OPEN LETTER

The Aquatic Symbiosis Genomics Project: probing the evolution of symbiosis across the tree of life [version 1; peer review: awaiting peer review]

Victoria McKenna, John M. Archibald, Roxanne Beinart, Michael N. Dawson, Ute Hentschel, Patrick J. Keeling, Jose V. Lopez, José M. Martín-Durán, Jillian M. Petersen, Julia D. Sigwart, Oleg Simakov, Kelly R. Sutherland, Michael Sweet, Nick Talbot, Anne W. Thompson, Peter W. Harrison, Jeena Rajan, Guy Cochrane, Matthew Berrian, Mara Lawniczak, Mark Blaxter

Abstract
We present the Aquatic Symbiosis Genomics Project, a global collaboration to generate high quality genome sequences for a wide range of eukaryotes and their microbial symbionts. Launched under the Symbiosis in Aquatic Systems Initiative of the Gordon and Betty Moore Foundation, the ASG Project brings together researchers from across the globe who hope to use these reference genomes to augment and extend their analyses of the dynamics, mechanisms and environmental importance of symbiosis. Applying large-scale, high-throughput sequencing and assembly technologies, the ASG
collaboration will assemble and annotate the genomes of 500 symbiotic organisms – both the “hosts” and the microbial symbionts with which they associate. These data will be released openly to benefit all who work on symbiosis, from conservation geneticists to those interested in the origin of the eukaryotic cell.

**Keywords**
Symbiosis, Marine, Freshwater, Genome Sequencing, Collaboration, Open Science
The genomics of symbiosis

Symbiosis, the living together of distinct organisms (Archibald, 2014; Oulhen et al., 2016), describes a spectrum of relationships from mutualistic to parasitic, and from obligate to temporary. Symbiosis has been and is fundamental to the evolution of life on Earth, from the deep origins of the eukaryotic cell and photosynthetic eukaryotes, through to the recent emergence of new partnerships. The power of symbiosis arises from the ability of the joint organism to draw from the independent, billion-year evolutionary histories of both partners. Symbiosis is a fact of life – it has arisen many, many times and new symbioses are constantly evolving (Figure 1). In this era of rapid climate change and biodiversity loss, many key-stone symbiotic systems are threatened, and their loss imperils the ecosystems they support.

Well-known mutualist symbioses permit colonisation of otherwise inaccessible habitats, are critical to ecosystem functioning, and support marine and freshwater diversity. For example, coral reefs, built through a photosymbiotic association between cnidarians and dinoflagellate algae (Weis, 2019), create biodiversity hotspots which house upwards of 25% of all described species in the oceans. The dominant animals colonising deep-sea hydrothermal vents are nutritionally dependent on chemosymbiotic associations with bacteria (Roeselers & Newton, 2012), allowing them to thrive in the food-limited dark ocean. For these symbioses, the biological fitness consequences

---

**Figure 1.** The phylogenetic diversity of eukaryotic symbioses. Symbiotic taxa, and Aquatic Symbiosis Genomics target species, are found across the diversity of the eukaryotic tree of life. Taxa highlighted with blue boxes include ASG targets. Within the tree, the small cartoons indicate the major event of plastid acquisition through symbiosis with a cyanobacterium (in the Archaeplastida; blue cell engulfed) and the several events of secondary and tertiary plastid acquisition in other lineages. Illustration by John Archibald and Mark Blaxter.
are largely understood, but in many less well-known symbioses, such as those between sponges and their bacterial collaborators, or partnerships in the diverse world of single celled eukaryotes, the basis of the relationships are not known in any detail.

**The aquatic symbiosis genomics project will transform symbiosis research**

The Gordon and Betty Moore Foundation has created a major funding initiative focused on investigating the biology of symbiosis in marine and freshwater ecosystems (see *Symbiosis in Aquatic Systems Initiative*). To support this global initiative, the Aquatic Symbiosis Genomics project (ASG; see *Aquatic Symbiosis Genomics Project – Wellcome Sanger Institute*) plans to generate high-quality genome sequences from a wide range of symbiotic systems. Our focus is on symbioses involving at least one microbial partner, and where there is likely to be co-evolving interplay between the species involved.

Like a symbiotic organism, the ASG project is more than the simple sum of its parts. ASG will merge the decades of ecological, evolutionary, taxonomic, and experimental expertise of researchers from diverse backgrounds with the decades of genomics experience of the Wellcome Sanger Institute. ASG works on a hub and spokes model, where communities of researchers nucleated on specific questions and/or species systems have come together as hubs to propose sets of taxa for sequencing (Table 1). These (currently) total ~450 distinct symbiotic organisms from the open ocean, the deep sea, coastal, littoral, and freshwater ecosystems, which are expected to include over 1000 nominal species of hosts and symbionts. The ASG target list includes species representing many phyla of animals, protists, algae and fungi, and encompasses ancient and recently-evolved partnerships.

