The evolution and implications of life history in marine invertebrates: a genomic and transcriptomic analysis

PI: Richard K. Grosberg
Co-PI: Anne Frances Armstrong
Co-PI: Serena A. Caplins
UC Davis Center for Population Biology

1. General introduction

A core challenge in biology is to characterize the processes creating and maintaining the diverse life history patterns seen throughout nature. The life history of an organism has sweeping ecological and evolutionary implications by affecting dispersal ability, population connectivity, speciation and extinction rates, as well as the potential for local adaptation (Vermeij 1982; Jeffery and Emlet 2003). A striking array of life histories exist throughout the metazoans—from parasitic trematodes whose life cycle can include infecting 3 different hosts, to urchins with larval periods that can extend for over a year, all the way through direct developing organisms such as mammals which entirely lack a larval phase (Moran 1994). Existing model organisms are poor candidates for studying the evolution and consequences of varied life histories; most model systems were chosen specifically for their simple life cycles and therefore are not representative of the naturally occurring diversity. Understanding the important ecological and evolutionary consequences of the vast array of life histories that exist in nature requires a varied set of model organisms. Therefore, a primary goal of our research program is to develop genetic tools in a diverse set of promising novel systems that vary with respect to their life history strategies.

Marine invertebrates are exceptional systems for investigating the evolution of life histories due to the varied life cycles they display, even between closely related species (Strathmann 1978; Wray 1996; Collin 2004). The majority of marine invertebrates undergo indirect development—a life cycle including a distinct larval phase (Strathmann 1978). This larval period generally results in large, well-connected populations that display low genetic divergence even over great distances. However, numerous species in every metazoan phyla have reduced or altogether lost this larval phase (Moran 1994). Current research shows that this reduced larval phase results in smaller species ranges, higher rates of extinction and higher rates of speciation (Vermeij 1982; Jeffery and Emlet 2003). However many questions still remain—What genetic mechanisms underlie the evolution of alternate developmental modes? How can local adaptation occur in a species that disperses thousands of kilometers? Do larvae preferentially settle with or near related individuals creating fine scale genetic structure even in widely dispersing organisms? These outstanding questions have previously been unanswerable due to a lack of available genetic tools. However, due to the advent of next-generation sequencing technologies, we are beginning to address these questions, among others.

Our lab currently has several ongoing projects in a diverse suite of organisms spanning phyla that will move us closer to a comprehensive understanding of both the generation of varied life histories and their evolutionary implications.

1. A transcriptomic analysis of development in the Clypeasteroids

The order *Clypeasteroida* includes sand dollars and sea biscuits. This order is notable as it displays more documented variation in life history patterns than any other order in the echinoderms. Non-feeding larval development has evolved several independent times in this group from species with long-lived feeding larval stages. This project primarily focuses on two closely related sea biscuit species (*Clypeaster rosaceus* and *Clypeaster subdepressus*); one species possesses long-lived feeding larvae, while the other has a reduced larval period where larvae do not need to feed. These two species present a remarkable opportunity as they can hybridize, allowing us to test how developmental mode is inherited. We have studied the development of these two species and their hybrids (Armstrong and Lessios 2015). We are using RNAseq to investigate differential gene expression between larvae of both parental species and each reciprocal hybrid cross to determine which genes are involved in regulating each developmental mode.

Recently we received a grant from the National Science Foundation to extend this project to twelve other species in Clypeasteroid family, which possess a wide range of developmental types. We have collected eggs from each of these species and will sequence the maternally deposited mRNA in each species eggs. Maternal influences on development are striking throughout animals and this project will allow us to test the hypothesis that this maternal effect is largely driven by differences in maternally deposited mRNA. Echinoderms, and sand dollars in particular, have long been used as a model system for studying development; this work will greatly extend the genomic resources available in this group while testing a core hypothesis that could explain the widespread evolution of alternate developmental programs.

2. **The genetic basis of phenotypic plastic for larval developmental mode**

The sacoglossan slug *Alderia willowi* exhibits a great degree of plastic control over larval developmental mode. Each adult can shift from producing many small eggs that develop into feeding larvae to producing few large eggs that develop into non-feeding larvae. *A. willowi* varies which larval developmental mode is produced on a seasonal basis, though not without individual and population level variation. The ramifications of these different developmental modes are extensive, and include dispersal distance, individual fitness, offspring survival, population and genetic connectivity, local population abundance, local mating system, and genetic diversity. Yet, the genomic and transcriptomic basis of different developmental modes remains largely unknown. There are only a few species that exhibit both developmental modes

(feeding and non-feeding larvae) and it is through study of these species that the underlying genetic basis of egg size and developmental mode can be understood.

