PhD project

Endogenous viral elements and the evolution of brown algal genomes

Context

The brown algae (Phaeophyceae) are key components of coastal ecosystems with important roles as primary producers and as habitats for a broad range of other species[1,2]. These seaweeds are attracting considerable interest as aquaculture crops as a consequence of their capacity to rapidly produce large quantities of biomass under sustainable culture conditions that do not require the use of arable terrestrial land nor freshwater resources[3]. Brown algae are also of fundamental interest as they represent the third most complex lineage of multicellular organisms after animals and land plants, to which there are only very distantly related, being members of the Stramenopile supergroup[4]. Despite these highly interesting features, the basic biology of these organisms is still quite poorly understood. The Phaeoexplorer project (https://phaeoexplorer.sb-roscoff.fr/home/), funded by the France Genomique large-scale sequencing program and coordinated by one of the two host laboratories in Roscoff, aims to address this knowledge gap using a genomics approach. Phaeoexplorer has recently completed the sequencing and annotation of 66 complete genomes, corresponding to 46 species of brown algae (Phaeophyceae) and closely-related sister species.

Genome sequencing has detected the presence of integrated viral DNA in the genomes of diverse eukaryotes and this integrated DNA is thought to play an important role in genome evolution[5,6]. In most cases, integration of viral sequences is thought to involve non-specific processes but a small number of eukaryotic viruses are lysogenic, so that integration into the genome of their host is actually part of the viral life cycle. This is the case for the phaeoviruses that infect brown algae[7,8]. The phaeoviruses are members of the phylum Nucleocytoviricota and the members of this phylum, known as nucleo-cytoplasmic large DNA viruses (NCLDVs), are remarkable for the large sizes of their genomes, often extending to several hundred kilobases and containing large numbers of genes. The large size of NCLDV genomes, and the diversity of genes they contain, means that they have a strong potential to influence the evolution of their host's genomes, particularly in the case of phaeoviruses, which spend a significant part of their life cycle as integrated sequences.

Objectives and relevance of the project to the call

The objective of this PhD project is use the newly-available Phaeoexplorer genome resource to analyse the role that integrating viral genomes have played in the evolution of brown algal genomes. The research proposed is relevant to axe 1 of the Institut de l'Océan because it aims to identify genomic processes that were important for the emergence of the brown algae over evolutionary time as key components (keystone primary producers) of many marine ecosystems. The project will also improve our knowledge of organisms that are of fundamental importance in understanding the resilience of coastal ecosystems in the face of climate change[1,2] (axe 2).

Description of the project

A pipeline has been developed to detect viral genes in the Phaeoexplorer genomes and application of this pipeline has identified a large number of diverse inserted viral sequences (referred to as endogenous viral elements or EVEs) in the algal genomes (postdoctoral fellowship of Dean Mckeown, October 2019 to February 2022). Analysis of the EVEs indicates that they include both complete, inserted viral genomes (proviruses) and fragments of viral genomes (small EVEs) that probably correspond to ancient insertions that have partially degenerated. There are also indications that some viral genes have been or are being assimilated into the alga host gene repertoire (for example, viral sequences that would normally be monoexonic and silent in the host genome acquiring introns and becoming transcriptionally active).

The aim of this PhD project is to analyse the EVEs in the Phaeoexplorer genomes to further our understanding of the interaction between the viruses and their algal host and to understand how the integrated viral sequences impact the evolution of the host genomes. Analysis of EVE diversity will be carried out by constructing phylogenetic trees using conserved genes such as DNA polymerase B and previously characterised NCLDV sequences from public databases such as the Reference Viral Database (RVDB). Phylogenetic groups will be defined based on these tree with the aim of identifying new viral taxa. The host specificity of taxonomic groups

of viruses will be analysed based on the dataset of EVEs identified across the entire brown algal genome dataset.

