ISSN 1607-6729, Doklady Biochemistry and Biophysics, 2015, Vol. 461, pp. 84–88. © Pleiades Publishing, Ltd., 2015.
Original Russian Text © Y.P. Galachyants, Yu.R. Zakharova, D.P. Petrova, A.A. Morozov, I.A. Sidorov, A.M. Marchenkov, M.D. Logacheva, M.L. Markelov, K.V. Khabudaev, Ye.V. Likhoshway, M.A. Grachev, 2015, published in Doklady Akademii Nauk, 2015, Vol. 461, No. 3, pp. 348–352.

Sequencing of the Complete Genome of an Araphid Pennate Diatom *Synedra acus* subsp. *radians* from Lake Baikal[1]

Y. P. Galachyants[a], Yu. R. Zakharova[a], D. P. Petrova[a], A. A. Morozov[a], I. A. Sidorov[b], A. M. Marchenkov[a], M. D. Logacheva[c], M. L. Markelov[d], K. V. Khabudaev[a], Ye. V. Likhoshway[a], and Academician M. A. Grachev[a]

Received November 12, 2014

Abstract—High-throughput method of sequencing was applied to determine the complete nucleotide sequence of an araphid pennate diatom *Synedra acus* subsp. *radians* from Lake Baikal (East Siberia). The assembled genome has a total length of 98 Mbp, the mean coverage is 33x. Structure-functional annotation of the genome was performed.

DOI: 10.1134/S1607672915020064

One of the key problems in modern biology is to reveal the way by which genetic information orchestrates morphogenesis. In this respect, diatom algae represent a convenient object as these unicellular eukaryotes belonging to Chromista possess siliceous exoskeleton with species-specific micro- and nanostructure. Molecular and cellular mechanisms of the exoskeleton morphogenesis are insufficiently studied, and sequencing the complete genomes of diatoms gives us new tools to explore these mechanisms.

Diatoms are widespread in all aquatic and humid environments and are responsible for approximately 20% of the Earth primary production [1], that undoubtedly makes them an important object of the global ecosystem. To date, two complete genomes of diatoms have been published: those of centric (Coscinodiscophyceae) *Thalassiosira pseudonana* Hasle and Heimdal [2] and raphid pennate (Bacillariophyceae) *Phaeodactylum tricornutum* Bohlin [3]. Several more diatom genome projects are under way. We have chosen a freshwater araphid pennate diatom *Synedra acus* subsp. *radians* (Kütz.) Skabitsch. from Lake Baikal. First, we deciphered mitochondrial [4] and chloroplast [5] genome sequences of *Synedra acus* subsp. *radians* in collaboration with the Center "Bioengineering," Russian Academy of Sciences. Comparative analysis allowed us to find the ORF-containing self-splicing introns that were acquired by mitochondrial genome of *Synedra acus* subsp. *radians via* horizontal gene transfer and a region of diatom mtDNAs with intensive gene rearrangements as opposed to the rest of mitochondrial DNA sequence where the order of genes is conserved [4]. It was shown that the gene order in chloroplast DNA could be used for phylogenetic analysis [5].

Here we report on *de novo* sequencing of *Synedra acus* subsp. *radians* nuclear genome.

Establishing of an axenic culture and isolation of the genomic DNA. Monoclonal axenic cultures of *S. acus* subsp. *radians* were established from a phytoplankton specimen sampled in Listvennichny Bay of Lake Baikal [6]. Strain G9 was grown in 20-liter bottles on DM medium for culturing diatoms under natural light at 18°C. DNA was isolated as described earlier [4] with several modifications.

The nucleotide sequence of nuclear genome was determined using Next-generation sequencing methods with standard approaches to DNA fragmentation, DNA library preparation, and processing the raw sequencing data.

The final assembly of nuclear genome contains reads from fragmented genomic libraries sequenced using 454/Roche GS FLX (LIN SB RAS) and Illumina MiSeq (CRIE). We also sequenced a series of mate-pair libraries with various insert sizes (3–4, 4–6, 6–8, 8–10 and >10 kbp) by Illumina MiSeq (MSU). Assembly was performed using GS De Novo Assem-

[1] The article was translated by the authors.

