Exception within exceptions – unusual organization of rRNA genes in *Euglena longa*

Paweł Hałakuc, Natalia Gumińska, Anna Karnkowska, Rafał Milanowski
Department of Molecular Phylogenetics and Evolution, Institute of Botany, Faculty of Biology, Biological and Chemical Research Centre, University of Warsaw
Żwirki i Wigury 101, 02-089 Warszawa
pawelhalakuc@biol.uw.edu.pl

ABSTRACT
Eukaryotic ribosomes are composed of a few dozen proteins and usually four RNA molecules. Genes coding three of them, 18S (SSU), 5.8S and 28S rRNA (LSU) are clustered in a single operon (rDNA) and transcribed together. In majority of known eukaryotes, rDNA occurs in several hundred copies, usually in tandem repeats. Several exceptions from standard are known, one of them is observed in euglenids.
The most known representatives of Euglenida (Euglenozoa, Excavata) are green euglenids and among them the euglenid model species – *Euglena gracilis*. It was revealed that all detectable copies of *E. gracilis* rDNA are located on thousands of extrachromosomal circular DNA molecules, each of them containing single rDNA copy. We analysed genomic sequences of *E. gracilis* and found continuous rDNA operon, in accordance with previous results. Similar picture was found in other euglenids, however organization of rDNA seems to be different in secondary osmothroph *Euglena longa*, sister species of *E. gracilis*. During analysis of its genome, we found two separate sequences containing the SSU and LSU rDNA respectively. The reason for this is the presence of a several hundred nucleotides long repeated sequence in both the internal transcribed spacer 1 (ITS-1) and the intergenic spacer (IGS). Based on this information, we predicted two possible structures of rDNA: the typical one with the SSU and LSU rDNA organised together, and the second with the SSU and the LSU rDNA separately. Data from long reads sequencing of *E. longa* genome and molecular analyses indicate that both structures are present in the cells of *E. longa* – an combination not known before in eukaryotes.

INTRODUCTION
In almost all eukaryotes 3 ribosomal RNAs (18S, 5.8S and 28S) are encoded within single operon – rDNA. In Euglenids only the model species *Euglena gracilis* was thoroughly investigated. All detected rDNA copies were localized on 11,5 kb extrachromosomal circular molecules, each of them containing single rDNA operon and the InterGenic Spacer (IGS). Furthermore, the 28S rRNA gene is fragmented into 13 smaller pieces which together with 5.8S rRNA are called LSU 1-14. It is believed that the operon is transcribed as a whole, with 13 Internal Transcribed Spacers (ITSs) removed posttranscriptionally (Schnare et al. 1990). Our analyses of whole genome sequencing data support this picture in majority of euglenids (Hałakuc *et al.* in prep). However, in secondary osmotrophic *Euglena longa*, sister species of *E. gracilis*, structure of rDNA could not be resolved using short read data. We acquired sequences of SSU rDNA and LSU 1-15 rDNA on two separate contigs, flanked by ~300 bp

long repeats. Thorough analysis of assembly graphs produced by SPAdes has shown that recreating of complete operon is possible with the assumption that two identical several hundred bp long repeats reside in both the ITS (between SSU and LSU) and the IGS (between LSU and SSU). However, alternative organization is also possible: the SSU and the LSU genes localized on separate molecules containing identical repeated region (Figures 1, 2). Arrangement was further complicated by presence of smaller repeat within the repeated region.

