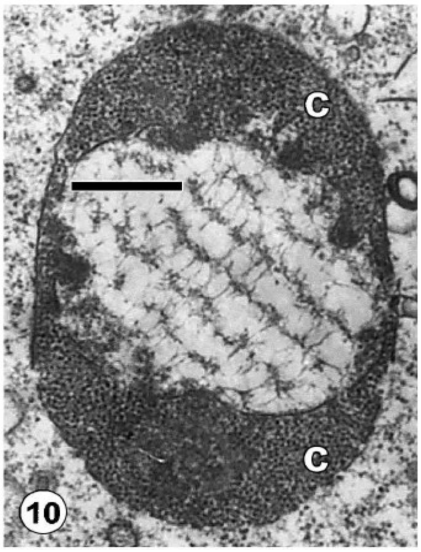


Figure 2: *Perkinsela* with kinetoplast, transmision electron microscopy by (Dyková et al., 2003)

Second, kinetoplast has a small gene content typical for reduced mitochondrial genomes. For example, in trypanosomatids the maxi-circle encodes just two rRNA subunits and 18 protein genes (Simpson et al., 2000), Nuclear genome also encodes all tRNAs for mitochondrial translation, mitochondrial RNA polymerase and other genes, which are transported into the mitochondria from the cytoplasm (Campbell et al., 2003). For complete maps of several maxi-circle sequences see Fig. 3.

Last, but definitely not least, specific feature of mitochondria in kinetoplastids is that a portion of mitochondria-encoded transcript undergoes editing to various extent by modification of uridine content.

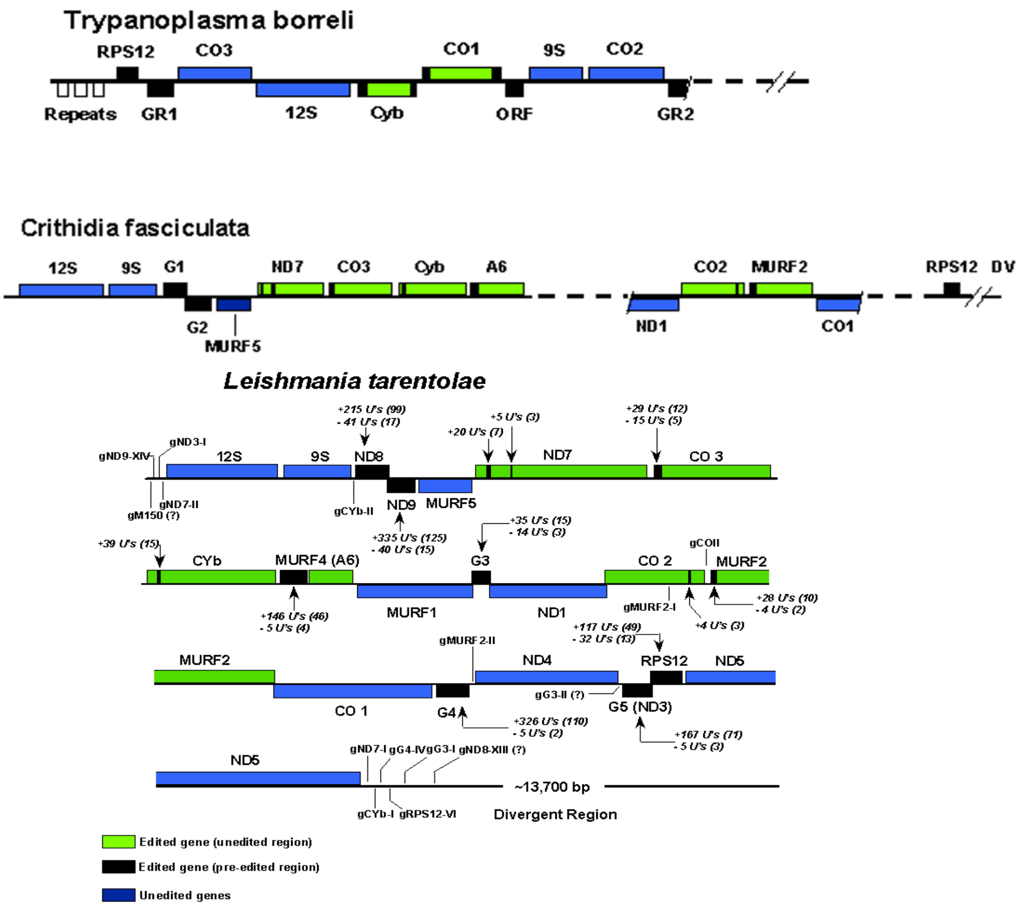


Figure 3: Example of maxicircle genes. (Composed on 27th March 2013 of publicly available content at: http://dna.kdna.ucla.edu/trypanosome/database.html: The RNA editing website by L. Simpson)

Uridine RNA editing in kinetoplastids

Primary role of RNA editing is to change mitochondrial transcripts by uridine insertion and deletion so that they encode functional template for translation. This U insertion/deletion editing is based on partial complementarity of transcript and gRNA molecules which form a hetero-duplex, allowing G to U pairing (Simpson et al., 2000). Illustration of the model can be found on Fig. 4. The mechanism itself is handled by dynamic 20S editosome complex, that catalyzes cleavage, insertion/ deletion and ligation step by step in 3' to 5' direction (Simpson et al., 2000). Some transcripts, such as ND9 in *L.tarentolae* are edited beyond recognition. In a single gene 340 insertions and 40 deletions creating 196 codons over the whole length, mediated by several partially overlapping gRNAs, have been observed (Simpson et al., 2000). Except for these pan-edited proteins, most proteins are edited in short regions at close proximity to the 3' end. There is one case of alternative editing described in *T. brucei*: *cox3* gene edited with two different gRNAs led to the creation of an alternative mRNA with a slightly altered sequence, resulting in two distinct protein products (Ochsenreiter and Hajduk, 2006; Ochsenreiter et al., 2008). The extent of editing in various kinetoplastids is shown on Fig. 3. Concerning this work, it is also important to note two additional features of mitochondrial editing in kietoplastids. First, there is usually a large amount of partially edited sequences in the the mitochondrion, with the last edited position carrying short junction region different both from pre-edited RNA and fully-edited RNA (Koslowsky et al., 1991). Second, generation of misedited molecules, that are edited by inappropriate gRNAs has been observed (Sturm et al., 1991).

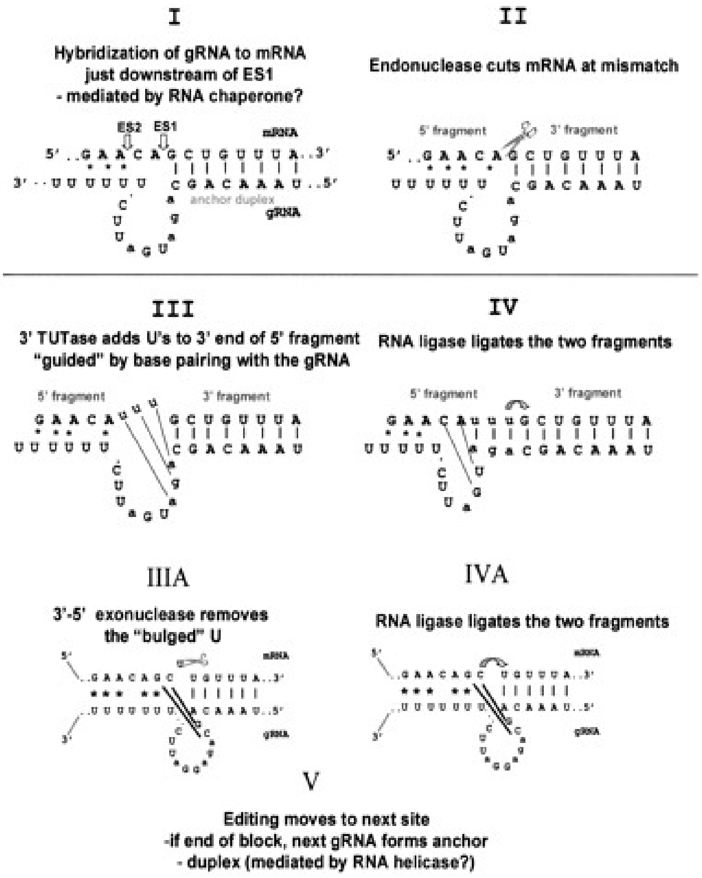

I
Hybridization of gRNA to mRNA just downstream of ES1 - mediated by RNA chaperone?
ES2
ES1
mRNA
anchor duplex
gRNA
II
Endonuclease cuts mRNA at mismatch
5' fragment
3' fragment
III
3' TUTase adds U's to 3' end of 5' fragment "guided" by base pairing with the gRNA
5' fragment
3' fragment
IV
RNA ligase ligates the two fragments
5' fragment
3' fragment
IIIA
3'-5' exonuclease removes the "bulged" U
mRNA
gRNA
IVA
RNA ligase ligates the two fragments
mRNA
gRNA
V
Editing moves to next site -if end of block, next gRNA forms anchor - duplex (mediated by RNA helicase?)

Figure 4: Enzyme cascade model of RNAediting (Simpson, 2003).

What is *Perkinsela*?

This organism is an obligate endosymbiont of *Neoparamoeba pemaquidensis* (Page 1987). The alternative name *"Ichthyobodo*-related organism" originates in Morrison et al., 2005, and is unfortunately not the only name that can be found in the literature. Former name *Perkinsiella* was initially used for more than one organism and currently is already established for the leaf-hoppers. For this reason the name *Perkinsela amoebae* (Hollande 1980) has been suggested for the single organism (Dyková et al., 2008) and for remaining endosymbionts, including strain CCAP1560/4, the name *Perkinsela*-like organisms is used. The unique species names are still not well defined.

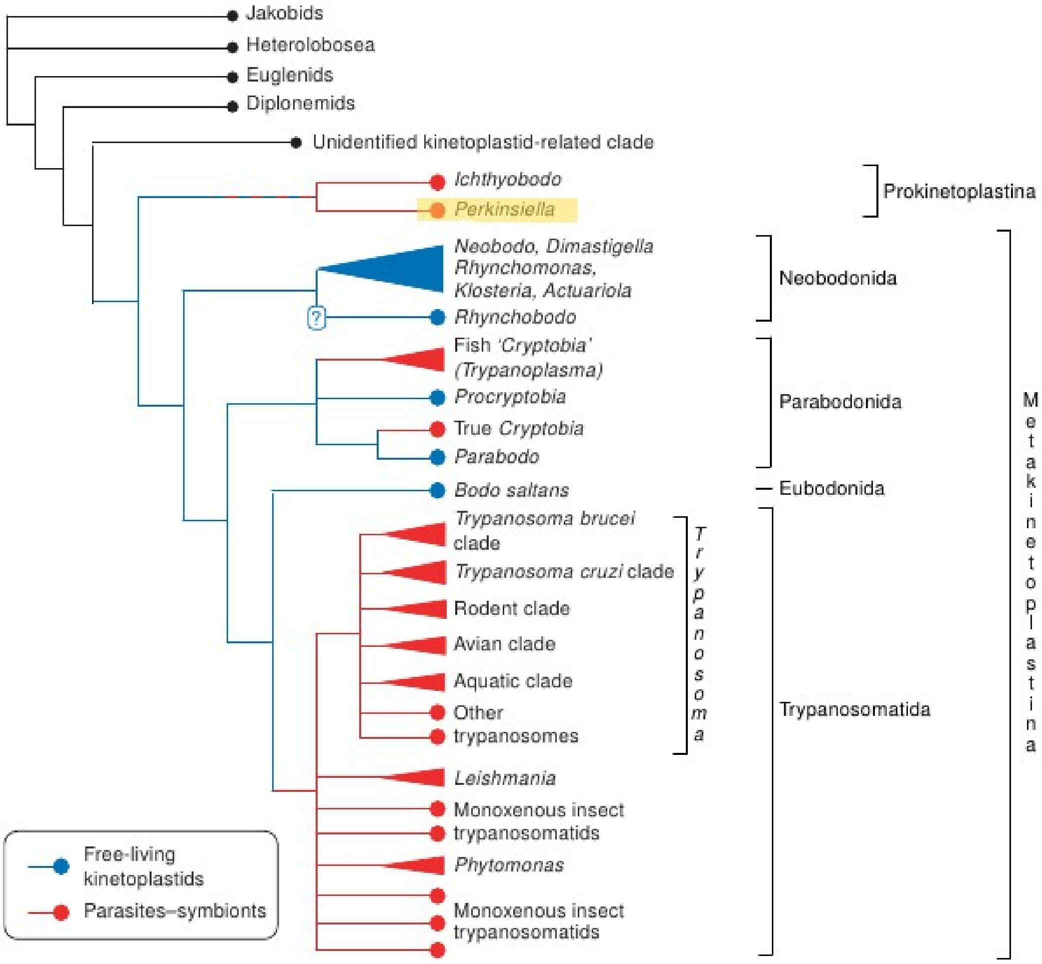


Figure 5: Evolution of kinetoplastid flagellates (Simpson et al., 2006).