The EVEs will also be analysed to understand the fates of inserted sequences, in particular the relative roles of degeneration and assimilation of genes by the algal host. Characteristics associated with the former include provirus fragmentation, transposon insertion and gene inactivation (e.g. frame shift mutations), whereas characteristics of the latter include intron acquisition and activation of transcription. The presence of many small EVEs in the Phaeoexplorer genomes indicates that inserted viral genomes become fragmented over time and analyses will be carried out to determine whether this fragmentation is simply due to degeneration of inserted viral genomes or whether this may also be one pathway for recruitment of viral genes by the host. For example, analyses will be carried out to determine whether genes in small EVEs are maintained by purifying selection and comparative genomics will be used to determine whether specific gene classes tend to be conserved in small EVEs across species.

Comparative genomics will also be used to carry out a large-scale analysis of viral integration sites to determine if these are conserved between genomes, either in terms of position or at the sequence level. Analysis of insertion sites will need to take into account the (unpublished) observation that EVEs, which are usually identified as clusters of monoexonic, silenced viral genes, may be flanked by expressed, multiexonic genes that appear to be of viral origin. This phenomenon has been observed for the provirus in the reference *Ectocarpus* genome (strain Ec32) and is of particular interest because it indicates a possible mechanisms of recruitment of viral genes by the host, i.e. by encroachment of host chromatin into the border regions of EVEs. The analysis of insertion site will therefore be strongly linked to the analyses aimed at identifying and quantifying recruitment of viral genes by the algal host. Genomes corresponding to 15 different species from the genus *Ectocarpus* will be particularly useful for comparative genomics, allowing precise mapping of EVE borders and identification of newly assimilated viral genes using genome synteny.

In conclusion, the analyses proposed for this PhD project will address several important outstanding questions including How diverse are the phaeoviruses? How do phaeoviruses insert into their algal hosts genomes? What is the long-term fate of proviruses after they have inserted into a genome? Are inserted viral genomes a source of important new genes via horizontal gene transfer (HGT) and, if so, are these genomes an important source of HGTs?

Supervision

The PhD project will be supervised jointly by Mark Cock of the UMR 8227 (Station Biologique de Roscoff) and Erwan Corre of the Analysis and Bioinformatics for Marine Science (ABiMS) platform (FR2424, Station Biologique de Roscoff; http://abims.sb-roscoff.fr/). Both supervisors will participate in all aspects of the PhD project, including regular, joint meetings with the student and concerted help with addressing both scientific and technical aspects of the project. In terms of complementarity, Mark Cock will provide input regarding brown algal biology and genomics, whereas Erwan Corre will provide high-level guidance for the implementation of bioinformatic approaches and coding.

Candidate

The PhD project will be carried by Patrick Jacques, who is currently working on a Masters 2 project looking at a specific aspect of brown algal EVEs, the presence of EVEs corresponding to small viral genome fragments. He is studying at the origins and evolutionary fates of these elements, which may represent a privileged intermediate for endogenisation of viral sequences by the brown algal hosts. The Masters 2 project will provide a solid basis for the broader analyses proposed for this PhD project. The current Masters 2 project is supervised by Mark Cock and Erwan Corre and financed by the Institut de l'Océan.

Feasibility

The co-supervision proposed for this project will ensure that the student has access to all of the data and facilities necessary to carry out the project. Mark Cock is coordinating the Phaeoexplorer project, which is being carried out in collaboration with Genoscope and involves a large, international consortium for the analysis of the Phaeoexplorer genome data (38 institutes in 14 countries on four continents). Erwan Corre leads

the Analysis and Bioinformatics for Marine Science (ABiMS; http://abims.sb-roscoff.fr/) platform within the FR2424, which is part of the French Bioinformatics Institute (IFB) and has extensive experience with providing bioinformatic support to diverse scientific projects both within and outside Roscoff Biological Station and with providing the calculation and storage environment for large-scale bioinformatics projects.