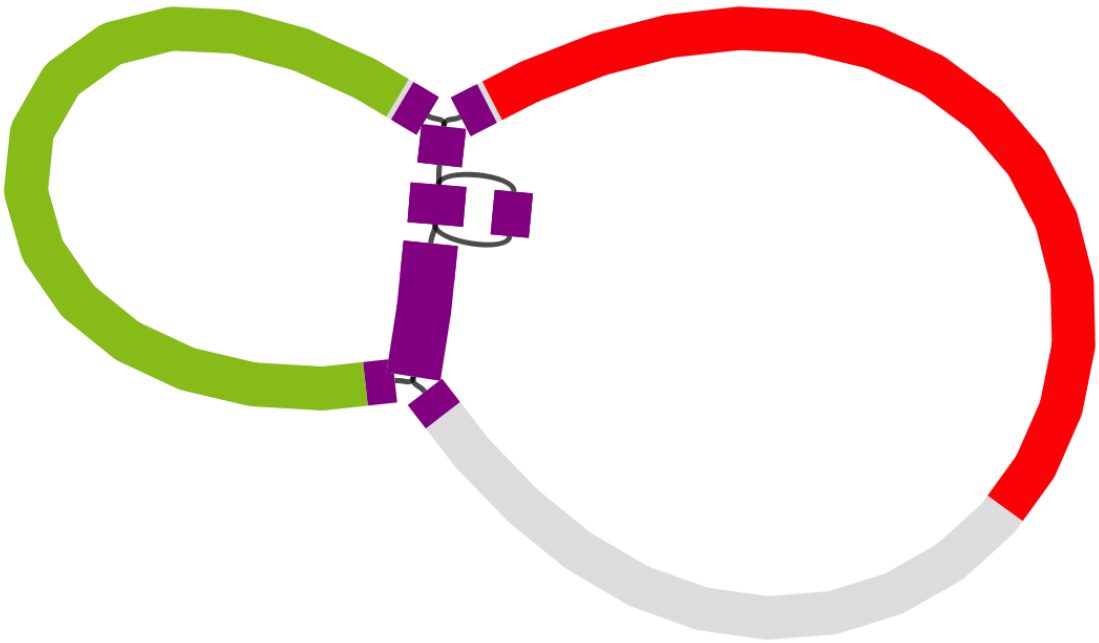

Figure 1
Fragment of assembly graph corresponding to the rDNA of *E. longa*. The SSU (green), the LSU (red) and repeated region (purple) are shown.