Broad phylogenetic context with *Perkinsiella* highlighted is showed on Fig 5. Phylogenetic relationship among closely related amoebae isolates and their endosymbionts have already been investigated (see Fig. 6) (Morrison et al., 2005; (Dyková et al., 2008). The latest name used was *Ichthyobodo*-related organism, that time for CCAP1560/4 (Tanifuji et al., 2011). I have decided to use the naming *Perkinsela* for the purpose of this thesis, referring only to the endosymbiont of *N. pemaquidensis, isolate* CCAP1560/4 (for simplicity). For other closely related organisms (Morrison et al., 2005, Dyková et al., 2008) the full name will be used.

Superficial genome investigation results for *Perkinsela* genome has already been published (Tanifuji et al., 2011). Pulse-field gel electrophoresis and SL gene RNA probe hybridization showed the presence of (at least) 11 SL gene containing chromosomes of total size of 25Mbp. Additionally, C+G content from several sequenced housekeeping genes does not show any C+G loss trend (often encountered in endosymbiotic organisms) when compared with *Bodo saltans* (Tanifuji et al., 2011).

Neoparamoeba: the host

Genus *Neoparamoeba* is known mostly as a causative agent of amoebic gill disease (AGD) which is currently the most decimating disease of sea-cultured fish. Although no attempt to evoke AGD by cultured amoebae was ever successful (Munday et al., 2002, Morrison et al., 2005), *Neoparamoeba* is still taken as an AGD causative agent. However, physiological aspects of the disease development have not been exhaustively described yet. The role of bacteria, which may take a part in the disease development has been tested by co-infection resulting in significantly more infections and more severe damage to the tissue (Embar-Gopinath et al., 2006). Mitchell and Rodger in their review concluded that the most likely hypothesis, due to the observations of AGD without apparent bacterial overgrowth, is that *Neoparamoeba* evolved as an opportunistic parasite and requires gill tissue damage to successfully invade fish (Mitchell and Rodger, 2011).

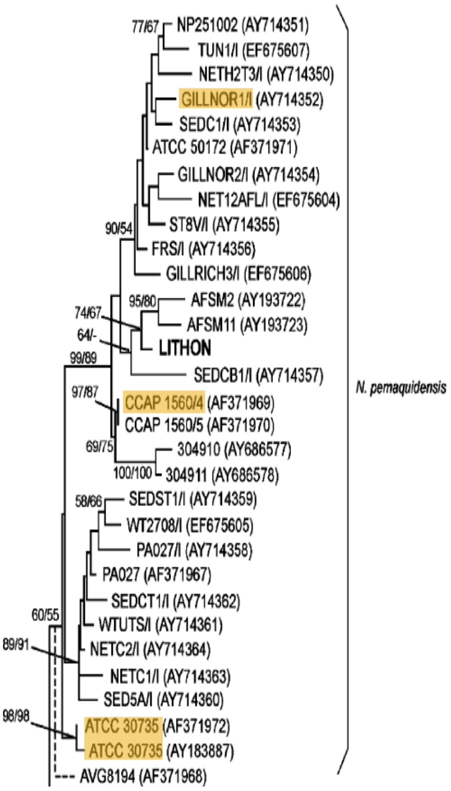


Figure 6: Endosymbionts of *N.pemaquidensis* with highlighted isolates involved in the genomic project so far (Dyková et al., 2008)

Since *Ichthyobodo* is also a fish ectoparasitic organism, the initiation of the endosymbiotic relationship seems to be similar to the mitochondrial or chloroplastic endosymbiosis. What remains elusive is the purpose of this novel *Perkinsela-Neoparamoeba* relationship. It has been shown that neither partner of this pair can survive alone (Dyková et al., 2000). The genome project might shed some light on the importance of *Perkinsela* for *Neoparamoeba* and *vice versa*.

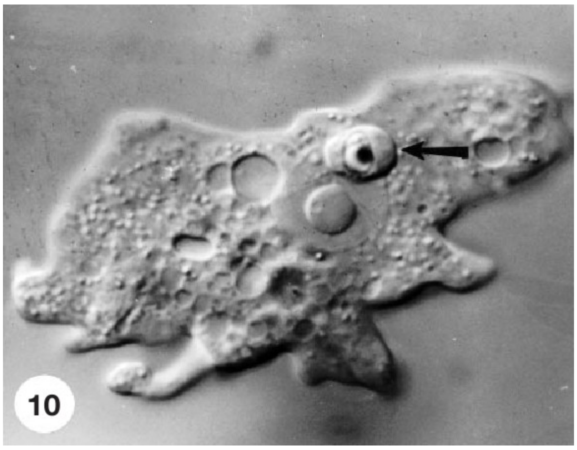


Figure 7: Phase contrast picture of *Neoparamoeba* (Dyková et al., 2000)

The overall cell structure of *Perkinsela* is interesting for several reasons. Each *Neoparamoeba* seems to harbor a single *Perkinsela* (Fig.7), with no more than two endosymbionts being observed so far (Fig.7, 8) (Dyková et al., 2000). The division pattern seems to fit the following duplication order: *Perkinsela* nucleus, kDNA, *Perkinsela*, *Neoparamoeba*. There is one picture of two nuclei, situated on the opposite poles of duplicated kinetoplast during its separation (Dyková et al., 2000), suggesting kDNA division being coordinated with the last phase of *Perkinsela* division. Fig. 9 shows DAPI stained *Neoparameoba* captured by Eva Horáková from our laboratory showing much the same results as found in literature (Tanifuji et al., 2011; (Dyková et al., 2000).

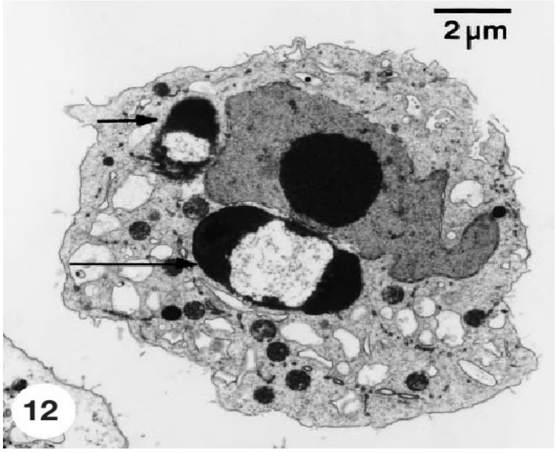


Figure8: TEM picture of *Neoparamoeba* (Dyková et al., 2000)

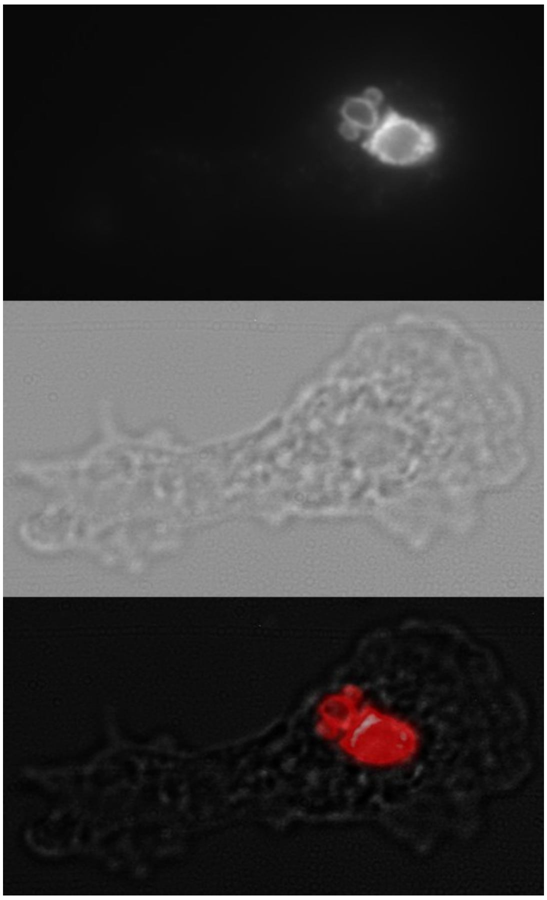


Figure 9: DAPI stained *Neoparamoeba*: 1st DAPI stained DNA, 2nd light microscope photo (E. Horáková 11/9/12 unpublished), 3rd photo merged noise reduced, colorized image of DAPI layer and the light microscope image.

Sequencing

Drop in the cost of recent, high-throughput, sequencing technologies has allowed a revolution in sequencing to occur. The most commonly used platforms, 454 (Roche) and Illumina along with various mass sequencing projects (i.e.: 1000 genomes project, aiming for a representative set of human genomes) have been at the forefront in the latest years of genomics. Nowadays, several scientific fields are extending the explanatory potential of so-called next-gen sequencing data by utilizing bioinformatics approaches together with the continual improvement of already established methods. In this part of my thesis, I am going to describe approaches which are tightly bound to my project.

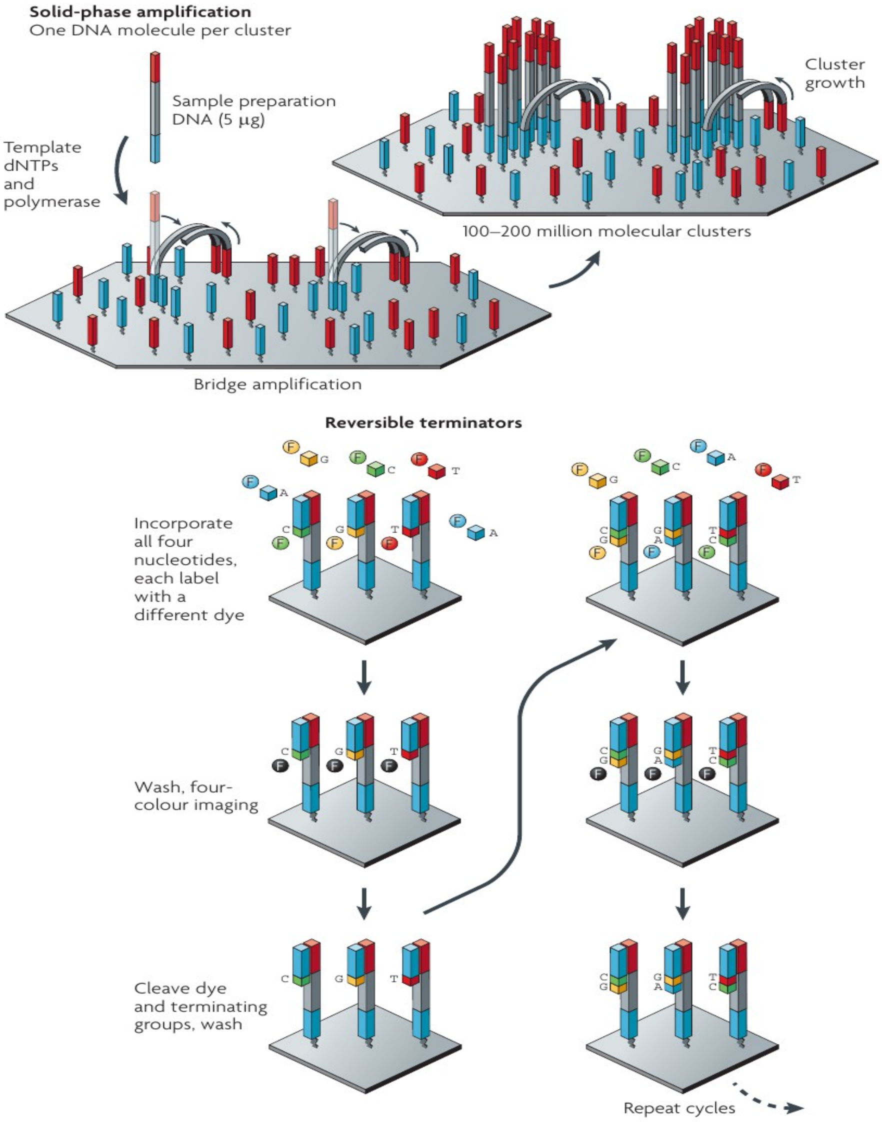

Solid-phase amplification
One DNA molecule per cluster
Sample preparation DNA (5 μg)
Template dNTPs and polymerase
Cluster growth
100–200 million molecular clusters
Bridge amplification
Reversible terminators
Incorporate all four nucleotides, each label with a different dye
Wash, four-colour imaging
Cleave dye and terminating groups, wash
Repeat cycles
F
A
C
G
T

Figure 10: Illumina seuencing (Metzker, 2009)

Illumina is the only sequencing platform that has been involved in our project so far. This approach has the advantage of low cost per base resulting in much higher coverage, compared to the alternatives, and low indel error rate (Metzker, 2009). The main drawback of this approach is relatively short read length up to 100bp (Metzker, 2009). Hoverer, in some recent kits the length have been extended to 250bp.

