A workflow for sequencing, annotating and curating aquatic symbiomes

Camilla Eldridge[1*], Michael Paulini[1*], Caroline Howard[1], Graeme Oathley[1], Radka Platte[1], Catherine McCarthy[1], Nancy Holroyd[1], Nicola Chapman[1], Neha Wali[1], Haoyu Niu[1], Peter Harrison[1], Alexey Sokolov[1], Talal Ibrahim[1], Claudia C. Weber[1], Emmelien Vancaester[1], Marcela Uliano-Silva[1], Ksenia Krasheninnikova[1], Eerik Aunin[1], Shane McCarthy[1], Will Chow, Ying Simms[1], James Torrance[1], Alan Tracey[1], Sarah Pelan[1], Joana Collins[1], Damon-Lee Pointon[1], Jonathan Wood[1], Kerstin Howe[1], Michael Sweet[2], Anne Thompson[3], Felipe Porto[4], Senjie Lin[4] Patrick Keeling[5], Helena Villela[6], Raquel Peixoto[6] , Victoria McKenna[1], Mara Lawniczak[1], Mark Blaxter[1], and members of the ASG consortium (see https://doi.org/10.5281/zenodo.5385621)

The Aquatic Symbiosis Genomics Project

The Aquatic Symbiosis Genomics Project, a collaboration between fourteen international research hubs, is sequencing more than 1000 species from 500 aquatic symbiotic systems to provide insights into complex ecological and evolutionary relationships. The goal is to generate annotated, gold-standard reference genomes for host and symbiont. An initial workflow, from sample preparation to *de novo* assembly, curation and annotation, has been developed to overcome the challenges associated with large-scale sequencing of diverse aquatic symbiotic systems.

Project status

- 92 species: arrived at Sanger
- 64 species: completed DNA extractions
- 39 species: in sequencing

Sequencing and primary assembly QC data are available on TOLQC https://tolqc.cog.sanger.ac.uk/

Conclusions and future work

- Aquatic symbiome samples have high levels of environmental contamination.
- Ongoing R & D needed for DNA extraction.
- Ongoing development of incorporating CobiontID into the workflow to enable separation of symbionts at read level.

Figure 1. STS and COPO sample management and tracking.

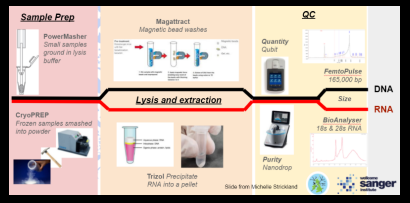
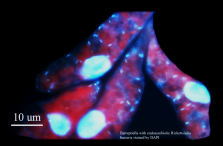
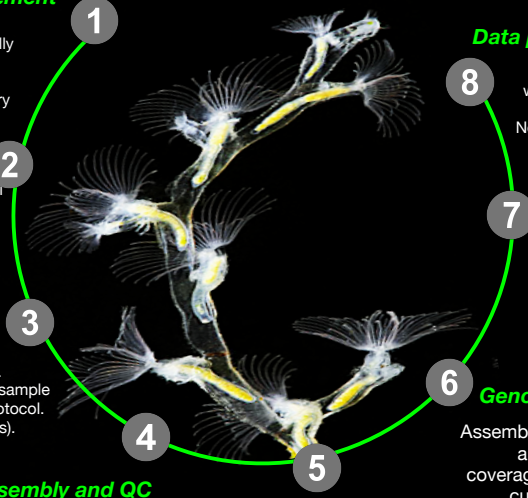


Figure 2. DNA extraction workflow.

Figure 3: The assembly pipeline.



ASG Workflow



1 Sample acquisition and management

Samples are collected by partners, snap frozen, and a detailed metadata manifest completed. We sample ethically and legally, including meeting all applicable regulatory compliance steps. These steps are completed before samples are shipped by specialist cold chain couriers in dry ice.

2 DNA extraction

Weighed subsamples are homogenised in lysis buffer/trizol with BioMasher & PowerMasher. A good quality extraction features High Molecular Weight DNA as the main component. The level of RNA degradation is recorded.