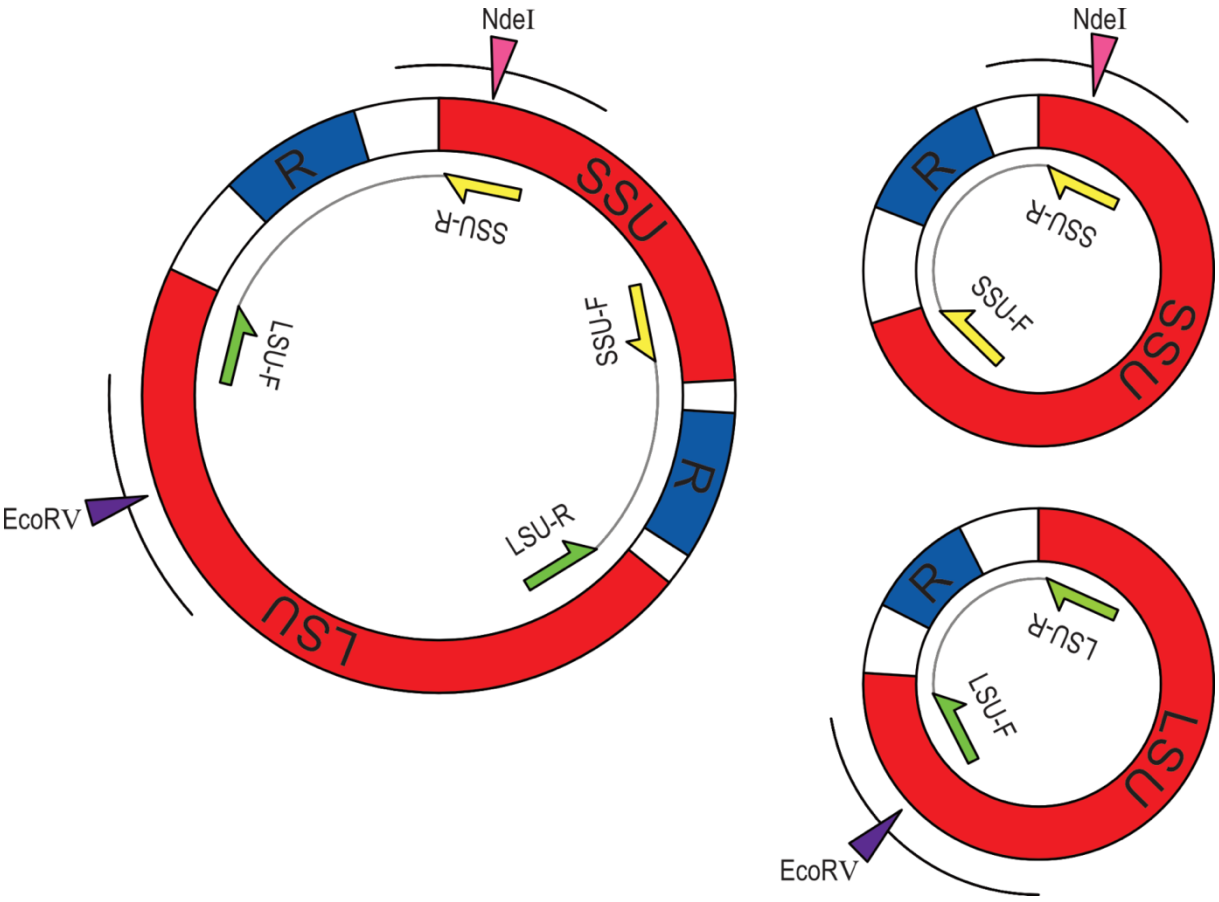

Figure 2
Schematic representation of hypothetical circular rDNA molecules in *E. longa*. The LSU 1-14 were concatenated for clarity. On scheme: the SSU and the LSU genes (red), repeated region (R, blue) inverse PCR primers (SSU-F and SSU-R in yellow; LSU-F and LSU-R in green), expected PCR products (gray arcs).

METHODS

Cultures of Euglena longa (CCAP 1204-17a) were cultivated statically in the Cramer- Myers medium (Cramer and Myers 1952), supplemented with ethanol (0.8% v/v) and aqueous soil extract (1% v/v). Cells were grown at 18 °C under white light exposure (16:8-h light/dark cycle, ca. 27 µmol photons m−2 s−1). Total DNA was extracted using modified CTAB method (Gumińska et al. 2018).

Preparation of a pair-end reads library was carried out externally using NEBNext DNA Library Prep Master Mix Set for Illumina (NEB) and sequenced commercially on an MiSeq (read length 250 bp) and HiSeq4000 (150bp) instruments at Genomed, Warsaw, Poland.

PacBio prep

The quality of raw short reads was evaluated using FastQC v0.11.5 (Andrews, 2010). Merging of paired reads was performed using PEAR v0.9.8 (Zhang et al. 2014) and the output was again evaluated using FastQC. Nucleotides of questionable quality were removed using Trimmomatic v0.36 (Bolger et al. 2014). Processed reads were assembled using SPAdes v3.12.0 (Nurk et al. 2013). The best results (as for rDNA) were acquired using `--meta` option.

Pacbio bioinf

Corrected PacBio reads were searched using local blastn (Camacho et al. 2009). Fragments of reapeted region, beginning and end of 18S rRNA, 5.8S rRNA and end of 28S rRNA genes were used as queries. Results were parsed using in-house script, reads corresponding to four arrangements (SSU-SSU, SSU-LSU, LSU-SSU and LSU-LSU) extracted and aligned in four separate groups using MAFFT v7.271 (Katoh and Standley 2013). Consensus sequences were used to resolve ambiguous regions.
To confirm bioinformatic results inverse PCR was performed. Starters were designed to match queries from blast serach (beginning and end of 18S rRNA, 5.8S rRNA and end of 28S rRNA genes). Four possible products were amplified: SSU-F/SSU-R, SSU-F/LSU-R, LSU-F/SSU-R and LSU-F/LSU-R (Figure 2). Products were put on x% agarose gel with MidoriGreen and visualized.