Input for this machine is fragmented DNA on which adapter sequences are added during the step called DNA library preparation. The library is then melted and loaded on the flow cell coated with adapter-complementary oligo-nucleotides. The cell then undergoes cyclic set of flushes in order to synthesize the reverse strand primed by the hybridized oligo-nucleotides and so covalently bound to the surface. The new strand hybridizes with the oligo-nucleotide again and forms a "bridge". Repetitions of the cycle creates a cluster of such identical bridges. All the DNA bridges than undergo a set of cleavage reactions and only single stranded molecules of identical orientation remain (Kircher and Kelso, 2010).

The next set of cycles refers to the reversible terminator sequencing itself. After the sequencing primer hybridizes with the adapter, modified polymerase adds a single fluorophore-terminator modified nucleotide to each DNA strand. Lasers then scan the cell with a nucleotide specific settings and the cycle repeats adding the next nucleotide, usually up to 100 nucleotides (Kircher and Kelso, 2010). Summary illustration for Illumina sequencing is situated above, Fig 10.

Genome and transcriptome reconstruction: *de novo* from Illumina data

Countless amount of sequencing data, raw reads, has to assembled into large pieces or whole chromosomes in order to provide comprehensible genetic information. To do so, several software solutions have been introduced. These, so called assemblers, differs both in performance and quality. Some assemblers are specialized in processing output from single platforms, some have separate modules to cover different sequencing platforms. Here, I will briefly describe assembly of short Illumina reads from scratch. Such reads are short (100bp or slightly less, if quality trimmed), and it is wise to assemble them with mate-pair information. Paired reads are two reads that originate from the opposite ends of the same insert (bridge) and may be separated by a gap of approximately known length.

First, every read that shares a part of the sequence with another read has to be linked with such a read. In order to prevent aligning all reads against all, the system of k-mers has been introduced. K-mer system of indexing uses short fragments of the read to look for sequences that share the same k-mer (Schatz et al., 2010), and only those are aligned. Alternatively, k-mers themselves often serve for further processing.

Second, the graph of overlapping units is constructed: the graph can be imagined as a network of sequence nodes and connections showing possible extensions. Here it's necessary to highlight the main issue of all recent high throughput sequencing data assemblies, repeats. Connecting the nodes through the true path on the graph is the main challenge for all assemblers. Here, the main aim is to identify as many true connections as possible, without creating an artificial extension. This can be achieved by combination of several strategies: using each k-mer only once, removing underrepresented paths or splitting them, removal of short dead-ends, validation by whole reads and preferring paths consistent with mate-pair information. If all linked nodes share a perfect unbiased path, nodes can be merged resulting in contiguous sequences, contigs. These are usually sufficient for some applications such as transcriptome assembly.

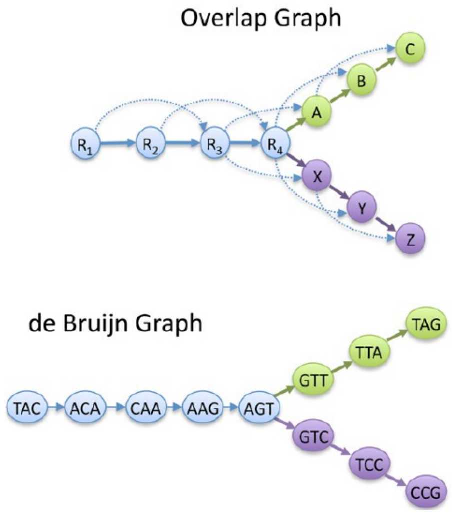


Figure 11: Assembly graphs (Schatz et al., 2010)

Otherwise, a scaffolding step is necessary, in which individual contigs are linked together utilizing mate-pair information. Adjacent contigs are those that share at least two (or significant amount) of connections through matches of mate-pair reads. Further improvements, aiming for the reconstruction of intact chromosomes, non-computational approaches such as restriction maps or hybridization can be used to connect remaining contigs.

Sequence similarity search software

BLAST is one of the most used bioinformatic tools for searching functionally and evolutionary related sequences. I have used BLAST in this particular case for the identification of mitochondrial rRNAs in the *Perkinsela* assembly. As a query for this search I have used all publicly available sequences for 9S and 12S mito-ribosomal RNAs which originated from kinetoplastids. In addition, BLAST was also used to confirm HMMER results.

HMMER software applies hidden Markov model based algorithm to search for query which is in my case an alignment (constructed with MUSCLE) of all available protein sequences from kinetoplastids for each mitochondrial protein. The alignment is turned by the software into a probabilistic "hidden" sequence which then serves as a query for searching translated DNA database in a BLAST-like manner (Eddy, 1998).

MUSCLE aligner

MUSCLE proved itself as an invaluable asset to our efforts. This progressive multiple sequence alignment algorithm is somewhat similar to the widespread CLUSTALW which it overcomes in speed and better alignment of the edges, especially if many sequences are aligned together. An overview of MUSCLE data handling can be found on Fig 12 (Edgar, 2004a).

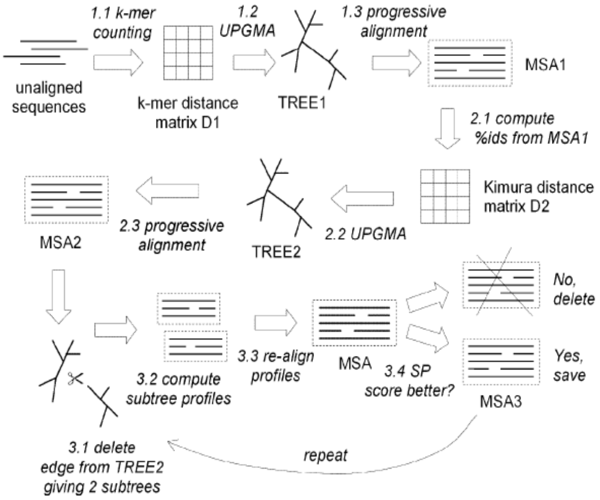

1.1 k-mer counting
1.2 UPGMA
1.3 progressive alignment
unaligned sequences
k-mer distance matrix D1
TREE1
MSA1
2.1 compute %ids from MSA1
Kimura distance matrix D2
2.2 UPGMA
TREE2
2.3 progressive alignment
MSA2
No, delete
Yes, save
3.3 re-align profiles
MSA
3.4 SP score better?
MSA3
3.2 compute subtree profiles
3.1 delete edge from TREE2 giving 2 subtrees
repeat

Fig.12: MUSCLE work-flow(Edgar, 2004b)

Read mapping

The term mapping is used in bioinformatics in a specific manner. It means creating large scale alignments by pairwise alignment of short reads with a corresponding longer template. These mapped reads can serve to characterize polymorphisms, repeats, errors, coverage and presumably other types of information "hidden" in sequencing data. If RNA reads are mapped, transcription pattern can be observed. Splicing and other post-transcriptional modifications like RNA editing can be captured if the settings and quality cutoffs allow reads from such RNAs to be present in the results. Otherwise this kind of information is either lost or overlooked.

The way all mapping software works is in principle, somehow similar to the way in which BLAST works. Each read is broken into several overlapping seeds of several nucleotides. Basically, the frame of defined seed length is shifted by a certain number of nucleotides creating a seed every time. When seeds building is finished, they are compared with query. Seeds that do perfectly match the query are accepted as hits and are passed to the aligner that attempts to construct the best alignment from any of the valid seeds

Most of the read mappers have two alignment modes. First, the local mode, where trimming of read ends is allowed. The mode has an advantage of recognition of poor quality for unfiltered reads data and is useful for read quantification and single nucleotide polymorphism identification. Second mode is called end-to-end mode where the whole read has to be aligned and pass certain score threshold to be accepted as valid. Such approach has an advantage of preserving information caused be gene duplication and many types of post-transcriptional processing when used on RNA data.

Read maps carry huge amount of information, inaccessible directly. For this reason, two basic solutions are at our disposal. There are browsers, generating graphically friendly and interactive environment for filtering and browsing. Otherwise simplified text outputs are produced for investigation by eye or by downstream applications.

Haplotype inference and read clustering

Another way of accessing diversity in heterogeneous read maps, is to estimate the population structure. ShoRAH, the software we have decided to use for our data, has been initially developed to analyze heterogeneous sequencing results of mixed microbial samples. It also reports good performance with simulated Illumina data (Zaghordi et al., 2011). The whole work-flow implements alignment, error correction, local haplotype reconstruction, global haplotype reconstruction and frequency estimation. Since we already had precise alignment and Sho-RAH is not constructed for global reconstruction of so extensively indel divergent data, We have used only error correction and local haplotype reconstruction (module diri_sampler). This module employs model-based probabilistic clustering algorithm to correct errors, infer both haplotypes and frequencies and finally estimates reconstruction quality using Bayesian approach in order to compute posterior probability distribution of parameters. Simplified picture of local haplotype reconstruction is showed on Fig.13.

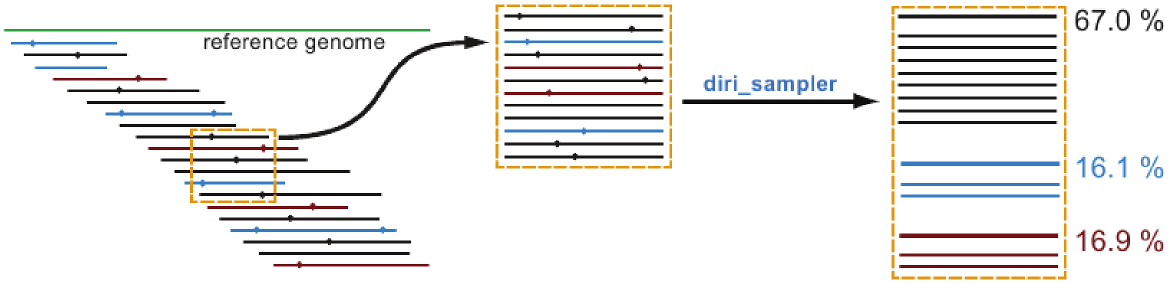


Figure 13: Local haplotype reconstruction by ShoRAH (Zaghordi et al., 2011).

The main variation in the results dwells in tuning two parameters: iteration number and parameter α. By utilizing this clustering step, we are getting quantitative representation of the alignment, that can be directly solved as it represents virtually error free editing with underrepresented versions distributed among the closest clusters.

Material and methods

All the data were processed with the UNIX versions of the software mentioned here. Complete list with versions and sources can be found in the supplementary data (S2) . Additionally, simple notepad and common office applications have been used. The initial processing design is shown on Fig 14.

Initial data

The initial dataset is composes of Illumina sequencing data for 2 isolates (4 libraries) and the corresponding assemblies (see Table 1). The most important libraries are CCAP1560/4 DNA (334.2M reads) and total RNA (275.6M reads), all composed of 100nt paired-end reads. SL-RNA library was prepared by *Perkinsela* spliced-leader primer. DNA library has been enriched for *Perkinsela* and host nuclear DNA using density gradient centrifugation.

Isolate	Type	library
ATCC 50172 (seawater, USA)	RNA	total RNA
CCAP 1560/4 (seawater, GB)	DNA	DNA fraction
	RNA	total RNA
	RNA	SL fraction

Table 1: Initial sequencing data (G. Tanifuji unpublished)

Gene identification with BLAST and HMMER

The “mixed” assembly of the host and endosymbiont represents some challenges in the identification of *Perkinsela* mitochondrial contigs. First, one has to identify as much mitochondrial content as possible by sequence similarity search. Second, phylogenetic analysis or similarity to closely related organisms should be used to distinguish between contigs originating from the host and those originating from the endosymbiont. I have employed HMMER software to search for typical kinetoplastid mitochondrion-encoded proteins and BLAST software in order to search for mitochondrial rRNA subunits of kinetoplatids. Lastly, HMMER search has been performed to look for proteins of electron transport chain complexes I, II, III, IV and V.

Resulting output tables have been merged and browsed in order to find valid hits. Sequences of hits were retained and confirmed using BLASTx search in whole NCBI nr database, that may have eventually discovered additional important regions. In order to distinguish between host and endosymbiont, three different approaches were used in the following order: co-localization in a contig with a non-ambiguous *Perkinsela* mitochondrial gene, E-values difference ($\Delta E>10^{-2}$ AND $\Delta\%^{identity}>10$, between valid top hits for ameoebozoa and kinetoplastida), and phylogenetic analysis (maximum likelihood with RAxML). Similarly, absence/presence of some nuclear-encoded proteins functionally linked with the mitochondrion-encoded proteins or RNA editing has been performed in the same way.

DNA seq mapping as assembly quality control

Perkinsela kDNA (putative mitochondrial) contigs sharing at least one valid hit have been assembled together by CLC Workbench assembly function and longest consensi have been retrained. Coherency of the retained contigs have been validated through mapping of raw DNA reads (SMALT, local alignment, 95% identity cutoff). As coherent contigs have been taken those contigs that had coverage of overlapping paired reads at each position, otherwise contig trimming or splitting has been introduced. Another candidates for kDNA contigs have been identified by BLAST search for typical terminal repetitive sequences in the first group of contigs.

