Short communication

ISSN 2658-3518
LIMNOLOGY FRESHWATER BIOLOGY
www.limnolfwbiol.com

Intragenic duplications in diatom algae: evolution and expression

Morozov A.A.*, Galachyants Yu.P., Marchenkov A.M., Bairamova E.M.

Limnological Institute, Siberian Branch of the Russian Academy of Sciences, Ulan-Batorskaya Str., 3, Irkutsk, 664033, Russia

ABSTRACT. In our earlier work, multiple diatom protein-coding genes were found to consist of long repeats, in some cases even including several copies of the entire protein within a single open reading frame. 28 of them were amplified and sequenced from *Fragilaria radians* total DNA, thus proving that they are not assembly or annotation artifacts. Out of these 28, 16 were successfully sequenced from cDNA of the same species, confirming their expression and, hypothetically, functionality. Much like the original set of unvalidated sequences, this group shows little functional and structural specificity. Most duplications are shared with other diatom species, implying either multiple independent duplications or an ancient duplication event and vertical inheritance.

Keywords: Duplications, diatoms, protein evolution

1. Introduction

In our previous works, multiple diatom genes were found to contain the genes created by the domain-scale duplications. These genes contain multiple copies of the domain which is usually single-copy in other organisms; in the most extreme cases, multiplicated genes contain literally several copies of a whole protein concatenated within a single open reading frame. Genomes and transcriptomes of most diatom algae contain tens to hundreds of such genes, roughly 0.1-1% of the predicted proteome. No functional or structural category was found to be enriched in multiplicated genes across the species. To investigate whether they are some sort of assembly or annotation artifacts, we have previously attempted to amplify and sequence 66 multiplicated genes (out of 342 found *in silic*o) from total DNA of *Fragilaria radians,* a major producer in lake Baikal. 28 of them were successfully validated (Morozov, 2019). The goal of current work was to find out whether they are expressed via amplification from cDNA matrices, as well as to find out whether they have multiplicated homologs in other diatoms.

2. Materials and Methods

2.1. Amplification and sequencing

To confirm that the sequences found in genome are indeed expressed, cDNA was prepared from axenic *F. radians* culture. It was used for PCR amplification with Encyclo kits (Evrogen, Russia) and primer pairs that successfully amplified these genes from total DNA. Sanger sequencing was performed with BigDye 3.1 (Applied Biosystems, USA) and 3130XL analyzer (Applied Biosystems, USA) at SB RAS Genomics Core Facility (Novosibirsk, Russia).

2.2. MCL clustering

To detect homology between the multiplicated genes of different diatoms, the entire sequence set for all species (both validated and not) was aligned to itself using DIAMOND and clustered with MCL at the inflation parameter of 4.0. Functional annotation of clusters and individual sequences was performed using HMMER and hidden Markov models from the PFAM database.

3. Results and Discussion

3.1. Expressed multiplicated genes

14 genes out of 28 tested were found to be expressed. Much like the original dataset of multiplicated genes detected *in silico,* this group shows no functional specificity. Strict GO enrichment analysis (or something similar) was not performed because of both small dataset size and the difficulty of formulating null hypothesis for such an artificial collection of sequences. However, validated gene set appears to be a random sample from the genome, including enzymes (glycosyl transferase, beta mannosidase, E3 ubiquitin ligase), transport proteins (ABC family transporter, Ca^{2+}/Mg^{2+} exporter), and diatom- or heterokont-specific proteins of unknown function.

*Corresponding author.
E-mail address: morozov@lin.irk.ru (A.A. Morozov)

3.2. Clustering and homology

This has produced 1164 clusters of 2 or more species have been produced, 704 of which had a common domain (as detected by HMMER with PFAM models) shared by at least half of the cluster. For most diatoms, 70% to 98% of sequences belonged to interspecies clusters. This implies that most of these genes either were created in a relatively ancient duplication, or duplicated independently in two or more taxa.

Among the validated genes of *F. radians,* two (glycosyl transferase and a protein of unknown function containing the cold shock domain) only have duplicated paralogs within the same species. For all the others, the same gene (and presumably protein) is also duplicated in other diatoms. Small cluster sizes for most proteins imply independent duplications, but phosphoglyceromutase, Ca^{2+}/Mg^{2+} exporter and ABC-family transporter subunits have tens of multiplicated homologs suggesting an ancient duplication. It is interesting to note that the latter two are olygomeric transmembrane transporters, a functional group known to have ungergone multiple duplications/merges in early eukaryotic history (Hennerdal et al., 2010). This group also includes SIT, whose evolution has followed a merger of two bacterial subunits into a single eukaryotic protein with a diatom-specific duplication (Durkin et al., 2016).

4. Conclusions

The set of 14 genes validated during this work has been proven to exist both in the genome and transcriptome of *F. radians*. Although it is by no means exhaustive, these genes can serve as models for further experiments into how genes affected by these multiplications work and interact with their single-domain counterparts.

Acknowledgements

This work was supported by the RFBR project 18-34-00441.

References

Durkin C.A., Mock Th., Armbrust E.V. 2016. The evolution of diatom-like silicon transporters in diatoms. Journal of Phycology 52: 716-731. DOI: 10.1111/jpy.12441

Hennerdal A., Falk J., Lindahl E. et al. 2010. Internal duplications in alpha-helical membrane protein topologies are common but the nonduplicated forms are rare. Protein Science 19(12): 2305-2318. DOI: 10.1002/pro.510

Morozov A.A. 2019. Intragenic duplications in diatom algae: search and validation. Trudy Instituta Biologii Vnutrennikh Vod RAN [Proceedings of the Institute for Biology of Inland Waters RAS] 88(91): 50-56. (in Russian)