The student will have full, pre-publication access to the Phaeoexplorer genome dataset and will carry out his analyses using the extensive server facilities available in Roscoff (http://abims.sb-roscoff.fr). The PhD project will build on the work carried out by the postdoctoral fellow Dean Mckeown (also co-supervised by Mark Cock and Erwan Corre), which involved identification and manual validation of the large number of EVEs in the Phaeoexplorer genomes. The PhD project will therefore be based on solid preliminary work and we therefore estimate the risk attached to the project to be minimal. The student will benefit from interactions with the extensive international consortium that has been formed around the Phaeoexplorer project, in particular from a collaboration with Declan Schroeder at the University of Minnesota (https://www.virology.umn.edu/bio/virology/declan-schroeder) who is providing extensive expertise in virology, including for phaeoviruses. Dean Mckeown is currently in Declan Schroeder's laboratory and will continue to work on the Phaeoexplorer dataset and provide support for the PhD student's analyses. The project will involve bioinformatic analysis of an existing dataset and no wet laboratory experiments are planned.

Timeline

The first aim of the thesis (months 1-8) will be to fully characterise the recently established EVE dataset for the Phaeoexplorer genomes, including phylogenetic analysis and comparative analysis of the gene contents of EVEs across genomes and across species. The project will then focus on using the Phaeoexplorer data to characterise the entire life cycle of an inserted viral genome (months 9-30), including analysing insertion sites and understanding what happens to inserted genomes after insertion in the long term, particularly provirus fragmentation, viral gene and genome degradation (provirus fragmentation, relaxation of selection on coding regions, accumulation of mutations and transposons, *etc*.). A particular focus of the work during this period will be to identify and quantify events representing recruitment of viral genes by the algal host (i.e. virus to alga HGT) and to understand what impact this process has had on the evolutionary history of the algal host. The final six months of the thesis (months 31-36) will be used to complete specific aspects of the study and write up the thesis for the defence.

References

1. Klinger, T. (2015). The role of seaweeds in the modern ocean. **Perspect Phycol** 2**,** 31–39
2. Ortega, A., Geraldi, N. R., Alam, I., Kamau, A. A., Acinas, S. G., Logares, R., Gasol, J. M., Massana, R., Krause-Jensen, D. & Duarte, C. M. (2019). Important contribution of macroalgae to oceanic carbon sequestration. **Nature Geoscience** 12**,** 748–754
3. Cai, J., Lovatelli, A., Aguilar-Manjarrez, J., Cornish, L., Dabbadie, L., Desrochers, A., Diffey, S., Garrido Gamarro, E., Geehan, J., Hurtado, A., Lucente, D., Mair, G., Miao, W., Potin, P., Przybyla, C., Reantaso, M., Roubach, R., Tauati, M. & Yuan, X. (2021). Seaweeds and microalgae: an overview for unlocking their potential in global aquaculture development. **FAO Fisheries and Aquaculture Circular No. 1229. Rome, FAO** doi:10.4060/cb5670en
4. Coelho, S. & Cock, J. (2020). Brown algal model organisms. **Ann Rev Genet** 54**,** 71–92
5. Moniruzzaman, M., Weinheimer, A. R., Martinez-Gutierrez, C. A. & Aylward, F. O. (2020). Widespread endogenization of giant viruses shapes genomes of green algae. **Nature** 588**,** 141–145
6. Feschotte, C. & Gilbert, C. (2012). Endogenous viruses: insights into viral evolution and impact on host biology. **Nat Rev Genet** 13**,** 283–296
7. Delaroque, N., Maier, I., Knippers, R. & Müller, D. (1999). Persistent virus integration into the genome of its algal host, *Ectocarpus siliculosus* (Phaeophyceae). **J Gen Virol** 80 (Pt 6)**,** 1367–70
8. Cock, J. M. *et al.* (2010). The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. **Nature** 465**,** 617–621