[a] *Limnological Institute, Siberian Branch of the Russian Academy of Sciences, Irkutsk*
[b] *Institute for System Dynamics and Control Theory, Siberian Branch of the Russian Academy of Sciences, Irkutsk*
[c] *Lomonosov Moscow State University, Moscow*
[d] *Central Research Institute of Epidemiology, Moscow*
e-mail: yuri.galachyants@lin.irk.ru

bler 2.8 ("F. Hoffmann-La Roche Ltd.") on a high-performance server "Tesla," and annotation—on a high-performance computing cluster "Academician V. M. Matrosov" at ISDCT SB RAS (http://hpc.icc.ru). Noticeably, inclusion of mate-pair reads increased significantly the assembly connectivity and allowed us to reach the following statistics: $N50_{contig}$ = 3.8 kbp, $N50_{scaffold}$ = 100.8 kbp; total length of scaffolds—98 Mbp, mean coverage—33x. There are long genomic regions in the assembly, possessing an increased coverage with peaks 60–80× and 100–180×.

To estimate the completeness of the assembly, we utilized genomic and transcriptomic data of *T. pseudonana* [2] and *Ph. tricornutum* [3] as well as *Pseudo-nitzschia multiseries* (Hasle) Hasle and *Fragilariopsis cylindrus* (Grunow) Krieger (http://genome.jgi.doe.gov).

Search for core eukaryotic genes (CEG) with CEGMA (Core Eukaryotic Gene Mapping Approach) [7] in genome assembly of *S. acus* subsp. *radians* revealed 391 out of 458 CEGs and 219 out of 248 highly-conserved CEGs (UCEGs). The draft assembly of the *S. acus* subsp. *radians* genome is comparable with other diatom genome assemblies (Table 1).

An extended amount of paralogous UCEGs (Table 1) and the presence of long regions with increased coverage in the assembly argue for multiple gene/genome duplications in the *S. acus* subsp. *radians* nuclear genome.

Nuclear genome annotation. We used Maker2 pipeline [8] to define the gene boundaries and the exon-intron structure of genes in the *S. acus* subsp. *radians* nuclear genome assembly. CEGMA-derived gene models were used to train GeneMark software that predicted gene models for the *S. acus* subsp. *radians* genome within Maker2. These models were then used to train the refined HMM in Augustus software and to perform the second round of gene models prediction with Maker2. For both Maker2 runs, nucleic acid sequences of protein-coding genes and predicted amino-acid sequences of proteins from one or several known diatom genomes were added to the input data. The best training dataset for gene callers turned out to be a united set of transcribed and predicted amino-acid sequences from *T. pseudonana, Ph. tricornutum* and *P. multiseries* (Fig. 1).

Table 1. Distribution of the complete sequences of highly-conserved core eukaryotic genes (UCEGs) in nuclear genomes of diatoms

Species	Number of UCEG groups, % of 248	Number of UCEG genes	Average number of genes per UCEG
T. pseudonana	232 (94)	301	1.30
Ph. tricornutum	235 (95)	303	1.29
F. cylindrus	229 (92)	401	1.91
P. multiseries	231 (93)	305	1.32
S. acus subsp. *radians*	219 (88)	375	1.71

We identified 27 337 gene models for *S. acus* subsp. *radians* protein-coding genes, with 11 184 of them having the annotation edit distance (AED) less than 1. The distribution of AED indices lower than 1 is strongly skewed to the left: more than 90% of AED values are less than 0.4.

It should be noted that three diatom genomes are comparable in terms of gene numbers, number of introns and exons and some other characteristics (Table 2). Somewhat larger size of the *S. acus* subsp. *radians* genome may be related to specific peculiarities of its structure which appeared in the process of evolution as well as to the larger cell volume [9].

