

Complete mitochondrial genome sequence of the freshwater diatom *Asterionella formosa*

Adrien Villain^a, Mila Kojadinovic^b, Carine Puppo^b, Laura Prioretti^b, Pierre Hubert^c, Yizhi Zhang^b, Gérald Grégori^d, Alain Roulet^{e,f}, Céline Roques^{e,f}, Jean-Michel Claverie^{a,g}, Brigitte Gontero^b and Guillaume Blanc^a

^aInformation Génomique & Structurale UMR 7256, Aix Marseille Univ CNRS, IMM FR 3479, Marseille, France; ^bBIP UMR 7281, IMM FR 3479, Aix Marseille Univ CNRS, Marseille Cedex 20, France; ^cLaboratoire d'Ingénierie des Systèmes Macromoléculaires, Aix Marseille Univ CNRS UMR 7255 (IMM FR 3479), Marseille, France; ^dMediterranean Institute of Oceanography, Aix Marseille Univ, Univ Toulon, CNRS, Marseille, France; ^eGeT-PlaGe, Genotoul, INRA, Castanet-Tolosan, France; ^fUAR1209, INRA, Castanet-Tolosan, France; ^gAssistance Publique des Hôpitaux de Marseille (APHM), Marseille, France

ABSTRACT

We report the complete mitochondrial genome sequence of the freshwater diatom *Asterionella formosa*. The large 61.9 kb circular sequence encodes 34 proteins and 25 tRNAs that are universally conserved in other sequenced diatoms. We fully resolved a unique 24 kb region containing highly conserved repeated sequence units, possibly collocating with an origin of replication.

ARTICLE HISTORY

Received 23 December 2016
Accepted 18 January 2017

KEYWORDS

Diatoms; *Asterionella*;
mitogenome; pacbio





Diatoms are one of the largest and ecologically most significant groups of organisms on the Earth. These unicellular stramenopile algae are broadly distributed in marine and freshwater habitats and studied for potential biotechnological applications as well. *Asterionella formosa* Hassall (Lund 1949) is a freshwater araphid pennate diatom species forming typical star-shaped colonies. A single colony was isolated from Esthwaite Water (54.4N, 2.9W) in the English Lake District in December 2014. DNA was extracted following a hexadecyltrimethylammonium bromide (CTAB)-based protocol, and sequenced using the Pacific Biosciences RSII instrument.

Genome assembly of data from 13 SMRT cells was performed using the HGAP 2.0 protocol (Chin et al. 2013) implemented in SMRT analysis (2.3.0.140936.p0.0). A 82,419-bp long contig was identified as the mitochondrial genome and manually circularized into a 61,877-bp chromosome. Protein-coding genes were predicted by retaining all open reading frames (ORFs) > 100 codons, whereas ORFs < 100 codons were only predicted as genes when exhibiting a BLASTP (Altschul 1997) match (E -value < $1E-5$) in the NCBI non-redundant (Nr) database. For all validated genes, start codon predictions were further refined by comparison with homologous sequences. Transfer RNAs were predicted using tRNAscan SE (Lowe & Eddy 1997) and ribosomal RNAs were predicted by alignment with diatom reference sequences. We annotated repeated sequences by combining the results of tandem repeats finder (Benson 1999) and local BLASTN searches. Functional annotations were gathered and manually validated within Unipro UGENE (the UGENE team 2012). The genome

sequence is available in DDBJ/EMBL/GenBank under the accession no. KY021079.

The gene content of the *A. formosa* mitogenome is almost identical to previously published diatoms mitochondrial genomes (Secq & Green 2011; Ravin et al. 2010). The 62 genes include small and large rRNAs subunits, 25 tRNAs, and 35 protein-coding genes encoding 16 ribosomal protein subunits (rps), 10 NADH dehydrogenase subunits (nad), 3 ATPase subunits (atp), 3 cytochrome oxidase subunits (cox), the apocytochrome B, and the Sec-independent translocase protein TatC. A single type II intron with an intronic reverse-transcriptase domain is located in the *cox1* gene, while two are found in its homolog in other diatoms. The gene cluster *rps10-rps8-rpl6-rps2-rps4-atp8-rps12-rps7-rpl14-rpl5-nad1-tatC-rps11-rpl2-rps19-rps3-rpl16-atp9-nad4l-nad11* is present and seems conserved among diatoms. As in *Phaeodactylum tricorutum*, but unlike *Thalassiosira pseudonana* and *Synedra acus*, most protein-coding genes (32/35 = 91%) are encoded on the same strand. This strong-bias is apparently species-specific as it does not correlate with the phylogenetic relationships shown in Figure 1. Gene density is high and illustrated by overlaps between *rpl2* and *rps19*, and *rps19* and *rps3*. In contrast, a 24.9 kb long region devoid of predicted coding sequences and composed of successive blocks of various tandemly arrayed repeats is located between *trnQ* and *nad11*.