3 Sequencing

Long read data are generated from a low input library protocol followed by sequencing on the PacBio Sequel IIe. For HiC, DNA-protein complexes are cross-linked prior to sample fragmentation and DNA extraction using the Arima-HiC protocol. Libraries are sequenced on Illumina NovaSeq (PE150 reads).

4 Genome assembly and QC

Genomes are assembled from PacBio HiFi reads using HiCanu or Hifiasm and then purged to separate haplotypes. MitoHifi assembles the mitochondrion and eventual plastids. Salsa scaffolds genomes using HiC reads (Arima or Qiagen).

5 Host-symbiont separation

The assembled scaffolds are separated based on taxa using a combination of the NCBI BLAST based decontamination pipeline, the CobiontID workflow and BlobToolkit.

6 Genome curation and postprocessing

Assembled genomes are curated using HiC maps and gEVAL where HiC contact information, coverage and assembly gaps are used in manual curation to scaffold long read assemblies to chromosomal level. Curated genomes are post-processed and submitted to the ENA.

7 Genome Annotation

Curated genomes are annotated using the ASG annotation pipeline that generates a repeat library for each species and uses RNA-seq to train gene model predictors in Braker2. Initial annotations are viewable on the UCSC browser.

8 Data publication and accessibility

Annotated genomes will be hosted by Ensembl where final annotations including protein domain information will be accessible and a Genome Note describing the assembly will be published in Wellcome Open Research.

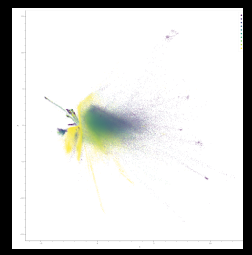


Fig 4: Read plot of Tetranucleotides coloured by coding density separating *Membranipora membranacea, a sea mat, from its cobionts.*

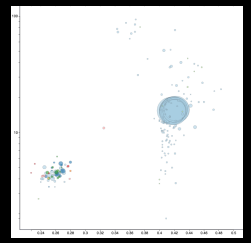


Figure 5: BlobToolKit plot (GC% x axis, coverage y axis) showing an assembly of a fish genome (upper right) also contains a myxozoan parasite (lower left)

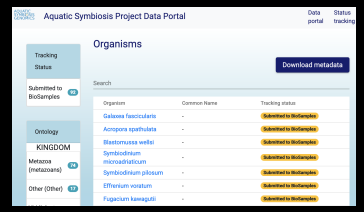


Figure 8: A dedicated data portal has been setup providing BioSample display, status tracking and taxonomy navigation.

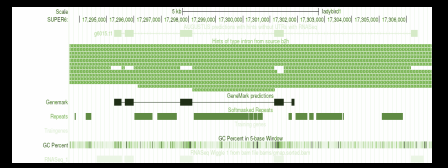


Figure 7: Viewing annotations

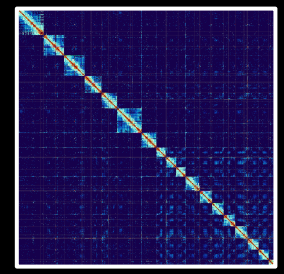


Figure 6. HiC contact map of a colonial sea squirt, *Aplidium turbinatum.*

Affiliations:
1 - Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, UK
2 - Aquatic Research Facility, Environmental Sustainability Research Centre, University of Derby, UK
3 - Department of Biology, Portland State University, USA
4 - Department of Marine Sciences & Institute of Systems Genomics, University of Connecticut, 1080, CT 06340, USA
5 - Botany Department, University of British Columbia, 3529-6270 University Boulevard, Vancouver, BC, V6T 1Z4, Canada
6 - Red Sea Research Center, King Abdullah University of Science and Technology, Bld 2, L2, Thuwal, SA

Images:
Coral, Nudibranch and ciliate images supplied by Patrick Keeling, Sea slugs (Elysia) taken by Patrick J.Krug, Eutreptiella figure provided by Senjie Lin. *Stentor pyriformis* image provided by Vincent Boudreau. Cnidarian photo taken from Fruška gora (Sremska Kamenica, Dunav), 2012. Copyright © Dragiša Savić, numbers taken from tagxOnsi on https://commons.wikimedia.org/wiki/File:Pink_sterile_sea_white_numbers.svg, Sea squirt image taken from https://gallerygenomes.org/specimen/chordata/sea-squirts-and-sea-pork/