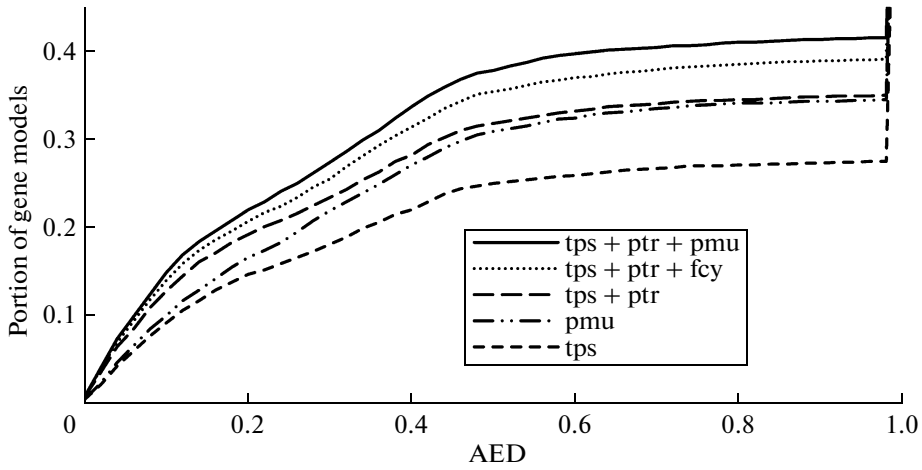

Fig. 1. Cumulative curves for annotation edit distance values of *S. acus* subsp. *radians* gene models built with different training sets: tps—*Th. pseudonana*, ptr—*Ph. tricornutum*, pmu—*P. multiseries*, fcy—*F. cylindrus.*

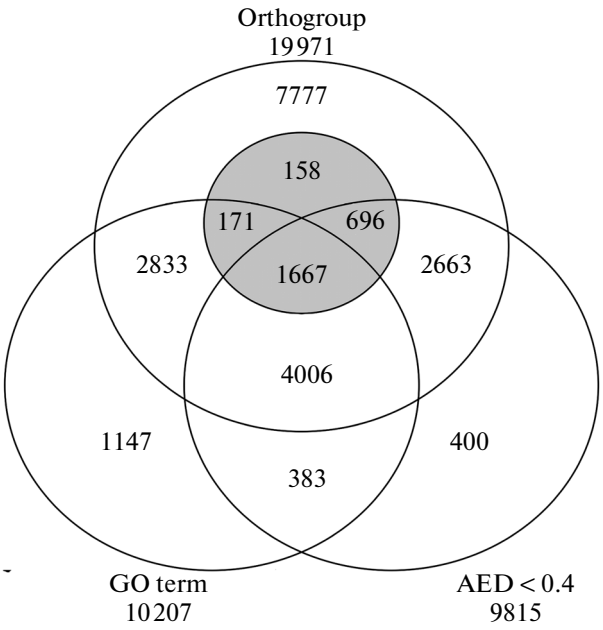

Fig. 2. Venn diagram for various subsets of 27 337 *S. acus* subsp. *radians* gene models. Orthologs shown in grey circle.

Functional annotation of *S. acus* subsp. *radians* gene models was performed with Blast2GO program [10] using the results of BLAST search against NR database and InterProScan [11] search for functional domains in amino-acid sequences. Gene Ontology (GO) terms were assigned for 10 207 gene models (Fig. 2, 3) with 2729 of them having EC number and enzyme name. Analysis of predicted amino-acid sequences in OrthoMCL [12] allowed us to find 13 584 groups of similar protein-coding genes of diatoms (orthogroups) that can include orthologous, coorthologous and inparalogous genes. Out of 27 337 *S. acus* subsp. *radians* gene models, 19 971 are included into 8860 orthogroups. More than 90% (9032) of *S. acus* subsp. *radians* genes having AED < 0.4 fall into some orthogroup (Fig. 2). Additionally, at least one GO term (Fig. 3) is assigned to more than a half of *S. acus* subsp. *radians* genes that belong to some orthogroup and have AED < 0.4. 2692 of *S. acus* subsp. *radians* genes have exactly one related sequence in all four diatom genomes, and function of 854 (31.7%) orthologs is unknown.

Silicon transporter (SIT) genes are present in the *S. acus* subsp. *radians* nuclear genome. The *S. acus* subsp. *radians* genome possesses 18 out of 26 genes that are coexpressed with SIT in *T. pseudonana* [13] as well as subunits of both coat protein (COT) complexes and clathrins that are also expressed with SIT in the same organism [14]. Genes of aquaporin, cyclins, frustulins and chitin synthase and other genes which may be involved in cell division and frustule morphogenesis are also found in the *S. acus* subsp. *radians* nuclear genome [15]. Given that diatoms have unique pathways of silica transport and deposition, there is a high probability to discover new genes participating in the formation of species-specific silica frustule among orthogroups and orthologs with unknown function.