A single, large repeat region has already been described in other mitogenomes and has been suggested to serve as a replication origin in two cryptophytes (Hauth 2005; Kim et al. 2008). This feature is also typical of diatom mitogenomes

CONTACT Guillaume Blanc  guillaume.blanc@igs.cnrs-mrs.fr  Information Génomique & Structurale UMR 7256, Aix Marseille Univ CNRS, IMM FR 3479, Marseille, France; Brigitte Gontero  bmeunier@imm.cnrs.fr  BIP UMR 7281, IMM FR 3479, Aix Marseille Univ CNRS, Marseille Cedex 20, France

© 2017 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

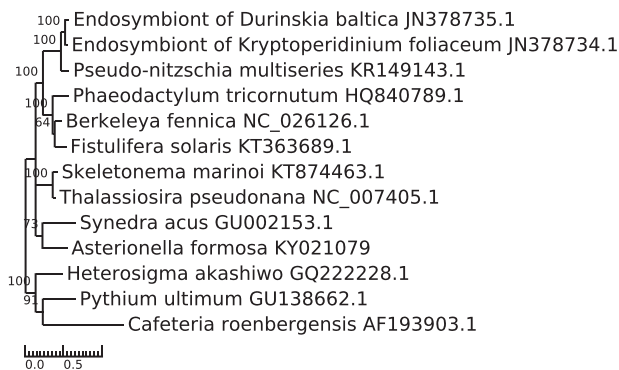


Figure 1. Visualization with TreeGraph (Stöver & Müller 2010) of a phylogenetic tree of selected diatoms and stramenopile mitogenomes based on the concatenation of the protein sequences of the following genes: *atp6*, *atp9*, *cox1*, *cox2*, *cox3*, *nad1*, *nad2*, *nad4*, *nad4l*, *nad5*, *nad6*, *nad7*, *rpl2*, *rps12*, *rps19*, *rps4*, *rps8*. Alignments were performed with MUSCLE (Edgar 2004) and the tree was constructed with PhyML (Guindon et al. 2009) 20120412 using the CpREV model, selected by ProtTest 3 (Darriba et al. 2011). Bootstrap values of 100 permutations are indicated at the nodes.

(Ravin et al. 2010; Secq & Green 2011), however, neither the sequences nor the general organizations of these repeats are conserved. A few ion torrent-sequenced diatom mitogenomes are reportedly lacking such a repeat region (An et al. 2016a, 2016b) but this absence may be due to the incapacity of short reads to resolve complex repeats, contrary to Sanger or Pacbio sequencing.

Disclosure statement

A declaration of interest statement reporting no conflict has been inserted. Please confirm the statement is accurate.

Funding

This work was supported by the A*MIDEX project (ANR-11-IDEX-0001-02) funded by the 'Investissements d'avenir' French Government program, managed by the French National Research Agency (ANR), the Centre national de la recherche scientifique interdisciplinaire PEPS project Exomod, the Aix-Marseille University, and la Région PACA. We acknowledge the use of the PACA-Bioinfo Platform, supported by IBISA and France-Génomique (ANR-10-INBS-0009). This work was performed in collaboration with the GeT core facility, Toulouse, France (<http://get.genotoul.fr>), and was supported by France Génomique National infrastructure, funded as part of 'Investissement d'avenir' program managed by Agence Nationale pour la Recherche (contract ANR-10-INBS-09).

References

- Altschul S. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389–3402.
- An SM, Kim SY, Noh JH, Yang EC. 2016a. Complete mitochondrial genome of *Skeletonema marinoi* (Mediophyceae, Bacillariophyta), a clonal chain forming diatom in the west coast of Korea. *Mitochondrial DNA Part B.* 1:549–550.
- An SM, Noh JH, Choi DH, Lee JH, Yang EC. 2016b. Repeat region absent in mitochondrial genome of tube-dwelling diatom *Berkeleya fennica* (Naviculales, Bacillariophyceae). *Mitochondrial DNA Part A.* 27:2137–2138.
- Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* 27:573–580.
- Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, et al. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods.* 10:563–569.
- Darriba D, Taboada GL, Doallo R, Posada D. 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics.* 27:1164–1165.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.
- Guindon S, Delsuc F, Dufayard J-F, Gascuel O. 2009. Estimating maximum likelihood phylogenies with PhyML, p. 113–137. In: Posada, D, ed. *Bioinformatics for DNA sequence analysis*. Totowa (NJ): Humana Press.
- Hauth AM. 2005. The *Rhodomonas salina* mitochondrial genome: bacteria-like operons, compact gene arrangement and complex repeat region. *Nucleic Acids Res.* 33:4433–4442.
- Kim E, Lane CE, Curtis BA, Kozera C, Bowman S, Archibald JM. 2008. Complete sequence and analysis of the mitochondrial genome of *Hemiselmis andersenii* CCMP644 (Cryptophyceae). *BMC Genomics.* 9:215
- Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25:955–964.
- Lund JWG. 1949. Studies on *Asterionella*: I. The origin and nature of the cells producing seasonal maxima. *J Ecol.* 37:389.
- Okonechnikov K, Golosova O, Fursov M, the UGENE team. 2012. Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics.* 28: 1166–1167.
- Ravin NV, Galachyants YP, Mardanov AV, Beletsky AV, Petrova DP, Sherbakova TA, Zakharova YR, Likhoshway YV, Skryabin KG, Grachev MA. 2010. Complete sequence of the mitochondrial genome of a diatom alga *Synedra acus* and comparative analysis of diatom mitochondrial genomes. *Curr Genet.* 56:215–223.
- Secq M-PO-L, Green BR. 2011. Complex repeat structures and novel features in the mitochondrial genomes of the diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana*. *Gene.* 476:20–26.
- Stöver BC, Müller KF. 2010. TreeGraph 2: combining and visualizing evidence from different phylogenetic analyses. *BMC Bioinformatics.* 11:7.