RESULTS
From 10 052 267 analyzed PacBio reads 22050 contained fragments corresponding to used queries. Specific considered variants and number of corresponding reads is shown in Table 1. In all cases specific order was required, e.g. in the SSU – repeat – LSU variant following hits needed to be present in order and on concordant strand: SSU-end, repeat and LSU-beginning. If only single type of hit was found in read it was classified as singular fragment. In case of unexpected arrangement (e.g. SSU-end, repeat, LSU-end) or discordant stand such read was classified as "other, possibly erroneous configuration". Lower number of reads in the (X –) repeat – SSU variants are most probably artifact, caused by longer sequence between repeated region and the beginning of the SSU than in other cases.

Table 1
Number of Pacbio reads classified to analysed variants. Reads corresponding to both hypothesis are bolded.

Variant	Number of reads
SSU – repeat – SSU	**5**
SSU – repeat – LSU	**614**
LSU – repeat – SSU	**155**
LSU – repeat – LSU	**11**
SSU – repeat	1 272
LSU – repeat	1 483
repeat – SSU	561
repeat – LSU	949
singular fragments	16 850
other, possibly erroneous configuration	150
total	22 050

Consensus sequences of all 4 informative groups contained only short indels compared to SPAdes results and indicated that in all cases additional small repeat was present within repeated region.
PCR products were acquired in all 4 variants (Fig. 3). Their lengths are in accordance with acquired PacBio consensus results.

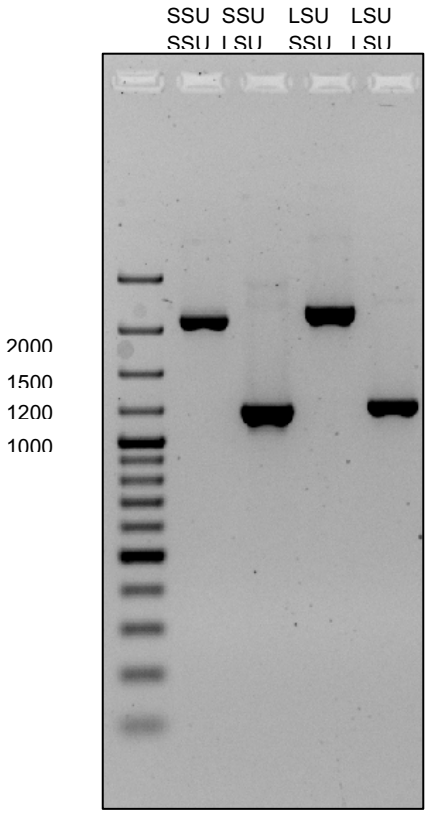

Figure 3
Picture of electrophoresis gel with PCR products.

DISCUSSION

PacBio results show that in fact 670 bp repeated region is located in both the SSU and the LSU. The PCR results confirm that at least small fraction of the SSU and the LSU genes are organized differently that in model species *E. gracilis*. Just a few PacBio reads correspond to those PCR products, however this may be the result of fractioning of the DNA size (>20 000 bp) before sequencing. It is much more probable that small fraction of big circles with both SSU and LSU (11 800 bp long) was not fully removed, than for smaller circles (4 400 bp for the SSU circle, 7 400 bp for the LSU circle).

It is also possible that other organizations are present in *E. longa* cells, e.g. small linear molecules, chromosomally located rDNA operons or mix of several types. Furthermore the structure may change in time, depending on the environmental conditions.

PERSPECTIVES

To confirm type of rDNA molecules in *E. longa* we are conducting Southern blot analysis with restriction enzyme digestion. Two restriction enzymes were chosen to include only one restriction site each in hypothetical molecules: NdeI in the SSU and EcoRV in the LSU. Probes were designed to much both sides of the restriction site (200-250 bp in both directions).