The hub partners have defined the major scientific questions they wish to explore, and will source and identify specimens that will deliver answers. ASG follows an ethical code of sampling practice, avoiding overcollection and respecting local and international laws and protocols, especially as ASG will be sampling from endangered ecosystems and in some cases endangered species. The project participants are fully committed to the Convention on Biological Diversity Nagoya Protocols on Access and Benefit Sharing, and only samples where express permission has been obtained will be sourced and sequenced. Samples may come from the wild, from mesocosms and aquaria, from explant lab cultures or from culture collections.

Genome sequencing and assembly will be delivered by the Tree of Life programme at the Sanger Institute using pipelines being developed for the *Darwin Tree of Life* and other major biodiversity genomics projects. Genomes will be assembled, annotated and released openly through the European Bioinformatics Institute (*EMBL-EBI*).

### Table 1. Aquatic symbiosis genomics project hubs.

<table>
<thead>
<tr>
<th>Lead researcher*</th>
<th>Project Title (short)</th>
<th>Major taxa represented</th>
<th>Hosts</th>
<th>Symbionts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Archibald</td>
<td>New symbioses in single-celled eukaryotes</td>
<td>Amoebozoa, Dinophyceae, Diplonemea (Euglenozoa), Haptophyta, Ochrophyta</td>
<td>Bacteria, Kinetoplastea, Ochrophyta</td>
<td></td>
</tr>
<tr>
<td>Beinart, Petersen, Sigwart</td>
<td>Molluscan symbioses</td>
<td>Mollusca</td>
<td>Arthropoda, Bacteria, Chlorophyta, Cnidaria, Dinophyceae, Platyhelminthes, Florideophyceae</td>
<td></td>
</tr>
<tr>
<td>Dawson, Sutherland, Thompson</td>
<td>Pelagic symbioses</td>
<td>Acoela, Ctenophora, Cnidaria, Tunicata</td>
<td>Bacteria, Chlorophyta, Dinophyceae, other Alveolata</td>
<td></td>
</tr>
<tr>
<td>Hentschel</td>
<td>Sponge symbioses</td>
<td>Porifera</td>
<td>Bacteria, Archaea, Viruses, <em>Symbiodinium</em> (Dinophyceae) and others</td>
<td></td>
</tr>
<tr>
<td>Keeling</td>
<td>Symbiosis in ciliates</td>
<td>Ciliophora</td>
<td>Archaea, Bacteria, Chlorophyta, Ciliophora, Dinophyceae</td>
<td></td>
</tr>
<tr>
<td>Lopez</td>
<td>Metazoan photosymbioses</td>
<td>Acoela, Cnidaria, Mollusca, Porifera, Tunicata</td>
<td>Bacteria, Chlorophyta, Cnidaria, Dinophyceae, Haptophyta, Myzozoa</td>
<td></td>
</tr>
<tr>
<td>Martín-Durán</td>
<td>Annelid chemosymbioses</td>
<td>Annelida</td>
<td>Bacteria, Archaea</td>
<td></td>
</tr>
<tr>
<td>Simakov</td>
<td>Cephalopod symbioses</td>
<td>Mollusca</td>
<td>Bacteria, Archaea</td>
<td></td>
</tr>
<tr>
<td>Sweet</td>
<td>Coral symbioses</td>
<td>Cnidaria</td>
<td><em>Symbiodinium</em> (Dinophyceae)</td>
<td></td>
</tr>
<tr>
<td>Talbot</td>
<td>Marine lichens</td>
<td>Fungi</td>
<td>Bacteria, Chlorophyta, ascomycete Fungi, Ochrophyta</td>
<td></td>
</tr>
</tbody>
</table>

* see author list for affiliations.
Sequencing symbionts: from sample to openly accessible genome assembly

Each ASG Hub (Table 1) has defined a set of taxa that it will sample for sequencing. We will sequence from single eukaryotic host specimens or clonal cultures rather than bulk samples whenever possible. While this can limit the mass of DNA and RNA available for sequencing, it has the very strong benefit of reducing allelic sequence complexity and enabling assembly. Importantly, we do not require that the symbiotic partners are separated before sequencing, as we will separate the host and symbiont genomes bioinformatically during assembly (Challis et al., 2020).

Each sample is formally identified and associated with rich metadata describing its collection location and other environmental features. We collate and validate these metadata through the COPO biodiversity data brokering system. Samples are shipped to the Sanger Institute for long DNA and RNA extraction and sequencing, with particular focus on low-input methods. We are generating a combination of long read and long range genomic data. For long reads we primarily use the Pacific Biosciences Sequel Ile circular consensus sequencing approach to generate high fidelity (HiFi) reads in the 15 to 20 kilobase range, and include Oxford Nanopore Technologies long reads where needed. For long range data we use chromatin conformation capture sequencing (known as Hi-C). These long range data generate important information that link sequences within chromosomes and organelles in the multi-kilobase to megabase range and will allow us to disentangle genomes from different species. The joint transcriptome of the symbioses will be sampled using RNA-Seq, both on Illumina short read and Pacific Biosciences long read platforms.