RNA mapping as an indicator of transcription and RNA editing

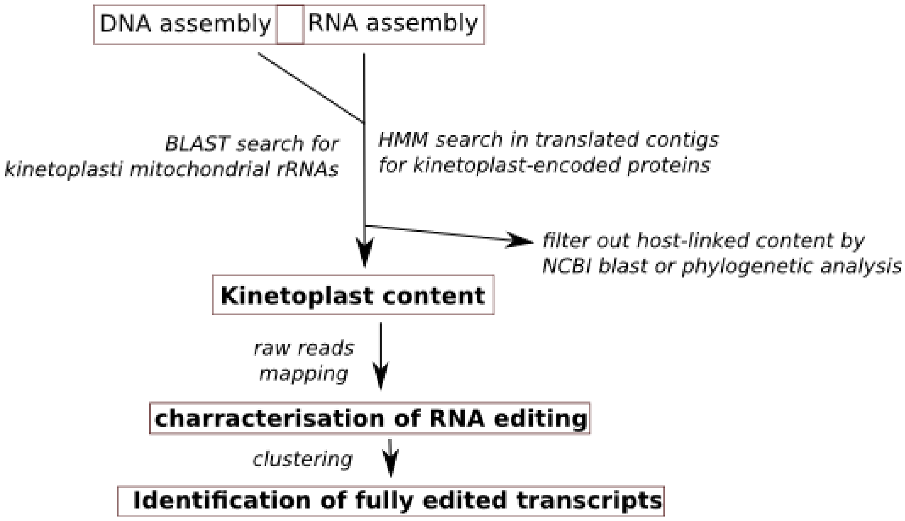
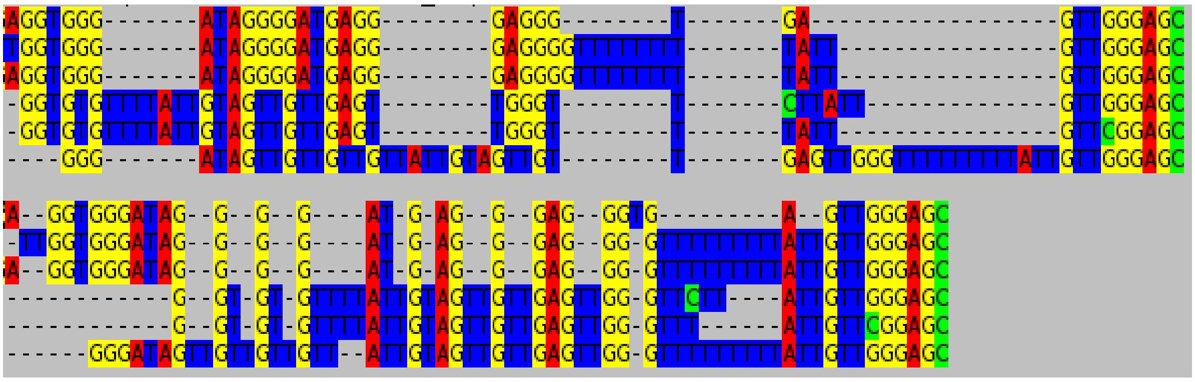

DNA assembly
RNA assembly
BLAST search for
kinetoplasti mitochondrial rRNAs
HMM search in translated contigs
for kinetoplast-encoded proteins
filter out host-linked content by
NCBI blast or phylogenetic analysis
Kinetoplast content
raw reads
mapping
charracterisation of RNA editing
clustering
Identification of fully edited transcripts

Figure 14: Initial work-flow design.

Transcription of all coherent kDNA and kDNA candidate contigs has been mapped in order to characterize their transcripts (SMALT, local alignment, 95% identity cutoff), and to reveal U-indel RNA editing (Bowtie2 and SMALT, various sensitivity options, see below). The quality of edited reads by SMALT appeared to be low. Since SMALT does not allow user to tune the Smith-Waterman aligning settings, Bowtie2 software has been used in order to study RNA editing phenomenon. After a set of testing and optimization runs, following settings have been used for mapping of putative 9S rRNA and cox1 transcripts: --end-to-end --very-sensitive -rdg 2,1 -rfg 2,1 -mp 3,4. This setting means low penalties for introduction of insertions and deletions and lowered penalties for mismatches. The main drawback of this setting is the fact, that the aligner still introduces artificially long insertions, followed by deletions and cumulative false insertions. Since there is currently no mapping software that allow nucleotide-specific gap penalties, we have decided to modify Bowtie 2 software, in the near the future. Without the modified software, a time consuming approach of manual correction of 9S rRNA and cox1 read alignments has been chosen to make the alignment of thousands of reads to accommodate only U-containing insertions and deletions, as implied on Fig 15.



```
AGGTGGG-------ATAGGGGATGAGG--------GAGGG--------T-------GA------------------GTTGGGAGC
TGGTGGG-------ATAGGGGATGAGG--------GAGGGGTTTTTTTT-------TATT----------------GTTGGGAGC
AGGTGGG-------ATAGGGGATGAGG--------GAGGGGTTTTTTTT-------TATT----------------GTTGGGAGC
-GGTGTGTTTTATTGTAGTTGTTGAGT--------TGGGT--------T-------CTTATT--------------GTTGGGAGC
-GGTGTGTTTTATTGTAGTTGTTGAGT--------TGGGT--------T-------TATT----------------GTTCGGAGC
----GGG-------ATAGTTGTTGTTGTTATTGTAGTTGT--------T-------GAGTTGGGTTTTTTTTTATTGTTGGGAGC

A--GGTGGGATAG--G--G--G----AT-G-AG--G--GAG--GGTG---------A--GTTGGGAGC
-TTGGTGGGATAG--G--G--G----AT-G-AG--G--GAG--GG-GTTTTTTTTTATTGTTGGGAGC
A--GGTGGGATAG--G--G--G----AT-G-AG--G--GAG--GG-GTTTTTTTTTATTGTTGGGAGC
------------G--GT-GT-GTTTTATTGTAGTTGTTGAGTTGG-GTTCTT----ATTGTTGGGAGC
------------G--GT-GT-GTTTTATTGTAGTTGTTGAGTTGG-GTTT------ATTGTTCGGAGC
------GGGATAGTTGTTGTTGTT--ATTGTAGTTGTTGAGTTGG-GTTTTTTTTTATTGTTGGGAGC
```

Figure 15: Comparison of manually uncorrected (upper block) and corrected (lower block) alignment.

Analyzing RNA editing by clustered mapping results.

Since RNA mapping results can consist of thousands of reads, it is not an easy task to draw conclusions from such data. This obstacle has been challenged by standard clustering approach, since no editing-specific software is available (and we are again thinking about introducing such solution). ShoRAH (diri_sampler module), clustering software has been used on alignments of reads fully covering the same regions with the following settings: -j 10000 -a 10.0 -t 10000 -K 100. Such settings produce sufficiently representative and reasonable amount of clusters. After removal of clusters with extremely low abundance (average number of reads<0.01%) and those with low posterior probability (P<90%), usually a set of tens of reads remains for explanation.

Results

As new results have been obtained in the project, the initial work-flow had to be updated. I have included a simplified schematics of current status of the project for better orientation of the reader (Fig. 16).

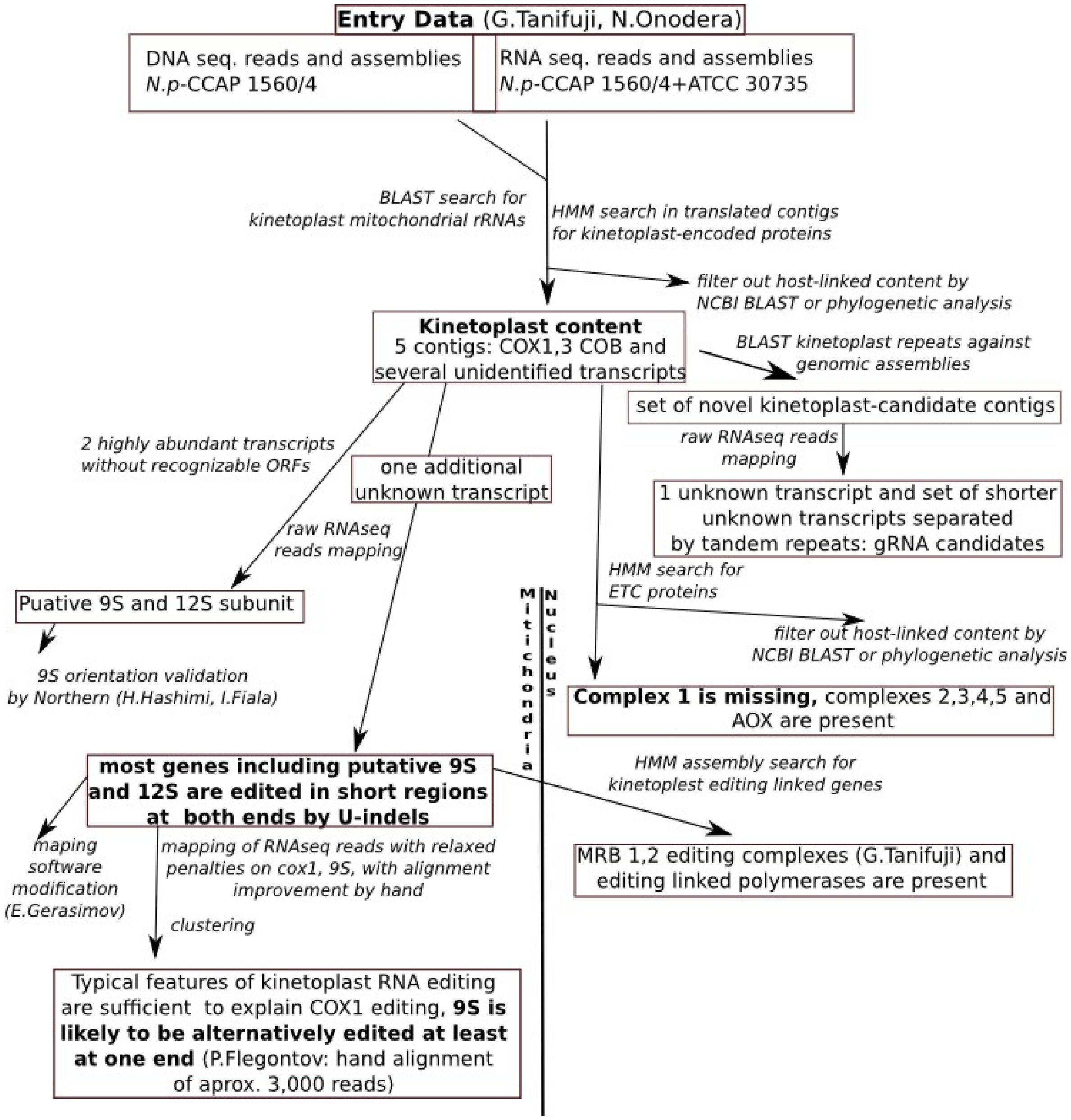

Entry Data (G.Tanifuji, N.Onodera)
DNA seq. reads and assemblies N.p-CCAP 1560/4
RNA seq. reads and assemblies N.p-CCAP 1560/4+ATCC 30735
BLAST search for kinetoplast mitochondrial rRNAs
HMM search in translated contigs for kinetoplast-encoded proteins
filter out host-linked content by NCBI BLAST or phylogenetic analysis
Kinetoplast content 5 contigs: COX1,3 COB and several unidentified transcripts
BLAST kinetoplast repeats against genomic assemblies
set of novel kinetoplast-candidate contigs
raw RNAseq reads mapping
1 unknown transcript and set of shorter unknown transcripts separated by tandem repeats: gRNA candidates
2 highly abundant transcripts without recognizable ORFs
one additional unknown transcript
raw RNAseq reads mapping
Puative 9S and 12S subunit
9S orientation validation by Northern (H.Hashimi, I.Fiala)
Mitichondria
Nucleus
HMM search for ETC proteins
filter out host-linked content by NCBI BLAST or phylogenetic analysis
Complex 1 is missing, complexes 2,3,4,5 and AOX are present
most genes including putative 9S and 12S are edited in short regions at both ends by U-indels
HMM assembly search for kinetoplest editing linked genes
MRB 1,2 editing complexes (G.Tanifuji) and editing linked polymerases are present
maping software modification (E.Gerasimov)
mapping of RNAseq reads with relaxed penalties on cox1, 9S, with alignment improvement by hand
clustering
Typical features of kinetoplast RNA editing are sufficient to explain COX1 editing, 9S is likely to be alternatively edited at least at one end (P.Flegontov: hand alignment of aprox. 3,000 reads)

Figure 16: Brief project overview.

kDNA content

Edges of contigs typically carry repetitive regions, which are, according to their coverage, extremely abundant. Overall DNA coverage of contigs fluctuates around 1,000 for unique regions and up to 100,000 for repetitive regions. Contig overview with transcripts is shown on Fig. 17.