Specific diatom protein sequences were shown to be highly divergent as revealed by phylogenetic analysis. For instance, genomes of seaborne *T. pseudonana, Ph. tricornutum* and freshwater *S. acus* subsp. *radians* contain different sets of aquaporins [15]. Interestingly, the predicted amino-acid sequences of highly conservative *rpb1* gene encoding β-subunit of RNA polymerase II have only 73% similarity between representatives of two classes of pennate diatoms—araphid *Fragilariophyceae* (*S. acus* subsp. *radians*) and raphid Bacillariophyceae (*Ph. tricornutum*). This is comparable with divergence of the *rpb1* between human and drosophila which is equal to 74%.

To summarize, we have assembled and annotated the draft version of the *S. acus* subsp. *radians* nuclear genome. This result makes it possible to proceed to a broad range of studies aimed at deciphering mechanisms of morphogenesis of diatoms using methods of molecular biology, genetic engineering and protein chemistry. In addition, comparative analysis of complete diatom genomes will allow detection of major genomic and genetic events in the evolution of such a divergent and ecologically important group.

Table 2. Statistics on nuclear genomes of diatoms

Feature	*T. pseudonana**	*Ph. tricornutum**	*S. acus* subsp. *radians*
Genome size, Mbp	32.4	27.4	98.4
		Protein-coding genes	
Predicted genes	11390	10025	11184
Number of exons	28910	17992	22633
Average number of exons per gene	2.54	1.79	2.02
Average gene length, bp	1561	1514	1782
Average exon length, bp	612	842	696

* According to annotation at http://genome.jgi.doe.gov.