We have strong expectations as to what we should find in the sequence data, and what we should be assembling, but biology is full of exceptions and surprises and organisms taken from the wild are frequently found in association with other symbionts. Each symbiosis contains a community of genomes that can be viewed as a low complexity metagenome: the “host” genome and the genomes of its organelles (mitochondrion and in some cases plastid), the symbiont genome (which if it is eukaryotic contains one or more organelar genomes) and the genomes of other commensals and symbionts. We separate data into presumed organismal and organelar subsets and assemble each independently. First we identify taxonomically informative marker loci, such as small subunit ribosomal RNAs (organelar 12S, prokaryotic 16S and eukaryotic 18S), cytochrome oxidase I genes, and ribulose-1,5-bisphosphate carboxylase-oxygenase genes, in the HiFi reads and primary assembly. These tell us which taxa are likely to be present and thus which genomes we should expect to assemble. To separate the data we use intrinsic features (GC and tetranucleotide composition, read coverage, coding capacity), sequence similarity to known genomes, and Hi-C linkage information. Binning contigs and their constituent reads into distinct subsets facilitates complete assembly of each organismal and organellar genome (Challis et al., 2020; Kumar & Blaxter, 2011). We aim to automate this cobiont identification and binning process, as it will be of utility in analyses of all tree of life genomes: many specimens harbour parasitic and other cobionts. Given 25- to 30-fold genome coverage in HiFi reads for each symbiont partner, we expect to generate primary assemblies with contig N50s in the multi-megabase range. The Hi-C data are used to scaffold these contigs into near-chromosomal pseudomolecules.

For each symbiotic system we will then curate the assemblies to improve accuracy (Howe et al., 2021) with particular attention to correct scaffolding of nuclear chromosomes and circularisation of organellar and prokaryotic genomes, and identification of remaining complex and unresolvable repetitive regions (such as ribosomal RNA and centromeric repeats). We aim to achieve or exceed the latest Earth BioGenome Project (Lewin et al., 2018) assembly standards. Curated assemblies and all raw data will be submitted to the European Nucleotide Archive (ENA) (Harrison et al., 2021) and from there to the rest of the International Nucleotide Sequence Database Consortium for immediate open release. The genomes will be annotated using the RNA-Seq transcriptomic data binned by species, and the annotations released openly. We have developed an ASG-specific data portal that collates all of the data generated by the project and promotes analysis. The Aquatic Symbiosis Genomics project relies on engagement and support from the whole of the Tree of Life production genomics team and of many colleagues who are participants in the ten Hubs. Each symbiotic system will be the subject of an open access publication, a Genome Note, that credits the full team that generated the assemblies, from collectors to annotators (Threlfall & Blaxter, 2021).

Building an aquatic symbiosis genomics community

The ASG project aims to generate a lasting resource in terms of the ~1000 genomes involved in ~500 symbiotic systems. To ensure this resource results in a flourishing ecosystem of postgenomic research, we are building community and expertise through a parallel programme of training and mentoring in genomics and bioinformatics. In collaboration with Wellcome Connecting Science and The Carpentries, the ASG project will deliver intensive and extensive collaborative training and investigative informatic analysis of symbiont genomes, to build collective genomics and bioinformatics capacity in the symbiosis community. Training will include core informatics, coding, and reproducible science, as well as deeper analytical dives into co-evolving genomes, detailed genome annotation, and prediction of the metabolic underpinnings of symbiotic cooperation.

Just as reefs built by corals and their symbiotic algae allow an exuberant and diverse ecology to thrive, the ASG project will build a lasting genomic foundation for flourishing and diverse analyses of symbiosis. Many of the fish that throng around coral reefs are open spawners, and their larvae spend their first weeks in the open ocean. They are recruited back around coral reefs because they can hear and smell it: the chatter generated by a healthy reef attracts, recruits, and builds the reef community (Gordon et al., 2019). Much like a healthy reef, our hope is that the high quality genomes we produce will generate the chatter that attracts new researchers and provides a
foundation for growth of fundamental research on the nature of symbiosis and conservation of habitats where symbioses abound.

**Data availability**
No data are associated with this article. ASG data will be released openly in the European Nucleotide Archive.

---

**Acknowledgements**
We thank Jonathan Threlfall for assistance with manuscript editing. This research was funded in part by Wellcome [grant 206194]. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

---

**References**


---

Page 6 of 6