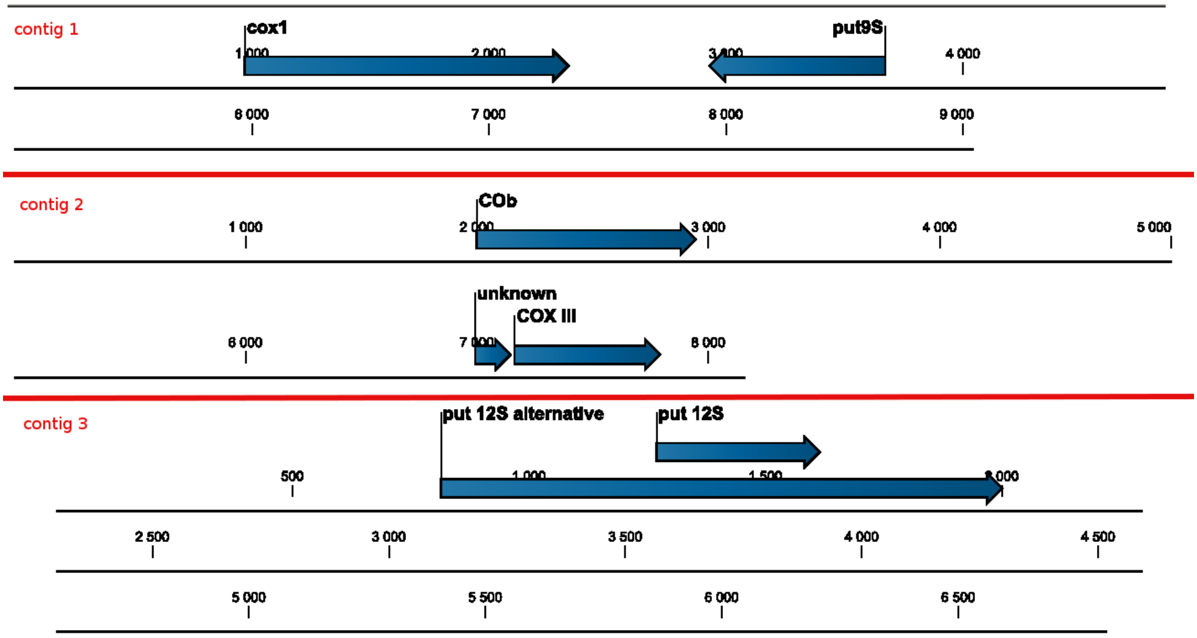


Figure 17: Minimal set of a kDNA contigs with annotated transcripts.

Additional set of 10 contigs share the same or similar repetitive sequence to those found in contigs 1,2, and 3. These kDNA contig candidates have no similarity to any RNA or protein, even though they carry transcribed regions. No transcript originating in these kDNA candidates is edited. However, 7 of these contigs bear ~160nt long transcripts bearing with polymorphic and homogenous parts, and they might correspond to gRNA transcripts, although no similarity to any of kDNA transcripts is recognizable. An example of putative gRNA contig with RNAseq mapping (Bowtie 2, -rdg 2,1 -rfg 2,1) coverage is shown on the Fig. 18.

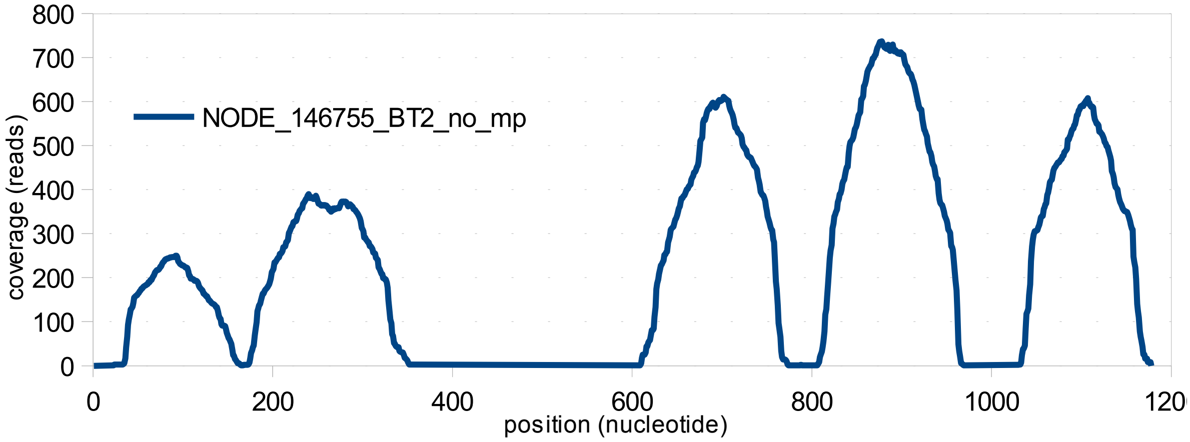


Figure 18: RNA coverage of putative kDNA contig bearing gRNA candidate transcripts.

RNA editing

RNAseq mapping on kDNA contigs shows that extensive U-insertion/deletion editing is present in the mitochondrion of *Perkinsela*. All transcripts of contigs 1, 2 and 3 are edited at both sides of the transcripts in relatively short regions, except for the short unknown transcript adjacent to the *cox3* transcript, that has only one editing site approximately in the middle of this short transcript. Coverage of edited transcripts (contig1 and 2), along with highlighted edited regions, is shown on Fig. 19 where RNAseq coverage maps made with various sensitivity show indirectly the intensity of editing.

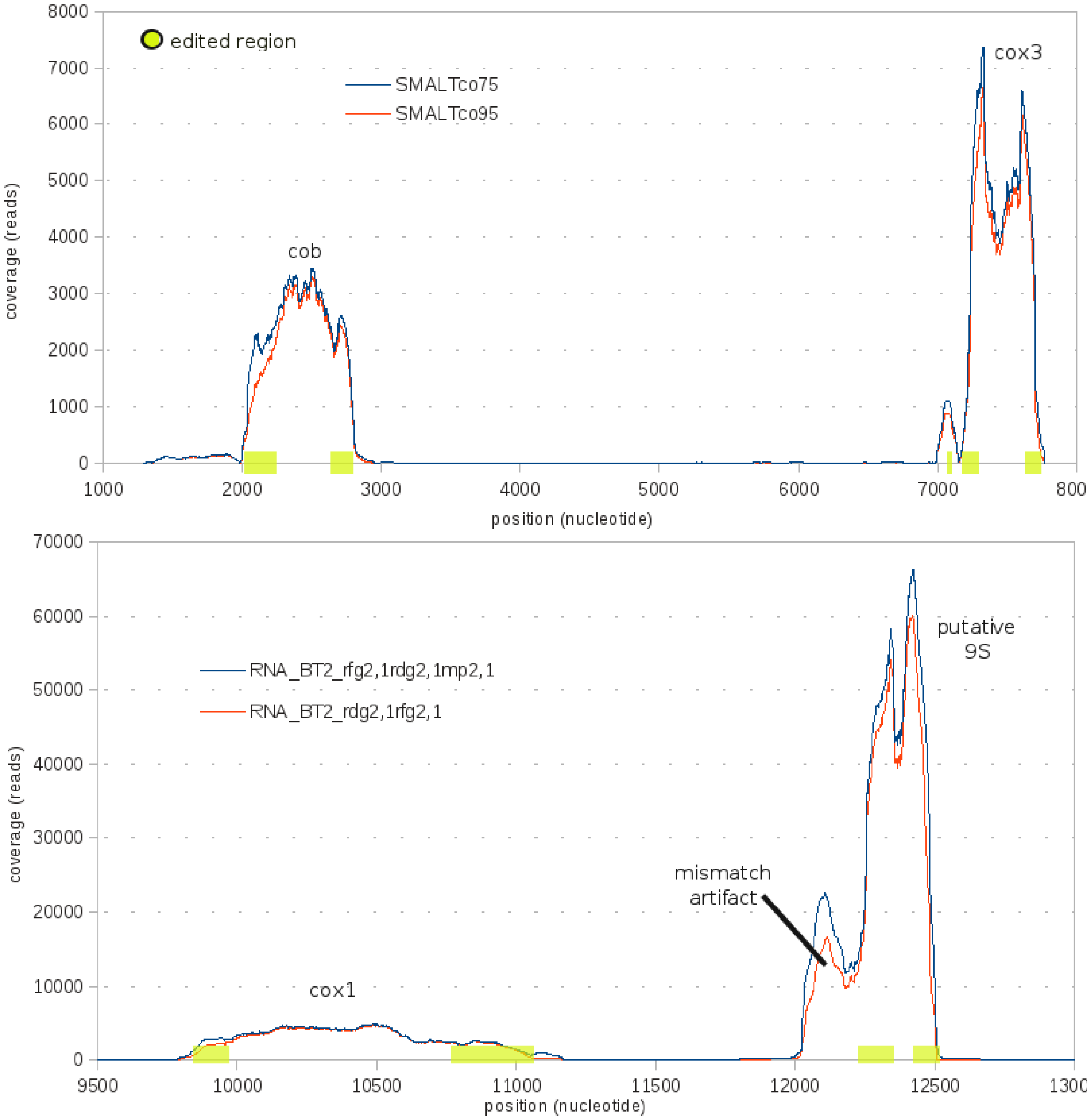


Figure 19: RNAseq coverage of contig 1 and 2 made with different sensitivity settings of SMALT (minimum identity either 75 or 95%) and different gap opening/extension and mismatch penalties for Bowtie2. Highlighted regions are edited with U-indels and the extent of editing is indirectly revealed by the change in coverage with higher mapping sensitivity

Manually finished realignment of all *cox1* RNAseq reads clustered with ShoRAH shows patterns coherent with known features of U-indel editing in kinetoplastids (Fig. 19): 3' to 5' progression, low-abundance misedited sequences (Sturm et al., 1991) and extensive change in uridine content. An edited 5' part of the cox1 alignment is shown on Fig.20, 3' results can be found in supplementary S1a

.

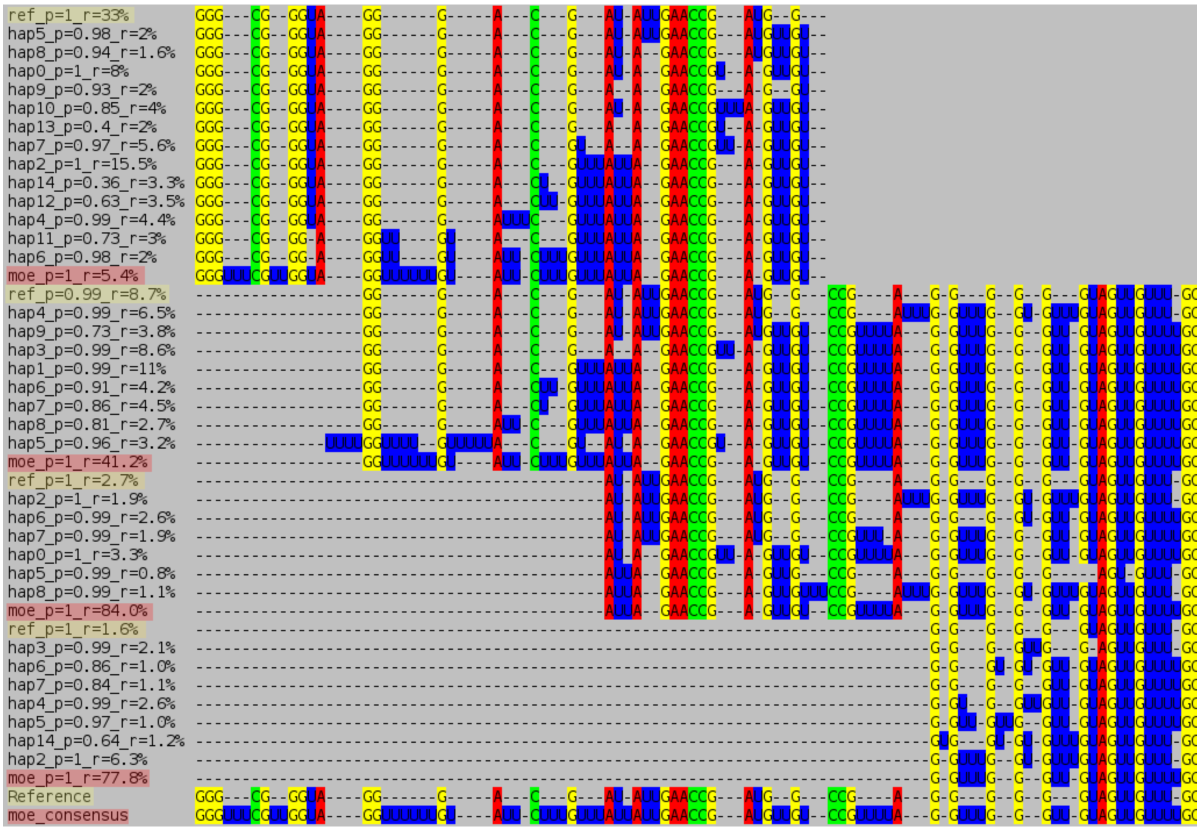


Figure 20: Manually sorted clustering results of RNAseq mapping for the 5' region of *cox1*. *r*=% of average number of reads, p=posterior probability of a 'haplotype'. Hypothetical fully edited sequences are highlighted in red, non-edited reference sequences are highlighted in yellow

RNA editing of putative 9S rRNA

Our investigation of putative 9S rRNA (with up to 60,000 RNAseq coverage) editing in manually realigned 3' and 5' edited regions, resulted in alignments of 6,961 and 4,044 reads, respectively. Editing of this transcript is significantly different from the patten described for *cox1*. First, the pre-edited sequence in the 5' region represents only 0.425% of the sample. Second, the 5' region shows branching pattern of editing (without a dominant fully edited sequence), which made us believe that at least this particular edited region undergoes alternative editing. Slightly trimmed clustered results for the 5' region can be found on Figure 21. Manually sorted clustered results of 3,407 reads alignment spanning the whole 5'' region is showed in the discussion (Fig. 22). The remaining 3' region is shown in supplementary S1b.