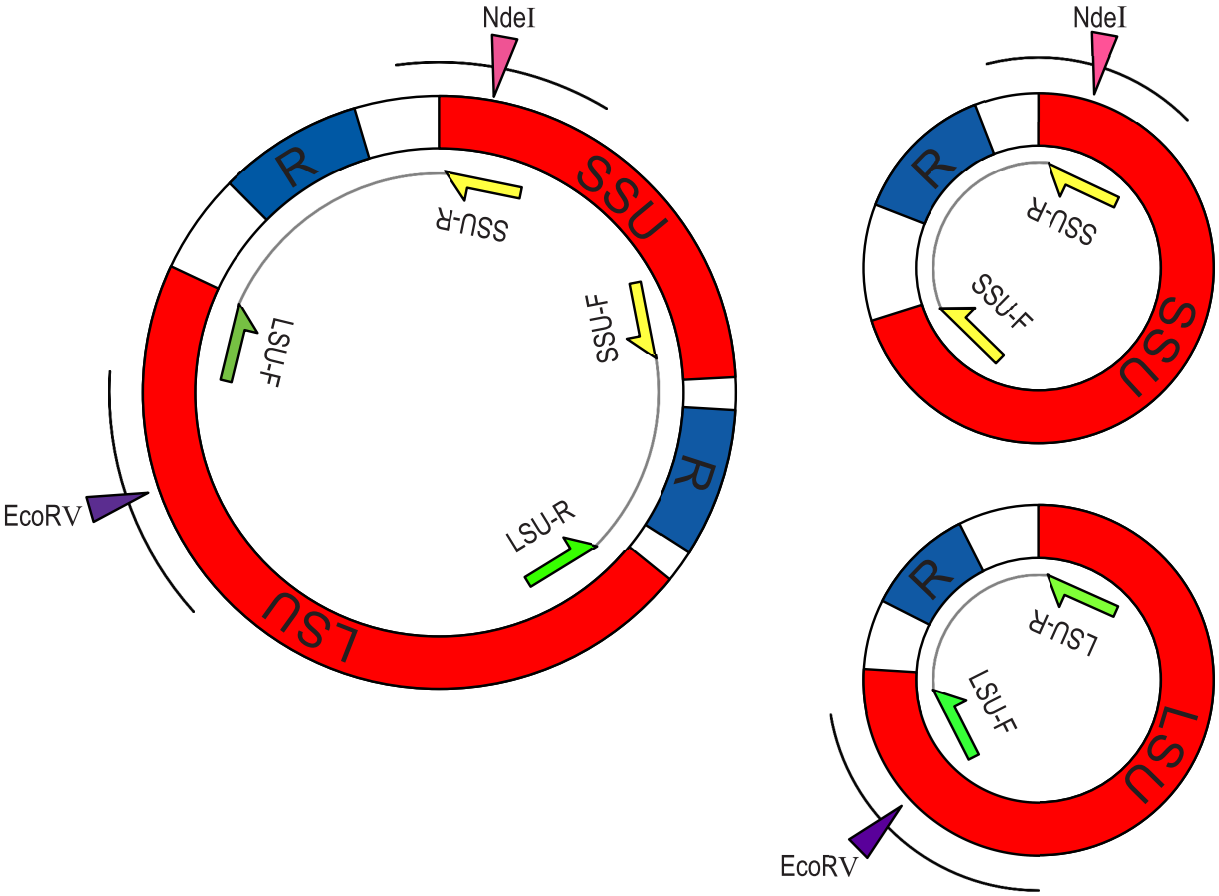

Figure 4
Scheme showing planned Southern blot analysis. Colors same as in Figure 2. Additionaly shown: restriction sites (pink for NdeI, purple for EcoRV), designed probes (black arcs).

Total DNA of *E. longa* will be digested in three variants: NdeI, EcoRV and NdeI + EcoRV. Combined with non-digested it will allow to unambiguously distinguish all possible variants (linear, circular, chromosomal) and ascertain ratio between different different forms.