We will evaluate changes in gene expression (via RNAseq) that occur during the production of each larval type. We will do this by examining gene expression in the central nervous system and ovaries of adult slugs that produce small-egg egg masses and large-egg egg masses. This will allow us to determine the regulatory changes that pre-empt egg mass type. Furthermore, we will characterize maternally deposited mRNA in small egg versus large eggs, to determine whether transcripts differ by amount or composition. We are currently assembling (*de novo* and genome guided) transcriptomes from the dimorphic *A. willowi* with that of the congeneric species *A. modesta*. *A. modesta* occupies similar habitats as *A. willowi*, but does not exhibit plasticity for larval developmental mode, and solely produces egg masses containing small eggs that develop into feeding larvae. Species that are plastic or dimorphic for egg size and larval type are key to uncovering the genetic basis of these traits, and will enable comparative studies of many more species in the future.

3. **Population genetic structure of *Anthopleura* and their symbionts**

Anemones in the genus *Anthopleura* are found along a large span of the North American coast, from Alaska to Baja California, Mexico, encountering varied habitats both locally (within the intertidal zone) and regionally (across multiple biogeographic barriers). Much like corals, these anemones also host the photosymbionts *Symbiodinium* spp., which provide their hosts with sugars in exchange for metaboloic waste and a consistently bright environment. The interaction between the host and symbiont (together referred to as the "holobiont") is remarkably persistent in the face of extreme environmental variation.

The goal of this project is to characterize the role of local adaptation and phenotypic plasticity in the host and symbiont that allow the relationship to persist across a wide range of environments. To do this we are using RAD-seq to generate population genetic datasets that allow us to determine the population structure of both the host and symbiont. In addition, these data will allow us to detect signals of coevolution between the host and symbiont, potentially lending insight into the role of the host, symbiont or a combination of both partners in allowing the holobiont to persist across variable environments.

In the course of analyzing the RAD-seq data from the three symbiotic species of *Anthopleura* sp. on the Pacific coast of the United States, we have also detected introgression between the host species. To investigate how gene flow between host species might impact their relationship with the symbiont *Symbiodinium* sp. we are now sequencing and assembling the genomes of *A. sola* and *A. xanthogrammica*. Coupled with the already-sequenced genome of *A. elegantissima*, these genomes will allow us to determine the rates of introgression between species, if introgression between host species is adaptive, and if host genetic background influences the outcome of the host-symbiont interaction.

4. **The effect of mating system on population structure and life history variation in the genus *Nucella***

The gastropod genus *Nucella* displays an usually wide range of mating systems within and between species. Our lab has previously documented that females vary in the number of males they mate with prior to producing egg capsules. This results in capsules with individuals of varying levels of relatedness resulting in increased levels of sibling-sibling conflict. Egg capsules from mixed paternity have significantly higher variation in offspring size than those from a single father. This increased variation in offspring size may result in differences in offspring growth and survival.

We plan to develop microsatellites for 9 species of *Nucella* to characterize how variation in mating system affects genetic structure across populations. Over the past year, we have assembled genomes and created microsatellites for two species over. In the upcoming year we plan on expanding this work to an additional four species in the genus. We will develop microsatelites from illumina shot-gun sequence data that we will assemble into contigs. We also plan to use this to test if variation in mating system is involved in the transitions between life history traits among species of *Nucella*.

2. COMPUTATIONAL METHODOLOGY

2.1 RNAseq

For our projects on *Clypeasteroida* and *Alderia* we will use XSEDE for RNAseq methods. We will assemble 14 transcriptomes (twelve *Clypeasteroids* and two *Alderia*) de novo with the software package Trinity (Grabherr et al. 2011). The two *Alderia* transcriptomes will be assembled *de novo* (using Trinity) and genome guided (using Oasis on Comet), as we suspect there is high heterozygosity for these species. After assembling each transcriptome we will cluster similar contigs using CD-hit. CD-hit will be run multiple times on each transcriptome with different parameters to determine what the appropriate clustering level. We will use Trinotate to analyze and annotate each assembled transcriptome including the two Clypeaster transcriptomes assembled last year. Additionally, to aid in annotation, we will search for orthologous transcripts by mapping contigs with BWA from each Clypeasteroid transcriptome to the sea urchin genome, *Strongylocentrotus purpuratus*, and *Alderia* contigs to the partially assembled genome of the sea slug *Aplesia californica*. Putative *Alderia* orthologous will be identified in OrthoMCL. Lastly we will use BLAT to compare each assembled transcriptome to one another and determine orthologous transcripts.