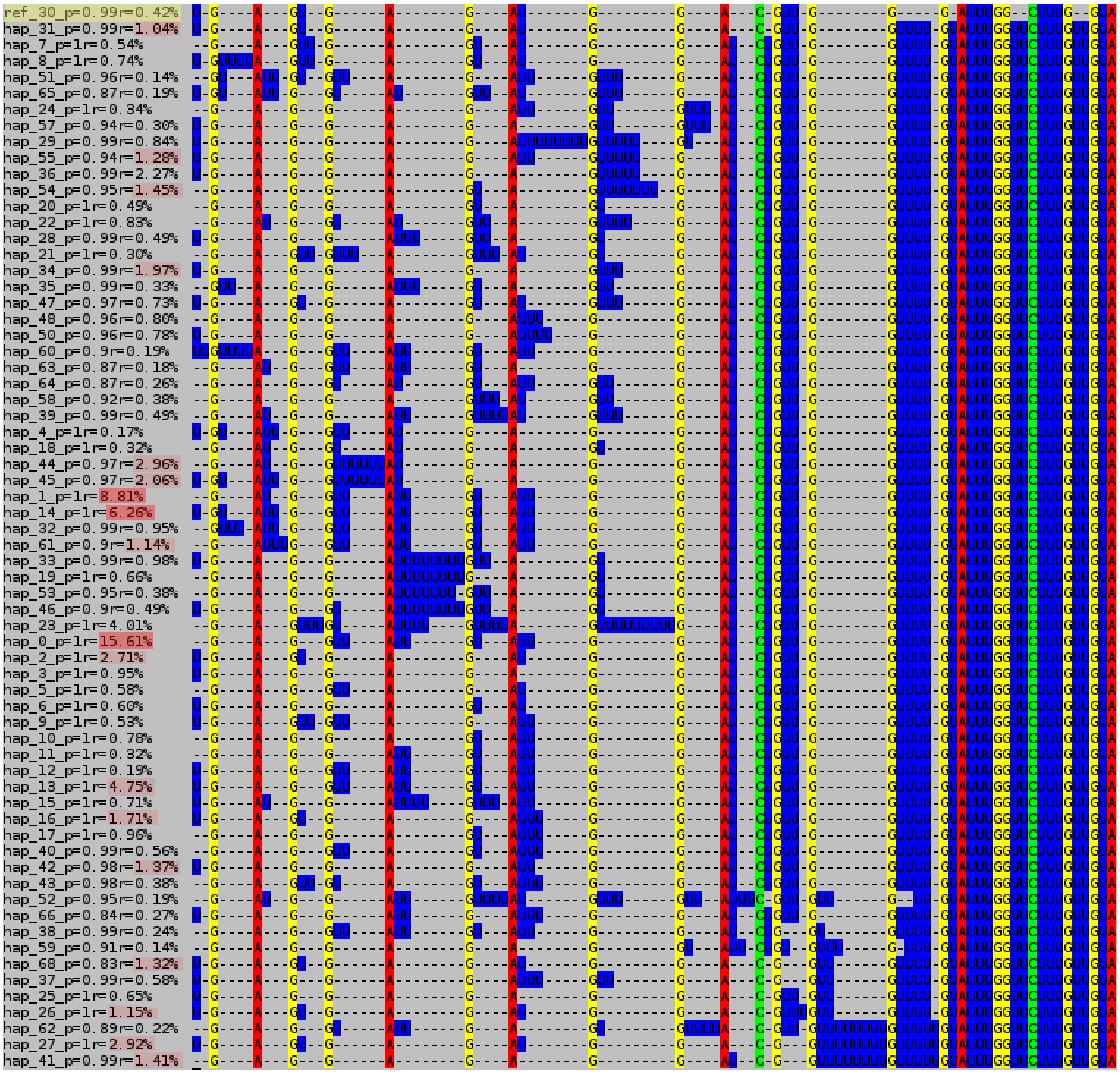


Figure 21: Manually sorted clustering results of RNAseq mapping for the 5' region of putative 9S rRNA transcript. R=% of average number of reads, p=posterior probability of a 'haplotype'. Non-edited reference sequences are highlighted in yellow, abundant sequences are highlighted in red.

Electron transport chain complexes

Lack of mitochondrial genes for NADH oxidoreductase protein subunits in kDNA is concordant with results of HMMER search for nuclear-encoded proteins of the electron transport chain (ETC). This means, that *Perkinsela* completely lost complex I of ETC. Subunits of complexes II, III, IV and V are present, as well as alternative oxidase acquired by kinetoplastids via horizontal gene transfer (See Table 2).

Targetname	**E value**	**score**
alternative oxidase		
AOX	8.6e-65	222.2
complex2		
SDH_Fe-S	5.7e-37	131.1
SDH_flavoprotein	3.5e-282	941.2
Succ_DH_flav_C	1.9e-45	158
complex3		
Rieske(precursor)	4.9e43	151.6
cytC1	6.9e68	233.4
complex4		
CG3-SCO1	5.6e-55	190.8
assembly protein	4e-53	184.6
cox11	1.5e-46	162.6
hemeA-farnesyltransferase	3.1e-43	152.5
complex5		
ATP-synt_C	8.6e-21	77.8
ATP-synt_ab	2e-110	371.6
ATP-synt	6.5e-26	95.6
ATP12	3.9e-19	72.9

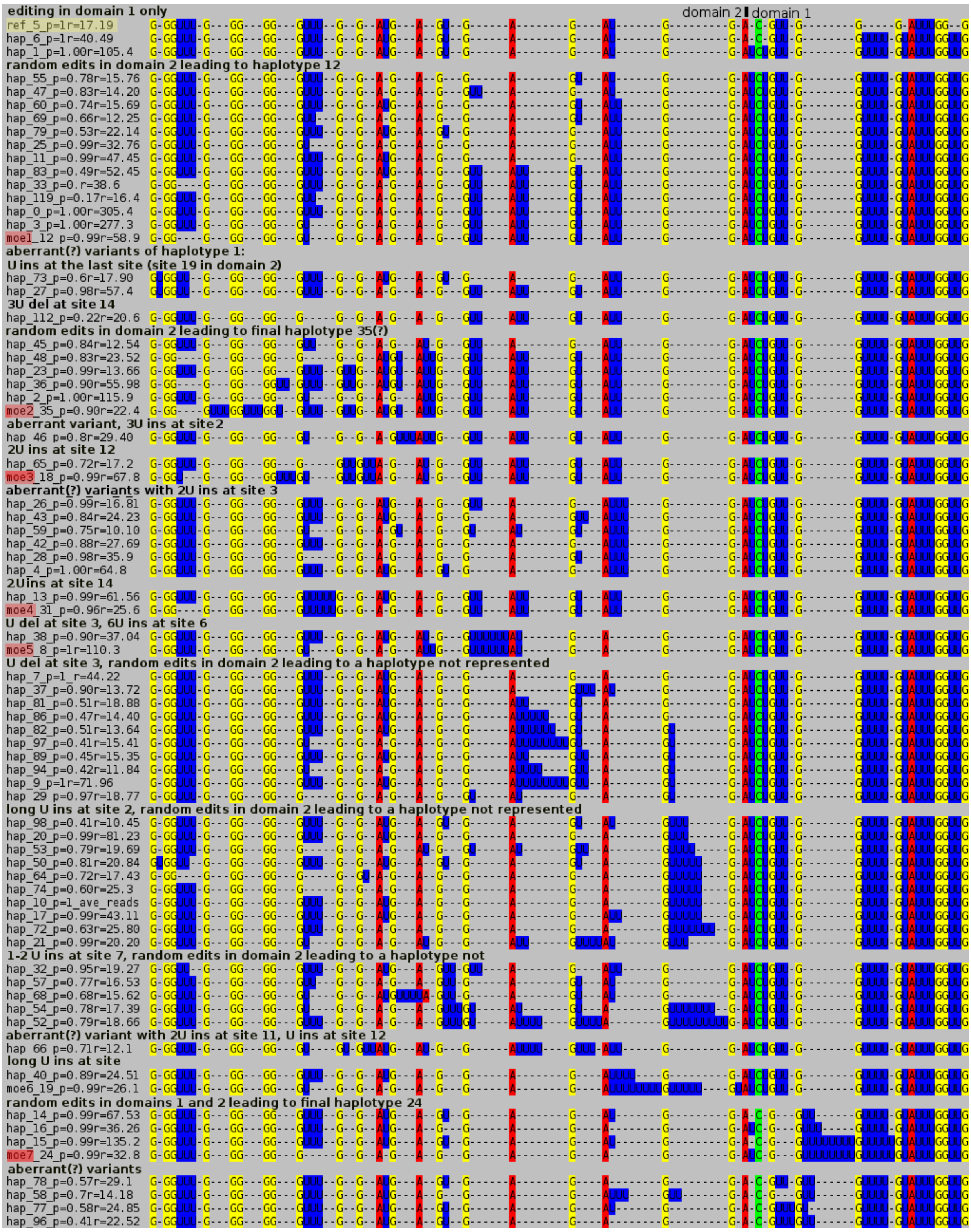
Table 2: HMMER identified proteins of ETC and alternative oxidase search.

Discussion

Mitochondrial gene content of *Perkinsela*

In our search we could not possibly identify pan-edited transcripts, diverged beyond recognition (if present), so the actual gene content of *Perkinsela* remains to be established. Final assembly and annotation will be finished beyond this thesis, because there will be more sequencing data with longer reads in near future, that might further enlarge current contigs and possibly close gaps between them. Among known kidnetoplsastids, the most similar gene order and RNA editing pattern encountered in *Trypanoplasma borreli* (Fig. 3, 17).

The lack of subunits of NADH dehydrogenase both in mitochondria and nucleus, along with *Perkinsela* nucleus-encoded complexes II, III, IV and V plus alternative oxidase protein, suggests the presence of functional electron transport chain with AOX reducing the oxidative stress supporting the situation, where the main purpose of mitochondria is rather biosynthesis than ATP production (Hannaert et al., 2003). This model would further support



Figue 22: Manually sorted clustering of the alignment composed of reads which span the whole 5' edited region of putative 9S rRNA transcript. r=average number of reads, p=posterior probability of a 'haplotype'. Putative fully edited sequences are highlighted in red, non-edited reference sequences are highlighted in yellow.

the endosymbiotic relationship between *Perkinsela* and *N. pemaquidensis*, because parasitic kinetoplastids rather rely on glycolysis and down-regulate mitochondrial activity when it comes to parasitic stages (Hannaert et al., 2003). For a big picture of *Perkinsela* mitochondrion metabolism, the complete set of nucleus-encoded mitochondrion-targeted proteins will be necessary.