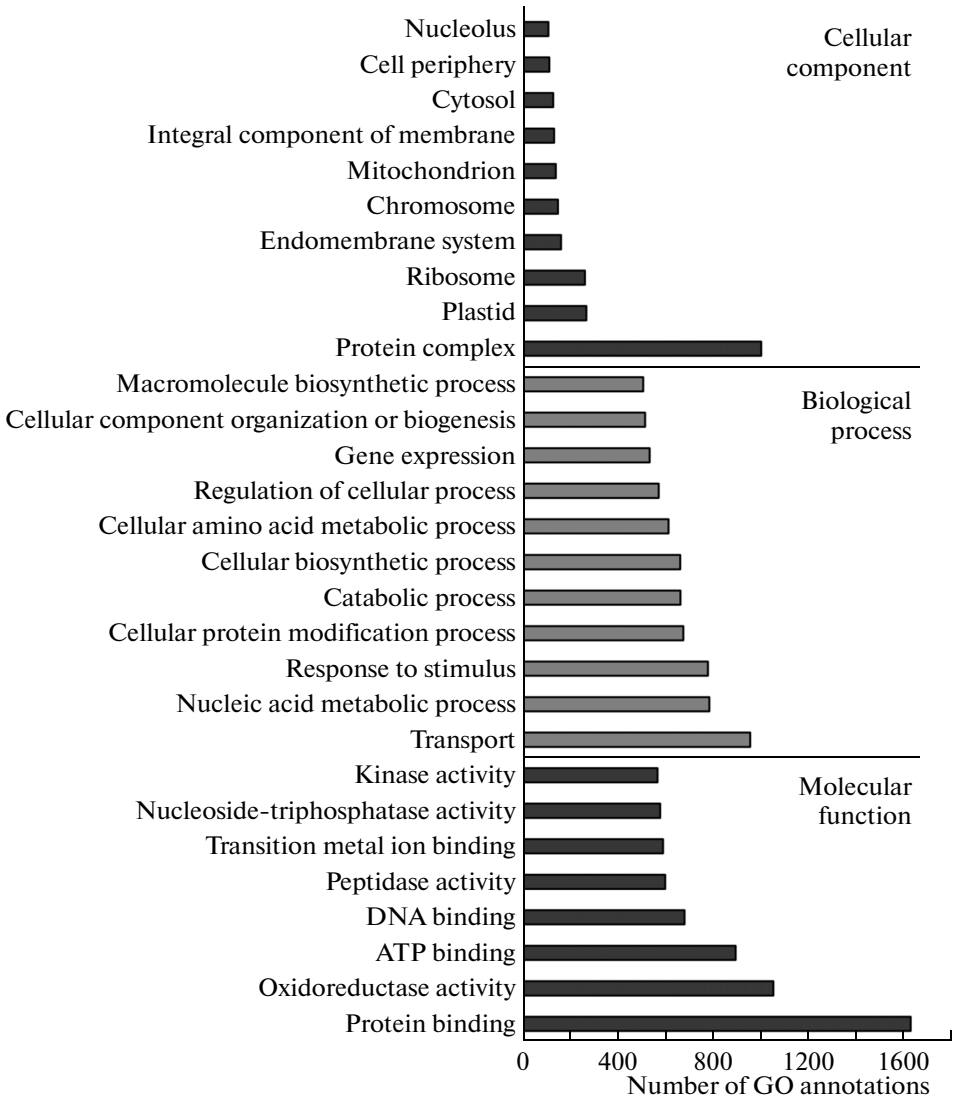

Fig. 3. Distribution of the most abundant GO-terms among 10 207 annotated *S. acus* subsp. *radians* gene models.

Assembly and annotation data for the *S. acus* subsp. *radians* complete genome are freely available at http://lin.irk.ru/sacus under the terms of CC-BY-NC4.0 license.

We are grateful to Professor A.S. Kondrashov (Moscow State University) for the opportunity to use the material and technical resources of the Laboratory of Evolutionary Genomics, Faculty of Bioengineering and Bioinformatics, Moscow State University. This work was supported by the Program of the Presidium of the Russian Academy of Sciences "Molecular and Cellular Biology" (project no. 6.9) "Investigation of mechanisms of assimilation and intracellular transmembrane transport of silicon and morphogenesis of genetically programmed siliceous ultrastructures (sequencing and comparative analysis of sequences); projects of the Federal Agency of Scientific Organizations no. VI.61.1.4 "Experimental Studies of Genomes and Proteomes" (annotation of the genome) and no. VI.50.1.3 "The Study of Genetic, Molecular, Evolutionary, and Ecological Aspects of Representatives of the Kingdom Chromista as Major Producers of Biogenic Silica and Participants in the Cycle of Biogenic Elements of Aquatic Ecosystems" (cultivation); Integration Project no. 137 of the Siberian Branch, Russian Academy of Sciences "Development of New Approaches to the Use of Supercomputers to Decrypt Nucleotide Sequences in Generation NEXT Sequencers (implementation of computationally intensive operations).

REFERENCES

1. Tréguer, P., Nelson, D.M., Van Bennekom, A.J., et al., *Science*, 1995, vol. 268, pp. 375–379.
2. Armbrust, E.V., Berges, J.A., Bowler, C., et al., *Science*, 2004, vol. 306, no. 5693, pp. 79–86.
3. Bowler, C., Allen, A.E., Badger, J.H., et al., *Nature*, 2008, vol. 456, no. 7219, pp. 239–244.

4. Ravin, N.V., Galachyants, Yu.P., Mardanov, A.V., et al., *Curr. Genet.*, 2010, vol. 56, pp. 215–223.
5. Galachyants, Yu.P., Morozov, A.A., Mardanov, A.V., et al., *Int. J. Biol.,* 2012, vol. 4, pp. 27–35.
6. Shishlyannikov, S.M., Zakharova, Yu.R., Volokitina, N.A., et al., *Limnol. Oceanogr. Meth.*, 2011, vol. 9, pp. 478–484.
7. Parra, G., Bradnam, K., and Korf, I., *Bioinformatics*, 2007, vol. 23, pp. 1061–1067.
8. Holt, C. and Yandell, M., *BMC Bioinformatics*, 2011, vol. 12, p. 491.
9. Connolly, J.A., Oliver, M.J., Beaulieu, J.M., et al., *J. Phycol.*, 2008, vol. 44, pp. 124–131.
10. Götz, S., García-Gómez, J.M., Terol, J., et al., *Nucl. Acids Res.*, 2008, vol. 36, pp. 3420–3435.
11. Jones, P., Binns, D., Chang, H.Y., et al., *Bioinformatics*, 2014, vol. 30, pp. 1236–1240.
12. Li, L., Stoeckert, C.J., and Roos, D.S., *Genome Res.,* 2003, vol. 13, pp. 2178–2189.
13. Shrestha, R.P., Tesson, B., Norden-Krichmar, T., et al., *BMC Genomics*, 2012, vol. 13 p. 499.
14. Du, C., Liang, J.-R., Chen, D.-D., et al., *J. Proteome Res.*, 2014, vol. 13, pp. 720–734.
15. Khabudaev, K.V., Petrova, D.P., Grachev, M.A., and Likhoshway, Ye.V., *BMC Genomics*, 2014, vol. 15, p. 173.