After assemblies and annotation are complete, we will measure differential gene expression using RSEM. We expect to have 250 replicate libraries for our Clypeaster project, and 150 for Alderia resulting in 400 total runs of RSEM. Read counts will be analyzed in the R package DESeq for each experiment.

Program	# runs	SUs/run	Resource	Total SUs
Trinity	14	37,500	Bridges	525,000
Oasis	2	50,000	Comet	100,000
Trinotate	16	2,500	Bridges	40,000
CD-Hit	36	3,000	Bridges	108,000
RSEM	300	300	Bridges	90,000
BWA	16	100	Bridges	16,000
BLAT	24	5,000	Bridges	120,000
R for DESeq	16	500	Bridges	8,000
Data Storage	10 lanes Illumina	1TB	Bridges-Pylon	

2.2 RADseq

We are using RADseq methods on both the *Anthepleura* system. RADseq allows us to generate sequence data from a large number of individuals at a restricted set of loci. This project is designed to detect population genetic structure on multiple spatial scales. In order to do this, approximately 700 individuals will be genotyped along the Pacific coast of the United States and Baja Mexico. Sequences will be obtained from the *Anthepleura* holobiont so as to investigate population structure in both the symbiont and the host.

For our work in *Anthepleura*, we must first distinguish between sequences obtained from the anemone host and symbiont. We will do this by mapping holobiont sequences to both host and symbiont draft genomes using BWA. Mapping symbiont reads is complicated by potentially high divergence between the draft genome and the symbiont populations along the Pacific coast, necessitating the use of BLAST to query reads against a custom database made up of the currently sequenced apicomplexan genomes (the closest sequenced relatives to *Symbiodinium* spp), the host and symbiont draft genomes and the RefSeq bacterial database available through NCBI.

Program	# runs	SUs/run	Resource	Total SUs
BWA	96	100	Bridges	9,600
dadi	6	9,500	Bridges	57,000
Data Storage	4 lanes	614GB	Bridges	

2.3 Genome assembly and microsatellite development

Genomes will be assembled *de novo* for four species of *Nucella* and two species of *Anthopleura*. Each assembly will be run in velvet at two separate hash lengths to find the hash length that achieves the largest n50 contig size. We will search for microsatelites using msatCommander on the assembly with the greatest n50 contig size. We will map microsatellites

back to the genomes using BWA to determine how where they are located in our assemblies. Python scripts will be run to select specific putative microsatellites from the contigs, for which primers will be designed.

In the case of the two *Anthopleura* genomes, we will use velvet as described above to assemble Illumina data that we sequence with two insert sizes. We will subsequently increase the Illumina assembly scaffold lengths with PacBio reads using PBJelly.

Program	**# runs**	**SUs/run**	**Resource**	**Total SUs**
Velvet	12	90,000	Bridges	1,080,000
msatCommander	4	1,000	Bridges	4,000
PBJelly	2	60,000	Bridges	120,000
BWA	6	100	Bridges	6,000
Data Storage	6 lanes	410GB	Bridges-Pylon	

3. JUSTIFICATION OF REQUESTS

Our lab group has been using an allocation on XSEDE's Greenfield and Mason for the past year. During that time we have explored the methodology proposed here and have streamlined pipelines to effectively use our computational time. We calculated the number of hours these tasks have taken in Greenfield and converted SUs to what we will need in Bridges using the XSEDE SU converter (see code performance and scaling document).

Resource	**Total SUs requested**
Comet	100,000 SUs
Bridges Regular Memory	2,183,600 SUs

We have several terabytes of storage on external hard drives in our lab, as well as data backed up by each student. However, we do not have any type of storage system that is accessible online. As much of this work is collaborative, we need to have storage that will be accessible to all researchers conducting projects. After completing analyses and projects, the data will be stored long term on the resources that we have locally in our laboratory.

Resource	**Total TB requested**
Bridges-Pylon	2 TB

4. ADDITIONAL CONSIDERATIONS

4.1 Outreach and Training

This proposal will directly benefit five current PhD students and one post-doc in the Grosberg lab. Access to the XSEDE resources will not only help them complete their research but will also increase their computational knowledge. Other members of the Grosberg lab will also use this allocation including one permanent lab technician and three undergraduates. This allocation will provide transformative training in computational science for these developing researchers.

4.2 Funding and local computing resources

The above-described projects are funded by a National Science Foundation OCE grant as well as three separate Graduate Research Fellowships. However, our local resources are not sufficient to complete the work nor do we have access to any other supercomputing clusters. Our laboratory-computing environment consists of two PCs and two Mac's. This has been enough for some small projects, but as the amount of data being generated in next generation sequencing rapidly increases, the majority of our work is quickly becoming infeasible on these devices.