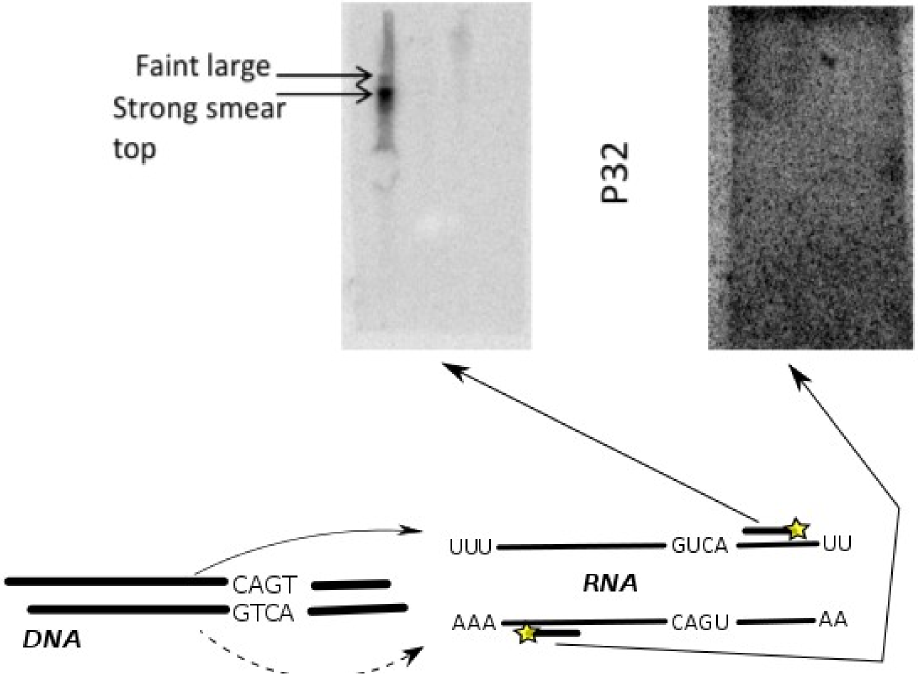
Enigmatic fate of mitochondrial rRNA subunits

We were surprised that 12S and 9S mitochondrial rRNA subunits remained unrecognized by BLAST in the genome assembly. However, when first mapping results appeared, two high-coverage peaks outnumbering identified protein-coding transcripts 100 and 60 times respectively, offered us the most likely explanation so far. Putative mitochondrial rRNA subunits are being edited by U-indel editing machinery like protein-the coding transcripts are. U-indel editing may have rapidly increased evolution speed of rRNA genes that finally have diverged beyond recognition.

The size heterogeneity of the alignment of 3' and 5' regions of putative 9S rRNA complicates the reconstruction of editing progress. As already shown on Fig. 21, the result consists of many read clusters without significant differences in their abundance. The branching pattern, that means the presence of alternative editing leading possibly to distinct functional transcripts (as in Ochsenreiter et al., 2008), is further supported by the sorted 5' edited alignment composed of 3,407 reads which span the whole 5' edited region of 9S on Fig. 22 on the next page. Remaining reads have mostly long inserts, and therefore they are unable to span the whole region.

According to the known mechanisms of RNA editing, it seems that 5' region of putative 9S transcript has two edited domains. First domain represents 3' part of the 5' region where most of the transcripts have 11Us inserted at 6 sites and first branching occurs with long insertion at site 5 (representing probably an alternative guide RNA for domain 1). In the upstream domain 2 the 3' to 5' directional pattern is lost as editing is likely directed by alternative long gRNAs matching the whole domain, and editing within each gRNA is apparently “random” with respect to the 3' to 5' coordinates. Editing in this particular region is so extensive,, that we might loose some reads that could overlap the reference sequence of the whole region. This and other limitations of our approach are discussed below.

Because our RNA library is not strand specific, Northern blot has been performed in order to confirm the orientation of putative 9S rRNA. Since this experiment has not been carried out by me, I will just briefly note that transcription of the U-indel edited strand has been unambiguously confirmed (see Fig. 23 for details).



Drawing 23: Northern blot results, showing orientation of the putative 9S rRNA transcript in *N.pemaquidensis* strainGILLNOR1/I endosymiont (I.Fiala, H.Hashimi unpublished).

Editing of *cox1*

Unlike the complicated situation with putative 9S transcript, *cox1* transcript is rather edited in a classical way (3' to 5' progression, relatively low-abundance editing errors, single final product,...), except for the obvious restriction of editing to transcript edges observed only in *T. borreli* so far. Other kinetoplastids edit only 3' ends or pan-edit the whole length. Interestingly, the most likely fully-edited product of *cox1* is edited with insertions only (28U at 5', 56U at 3'), although deletions are present in some mis-edited reads. The rest of the protein-coding transcripts, based on preliminary observations, seems to be edited in a very similar fashion to the *cox1* transcript and will be analyzed beyond the scope of this work.

Limitations of the mapping software

Here I will describe two major reasons, why some reads might be missing in our mapping results based on the experience with Bowtie2 software. First limitation is the fact, that exhaustive mapping is extremely time consuming and is substituted by seed search. Bowtie2 very-sensitive option uses seeds 20 nucleotides long. If a read is extensively edited over full length, no seed will be found and the read will be discarded. Second limitation is the balance of the scoring of alignments. Bowtie2 (in the end-to-end mode) uses by default ~60% read length read as the similarity cutoff (negative value). Combination of these rules means that only reads, which alignment to the template has 20 perfect matches in row (1/5 of the usual read length), and which sum of penalties for indels and mismatch errors is smaller than the cutoff (-60 for usual length), are included. The mapping efficiency of the current version of Bowtie2 can not be further enhanced, since both shorter seeds and looser cutoffs rapidly increase both abundance of artifacts and time required for mapping.

For this reason we are preparing an “editing-friendly” version of Bowtie2 software where nucleotide-specific penalties will be allowed, improving both the quality of the alignments and chance of artefactual sequences to be mapped. Testing the modified software shows that we will be able to map in reasonable time all reads that carry 16 nucleotide seeds and align reads with U/indels perfectly, except for the the very ends of reads. Alignment of read ends will be further corrected manually.

Summary

I have processed DNA and RNA sequencing data from *Neoparamoeba pemaquidensis* isolate CCAP 1560/4 and identified contigs corresponding to the mitochondrion of its endosymbiont, *Perkinsela.* Transcripts have been identified, and their U- insertion/deletion editing analyzed to some extent. Three protein-coding transcripts are edited at both ends (*cox1*, *cox3*, *cob*), one unidentified transcript is unedited, two high-coverage non-coding transcripts are edited at both ends (putative divergent rRNAs). This editing pattern is somewhat similar to that of *Trypanoplasma borelli*. Putative 9S rRNA transcript is probably alternatively edited giving rise to several variants at the 5' end. In my opinion, results of this work and the whole *N. pemaquidensis/Perkinsela* genome project will shed some light on the symbiotic relations of these organisms as well as evolution of kinetoplastid mitochondrial genome and RNA editing.

References

Blom, D., de Haan, A., van den Burg, J., van den Berg, M., Sloof, P., Jirku, M., Lukes, J., and Benne, R. (2000). Mitochondrial minicircles in the free-living bodonid Bodo saltans contain two gRNA gene cassettes and are not found in large networks. RNA *6*, 121–135.

Campbell, D.A., Thomas, S., and Sturm, N.R. (2003). Transcription in kinetoplastid protozoa: why be normal? Microbes and Infection *5*, 1231–1240.

Dyková, I., Figueras, A., and Peric, Z. (2000). Neoparamoeba Page, 1987: light and electron microscopic observations on six strains of different origin. Diseases of Aquatic Organisms *43*, 217–223.

Dyková, I., Fiala, I., Lom, J., and Lukes, J. (2003). Perkinsiella amoebae-like endosymbionts of Neoparamoeba spp., relatives of the kinetoplastid Ichthyobodo. European Journal of Protistology *39*, 37–52.

Dyková, I., Fiala, I., and Pecková, H. (2008). Neoparamoeba spp. and their eukaryotic endosymbionts similar to Perkinsela amoebae(Hollande, 1980): Coevolution demonstrated by SSU rRNA gene phylogenies. European Journal of Protistology *44*, 269–277.

Eddy, S. (1998). Profile hidden Markov models. Bioinformatics *14*, 755–763.

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

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

Embar-Gopinath, S., Crosbie, P., and Nowak, B.F. (2006). Concentration effects of Winogradskyella sp. on the incidence and severity of amoebic gill disease. Diseases of Aquatic Organisms *73*, 43–47.

Hannaert, V., Bringaud, F., Opperdoes, F.R., and Michels, P.A. (2003). Evolution of energy metabolism and its compartmentation in Kinetoplastida. Kinetoplastid Biology and Disease *2*, 11.

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

Koslowsky, D.J., Bhat, G.J., Read, L.K., and Stuart, K. (1991). Cycles of progressive realignment of gRNA with mRNA in RNA editing. Cell *67*, 537–546.

Lukes, J., Guilbride, D.L., Votỳpka, J., Zíková, A., Benne, R., and Englund, P.T. (2002). Kinetoplast DNA network: evolution of an improbable structure. Eukaryotic Cell *1*, 495–502.

Lun, Z.-R., Lai, D.-H., Li, F.-J., Lukeš, J., and Ayala, F.J. (2010). Trypanosoma brucei: two steps to spread out from Africa. Trends in Parasitology *26*, 424–427.

Metzker, M.L. (2009). Sequencing technologies—the next generation. Nature Reviews

Genetics *11*, 31–46.

Mitchell, S.O., and Rodger, H.D. (2011). A review of infectious gill disease in marine salmonid fish. Journal of Fish Diseases *34*, 411–432.

Morrison, R.N., Crosbie, P.B.B., Cook, M.T., and Nowak, B.F. (2005). Cultured gill-derived Neoparamoebapemaquidensis fails to elicit amoebic gill disease (AGD) in Atlantic salmon Salmo salar. Diseases of Aquatic Organisms *66*, 135–144.

Munday, B.L., Zilberg, D., and Findlay, V. (2002). Gill disease of marine fish caused by infection with Neoparamoeba pemaquidensis. Journal of Fish Diseases *24*, 497–507.

Ochsenreiter, T., and Hajduk, S.L. (2006). Alternative editing of cytochrome c oxidase III mRNA in trypanosome mitochondria generates protein diversity. EMBO Reports *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. Molecular and Cellular Biology *28*, 5595–5604.

Schatz, M.C., Delcher, A.L., and Salzberg, S.L. (2010). Assembly of large genomes using second-generation sequencing. Genome Research *20*, 1165–1173.

Schnaufer, A. (2010). Evolution of dyskinetoplastic trypanosomes: how, and how often? Trends in Parasitology *26*, 557–558.

Simpson, L. (2003). Uridine insertion/deletion RNA editing in trypanosome mitochondria: A complex business. RNA *9*, 265–276.

Simpson, A.G.B., Stevens, J.R., and Luke\vs, J. (2006). The evolution and diversity of kinetoplastid flagellates. Trends in Parasitology *22*, 168–174.

Simpson, L., Thiemann, O.H., Savill, N.J., Alfonzo, J.D., and Maslov, D.A. (2000). Evolution of RNA editing in trypanosome mitochondria. Proceedings of the National Academy of Sciences *97*, 6986–6993.

Sturm, N.R., Maslov, D.A., Blum, B., and Simpson, L. (1991). Generation of unexpected editing patterns in Leishmania tarentolae mitochondrial mRNAs: Misediting produced by misguiding. Cell *70*, 469–476.

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

Zaghordi et al. (2011). ShoRAH: estimating the genetic diversity of a mixed sample from next-generation sequencing data. BMC Bioinformatics *12*, 119.

S1:

Remaining manually sorted clustering results of RNAseq mapping

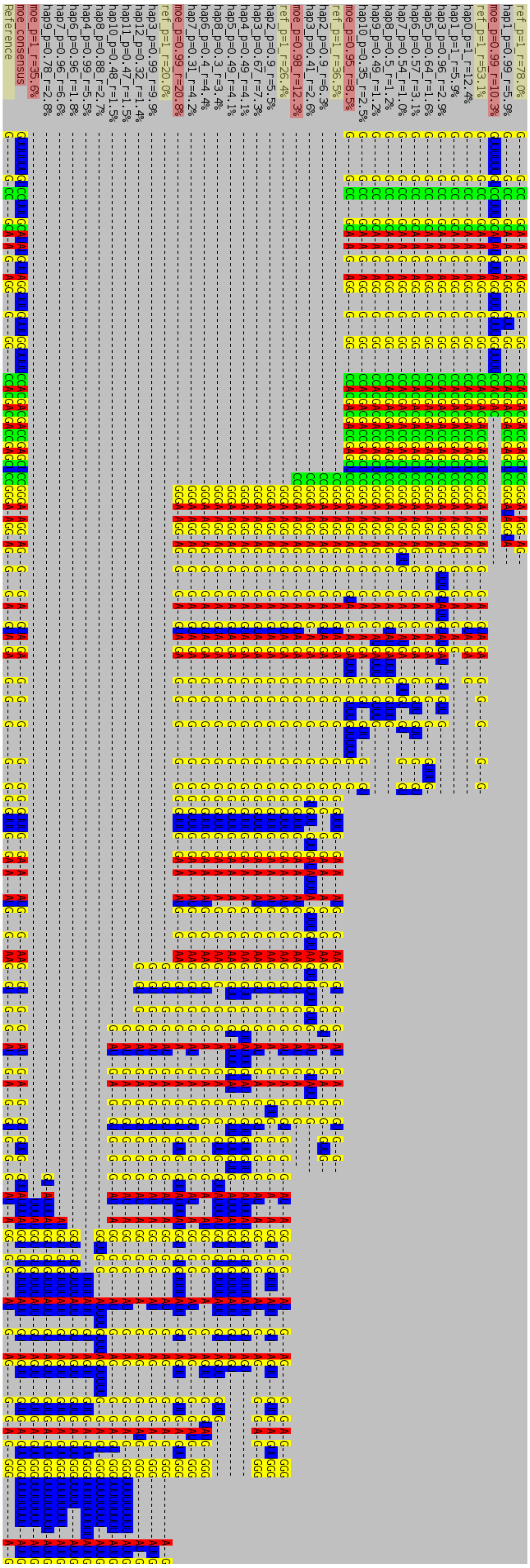
- **S1a:**

  Manually sorted clustering results of RNAseq mapping for the 3' region of *cox1*. r=% of average number of reads, p=posterior probability of a 'haplotype'. Hypothetical fully edited sequences are highlighted in red, non-edited reference sequences are highlighted in yellow

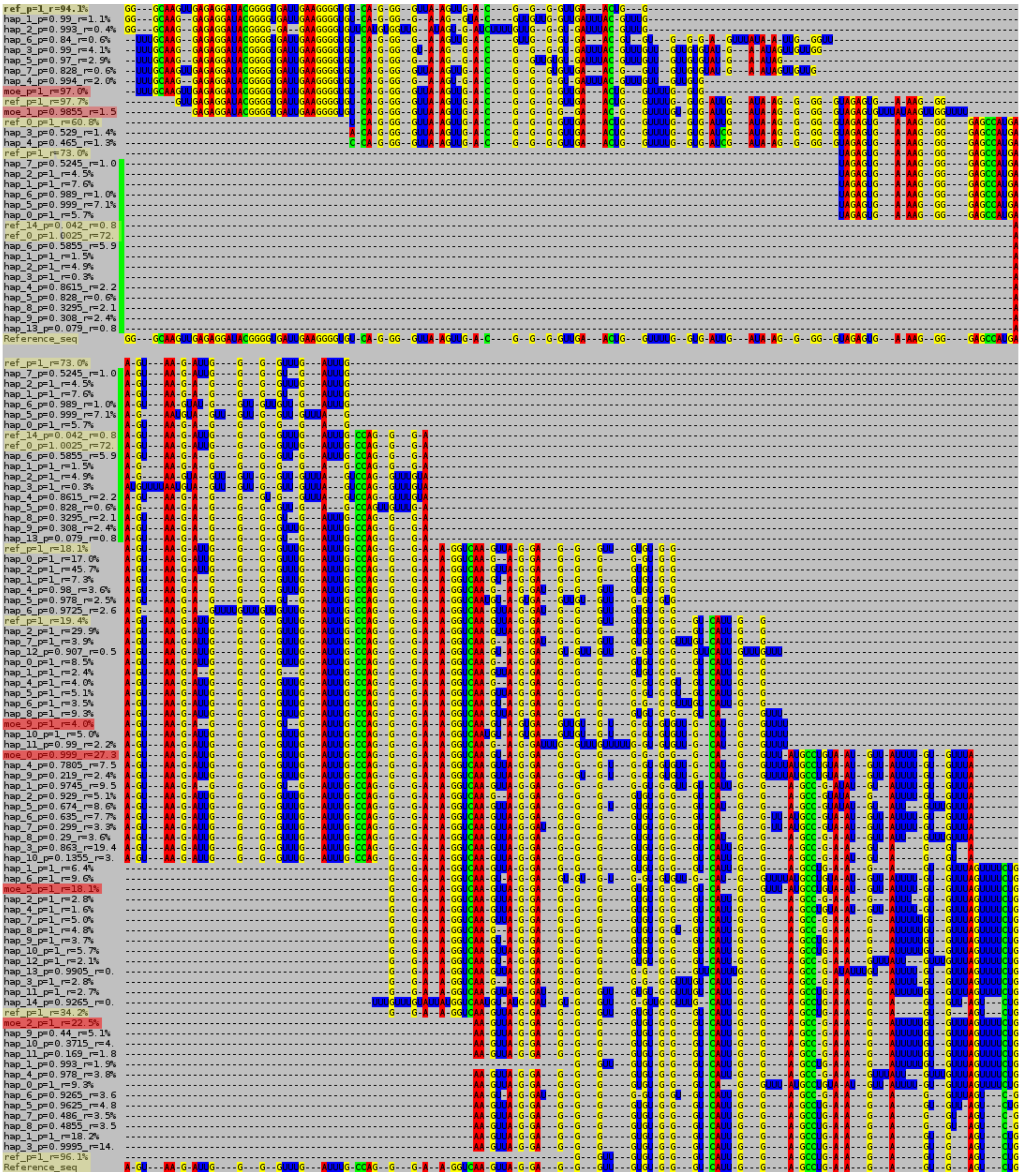
- **S2b:**

  Manually sorted clustering results of RNAseq mapping for the 3' region of putative 9S rRNA transcript split in two at the never-edited site (46nt). r=% of average number of reads, p=posterior probability of a 'haplotype'. Non-edited reference sequences are highlighted in yellow, abundant sequences are highlighted in red. Overlapping part of the alignment is represented by green bar.

S1a


Reference
moe_consensus
moe_p=1_r=35.6%
hap9_p=0.78_r=2.8%
hap7_p=0.96_r=6.6%
hap6_p=0.96_r=1.8%
hap4_p=0.99_r=5.5%
hap8_p=0.88_r=2.7%
hap10_p=0.48_r=1.5%
hap11_p=0.47_r=1.5%
hap12_p=0.32_r=1.4%
hap3_p=0.99_r=9.9%
ref_p=1_r=20.0%
moe_p=0.99_r=20.8%
hap7_p=0.31_r=4.2%
hap6_p=0.4_r=4.4%
hap8_p=0.3_r=3.4%
hap5_p=0.49_r=4.1%
hap4_p=0.49_r=4.1%
hap3_p=0.67_r=7.3%
hap2_p=0.9_r=5.5%
ref_p=1_r=26.4%
moe_p=0.98_r=12.3%
hap3_p=0.41_r=2.6%
hap2_p=0.9_r=6.3%
ref_p=1_r=36.5%
moe_p=0.95_r=8.5%
hap10_p=0.35_r=2.5%
hap9_p=0.49_r=1.2%
hap8_p=0.5_r=1.2%
hap7_p=0.54_r=1.0%
hap6_p=0.57_r=3.1%
hap5_p=0.64_r=1.6%
hap3_p=0.96_r=2.9%
hap1_p=1_r=5.9%
hap0_p=1_r=12.4%
ref_p=1_r=53.1%
moe_p=0.99_r=10.3%
hap1_p=0.99_r=5.9%
ref_p=1_r=78.0%

S1b:


ref_p=1_r=94.1%
hap_1_p=0.99_r=1.1%
hap_2_p=0.993_r=0.4%
hap_6_p=0.84_r=0.6%
hap_3_p=0.99_r=4.1%
hap_5_p=0.97_r=2.9%
hap_7_p=0.828_r=0.6%
hap_4_p=0.994_r=2.0%
moe_p=1_r=97.0%
ref_p=1_r=97.7%
moe_1_p=0.9855_r=1.5
ref_0_p=1_r=60.8%
hap_3_p=0.529_r=1.4%
hap_4_p=0.465_r=1.3%
ref_p=1_r=73.0%
hap_7_p=0.5245_r=1.0
hap_2_p=1_r=4.5%
hap_1_p=1_r=7.6%
hap_6_p=0.989_r=1.0%
hap_5_p=0.999_r=7.1%
hap_0_p=1_r=5.7%
ref_14_p=0.042_r=0.8
ref_0_p=1.0025_r=72.
hap_6_p=0.5855_r=5.9
hap_1_p=1_r=1.5%
hap_2_p=1_r=4.9%
hap_3_p=1_r=0.3%
hap_4_p=0.8615_r=2.2
hap_5_p=0.828_r=0.6%
hap_8_p=0.3295_r=2.1
hap_9_p=0.308_r=2.4%
hap_13_p=0.079_r=0.8
Reference_seq
ref_p=1_r=73.0%
hap_7_p=0.5245_r=1.0
hap_2_p=1_r=4.5%
hap_1_p=1_r=7.6%
hap_6_p=0.989_r=1.0%
hap_5_p=0.999_r=7.1%
hap_0_p=1_r=5.7%
ref_14_p=0.042_r=0.8
ref_0_p=1.0025_r=72.
hap_6_p=0.5855_r=5.9
hap_1_p=1_r=1.5%
hap_2_p=1_r=4.9%
hap_3_p=1_r=0.3%
hap_4_p=0.8615_r=2.2
hap_5_p=0.828_r=0.6%
hap_8_p=0.3295_r=2.1
hap_9_p=0.308_r=2.4%
hap_13_p=0.079_r=0.8
ref_p=1_r=18.1%
hap_0_p=1_r=17.0%
hap_2_p=1_r=45.7%
hap_1_p=1_r=7.3%
hap_4_p=0.98_r=3.6%
hap_5_p=0.978_r=2.5%
hap_6_p=0.9725_r=2.6
ref_p=1_r=19.4%
hap_2_p=1_r=29.9%
hap_7_p=1_r=3.9%
hap_12_p=0.907_r=0.5
hap_0_p=1_r=8.5%
hap_1_p=1_r=2.4%
hap_4_p=1_r=4.0%
hap_5_p=1_r=5.1%
hap_6_p=1_r=3.5%
hap_8_p=1_r=9.3%
moe_9_p=1_r=4.0%
hap_10_p=1_r=5.0%
hap_11_p=0.99_r=2.2%
moe_0_p=0.999_r=27.3
hap_4_p=0.7805_r=7.5
hap_9_p=0.219_r=2.4%
hap_1_p=0.9745_r=9.5
hap_2_p=0.929_r=5.1%
hap_5_p=0.674_r=8.6%
hap_6_p=0.635_r=7.7%
hap_7_p=0.299_r=3.3%
hap_8_p=0.29_r=3.6%
hap_3_p=0.863_r=19.4
hap_10_p=0.1355_r=3.
hap_1_p=1_r=6.4%
hap_6_p=1_r=9.6%
moe_5_p=1_r=18.1%
hap_2_p=1_r=2.8%
hap_4_p=1_r=1.6%
hap_7_p=1_r=5.0%
hap_8_p=1_r=4.8%
hap_9_p=1_r=3.7%
hap_10_p=1_r=5.7%
hap_12_p=1_r=2.1%
hap_13_p=0.9905_r=0.
hap_3_p=1_r=2.8%
hap_11_p=1_r=2.7%
hap_14_p=0.9265_r=0.
ref_p=1_r=34.2%
moe_2_p=1_r=22.5%
hap_9_p=0.44_r=5.1%
hap_10_p=0.3715_r=4.
hap_11_p=0.169_r=1.8
hap_1_p=0.993_r=1.9%
hap_4_p=0.978_r=3.8%
hap_0_p=1_r=9.3%
hap_6_p=0.9265_r=3.6
hap_5_p=0.9625_r=4.8
hap_7_p=0.486_r=3.5%
hap_8_p=0.4855_r=3.5
hap_1_p=1_r=18.2%
hap_3_p=0.9995_r=14.
ref_p=1_r=96.1%
Reference_seq

S2:
List of software

- **Samtools package**
  Samtools 0.1.18 is an open-source set of programs used for further processing of raw sam files which are primary output of nucleotide mapping to genome template. Generally individual parts can be arranged to software pipeline aiming for compression, general visualization or specific information extraction. (homepage: http://samtools.sourceforge.net/)

- **EMBOSS package**
  EMBOSS 6.4.0 is an open-source collection of simple command line based tools which are used both for basic and advanced treating of biological data. Its components can be found implemented in common user-friendly software which translates sequences, builds alignments by various algorithms, has ability to use BLAST for browsing local databases or NCBI resources. (homepage: http://emboss.sourceforge.net/)

- **SMALT**
  Smalt 0.7.0.1 is a software capable of mapping raw reads from sequencing on the assembled genomes or generally any sequence. Software is using Smith-Waterman algorithm to find best alignment form for each read. (homepage: http://www.sanger.ac.uk/resources/software/smalt/)

- **Bowtie2**
  Bowtie 2.0.6 is like SMALT a free to purchase mapping software. However, it has an advantage of modularity which is almost a need for RNA mapping that is affected by various post transcriptional processes. (homepage: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml)
- **HMMER**
  HMMER 3.0 is an advanced tool for protein similarity search using alignment profile as query. Search itself is using hidden Markov models to determine homologous sequences among supplemented database of sequences. (homepage: http://hmmer.janelia.org/)

- **CLC workbench**
  CLC Main Workbench 6.8.1 in an integrated collection of sequence analysis tools (Functionally CLCwb almost overlaps with EMBOSS package) with friendly interface that fits for brief analysis and comprehensible visualization of data. (homepage: http://www.clcbio.com/products/clc-main-workbench/)

- **ShoRah**
  ShoRah 0.6 is an open-source tool chain for cluster analysis of mapping results. This software originally meant to be used for analysis of mixed-sample data and estimate of genetic diversity suits well for virtually any heterogeneity or artifact estimation among properly aligned maping results. (homepage: http://www.bsse.ethz.ch/cbg/software/shorah)