Okinawa Institute of Science and Technology

Graduate University

Thesis submitted for the degree

Doctor of Philosophy

Decoding and Analysis of the Crown-of-Thorns Starfish *Acanthaster planci* Genome

by

Kenneth William Baughman

Supervisor: Prof. Noriyuki Satoh

submit date: Friday, April 14, 2017

Declaration of Original and Sole Authorship

I, Kenneth William Baughman declare that this thesis entitled "Decoding and Analysis of the Crown-of-Thorns Starfish Acanthaster planci Genome" and the data presented in it are original and my own work. I confirm that:

- No part of this work has previously been submitted for a degree at this or any other university.
- References to the work of others have been clearly acknowledged. Quotations from the work of others have been clearly indicated, and attributed to them.
- In cases where others have contributed to part of this work, such contribution has been clearly acknowledged and distinguished from my own work.
- None of this work has been previously published elsewhere, with the exception of the following:

- **Baughman, Kenneth W,** Carmel McDougall, Scott F Cummins, Mike Hall, Bernard M Degnan, Nori Satoh, and Eiichi Shoguchi. 2014. "Genomic organization of Hox and ParaHox clusters in the Echinoderm, *Acanthaster Planci*." *Genesis* 52 (12): 952–58. doi:10.1002/dvg.22840.

- Oleg Simakov*, Takeshi Kawashima*, Ferdinand Marlétaz, Jerry Jenkins, Ryo Koyanagi, Therese Mitros, Kanako Hisata, Jessen Bredeson, Eiichi Shoguchi, Fuki Gyoja, Jia-Xing Yue†, Yi-Chih Chen, Robert M. Freeman Jr, Akane Sasaki, Tomoe Hikosaka-Katayama, Atsuko Sato, Manabu Fujie, **Kenneth W. Baughman**, Judith Levine, Paul Gonzalez, Christopher Cameron, Jens H. Fritzenwanker, Ariel M. Pani, Hiroki Goto, Miyuki Kanda, Nana Arakaki, Shinichi Yamasaki, Jiaxin Qu, Andrew Cree, Yan Ding, Huyen H. Dinh, Shannon Dugan, Michael Holder, Shalini N. Jhangiani, Christie L. Kovar, Sandra L. Lee, Lora R. Lewis, Donna Morton, Lynne V. Nazareth, Geoffrey Okwuonu, Jireh Santibanez, Rui Chen, Stephen Richards, Donna M. Muzny, Andrew Gillis, Leonid Peshkin, Michael Wu, Tom Humphreys, Yi-Hsien Su, Nicholas H. Putnam, Jeremy Schmutz, Asao Fujiyama, Jr-Kai Yu, Kunifumi Tagawa, Kim C. Worley, Richard A. Gibbs, Marc W. Kirschner, Christopher J. Lowe, Noriyuki Satoh, Daniel S. Rokhsar, and John Gerhart 2015. "Hemichordate Genomes and Deuterostome Origins." *Nature* 527 (7579): 459–65. doi:10.1038/nature16150. (*co-first authors)
- Hall, Michael R.*, Kevin M. Kocot*, **Kenneth W. Baughman*,** Selene L. Fernandez-Valverde, Marie E. A. Gauthier, William L. Hatleberg, Arunkumar Krishnan, Carmel McDougall, Cherie A. Motti, Eiichi Shoguchi, Tianfang Wang, Xueyan Xiang, Min Zhao, Utpal Bose, Chuya Shinzato, Kanako Hisata, Manabu Fujie, Miyuki Kanda, Scott F. Cummins, Noriyuki Satoh, Sandie M. Degnan and Bernard M. Degnan. (2017) "The crown-of-thorns starfish genome as a tool for biocontrol of a coral reef pest" *Nature* 544, 231–234 (*co-first authors)

- Signature:

Date: Friday, April 14, 2017

Abstract

Echinoderms are at the base of the deuterostome clade, yet have radial body plans, a water-vascular system, and exoskeletons. In order to investigate how genomes control development, I studied the "Crown-of-Thorns Starfish" (COTS) or *Acanthaster planci* genome. I made four discoveries from sequencing two COTS specimens, one from the Great Barrier Reef, Australia ('GBR') and the other from Okinawa, Japan ('OKI'). Separate 384 megabase (Mb) assemblies containing ~24,500 genes were generated. First, I discovered that both genomes displayed unexpectedly low heterozygosity; reciprocal BLAST alignment of scaffolds longer than 10 kilobases (Kb) revealed 98.8% nucleotide identity, consistent with a single pacific COTS clade undergoing a recent population expansion. Second, although the unique Hox gene order in sea urchins was hypothesized to be related to pentaradial body plans, I discovered that COTS Hox and ParaHox clusters resemble hemichordate and chordate clusters. The COTS Hox cluster shares with sea urchins the transposition of *even-skipped* (*Evx)*, as well as posterior Hox reorganization. I thus proposed an evolutionary scenario for how shuffling of the Hox cluster in urchins may have arisen. Third, recent studies show that hemichordates possess a deuterostome-specific cluster of transcription factors associated with development of pharyngeal gill slits. Although extant echinoderms do not have pharyngeal gill slits, I found the cluster in the COTS genome, supporting an ancient origin for pharyngeal gill slits as a deuterostome-defining morphological feature. Fourth, using systems biology notation, I mapped COTS candidate genes for 1-methlyadenine (1-MA)-mediated oocyte maturation. This thesis confirms that the high quality of the COTS genome is biologically significant, and amendable to future studies. Although COTS are famous for decimating coral reefs, this thesis shows that COTS can also be used for genomic and evolutionary developmental research.

Date: Friday, April 14, 2017

Acknowledgements: I would like to acknowledge assistance received in the following areas:

Designing the research: Prof. Nori Satoh and Dr. Eiichi Shoguchi

Executing the research: Dr. Eiichi Shoguchi, Prof. Chuya Shinzato, Kanako Hisata, Koichi Toda, Dr. Manabu Fujie, Miyuki Kanda, Prof. Oleg Simakov, Dr. Takeshi Takeuchi, and Keita Ikegami

Analyzing the data: Prof. Nori Satoh, Dr. Eiichi Shoguchi, Kanako Hisata, Dr. Konstantin Khalturin, Prof Chuya Shinzato, Prof. Oleg Simakov, and Yi-jyun Luo

Interpreting the data/research: Prof. Nori Satoh, Dr. Eiichi Shoguchi, Kanako Hisata, Prof Chuya Shinzato, Prof. Oleg Simakov, Yi-jyun Luo, and Dr. Ferdinand Marletaz

Writing, proofing, or copyediting the manuscript: Prof. Nori Satoh and Dr. Eiichi Shoguchi

Table of Contents

List of Figures

List of Tables

Chapter 1 : Introduction

1.1 Echinoderms, ambulacraria, and deuterostomes

The word ‘Echinodermata’ derives from the ancient Greek for ‘porcupine’ or ‘hedgehog’ (*ekhinos*) plus ‘skin’ (*derma*) and refers to the clade’s most obvious synapomorphy; adult calcium carbonate exoskeletons. The 5 non-extinct subphyla include: sea lilies or feather stars (Crinoidea), brittle stars (Ophiuroidea), sea stars (Asteroidea), sea cucumbers (Holothuroidea), and sea urchins (Echinoidea) (Figure 1.1a). Echinoderms are identified by their four major synapomorphies: 1) calcium carbonate exoskeletons, 2) water vascular systems, 3) controllable collagenous connective tissue, and 4) radial, often pentameric adult body plans (Figure 1.1b). Conversely, echinoderms share developmental traits with hemichordates, together forming the monophyletic clade ‘Ambulacraria’. Ambulacraria, along with Chordata composes Dueterostomia (Satoh 2016). Sea urchins in particular, have long served as a model system for studying deuterostome development and the evolution of the chordate body plan (Davidson 1997; De Robertis 2008; Davidson 2010; Sea Urchin Genome Sequencing Consortium et al. 2006).

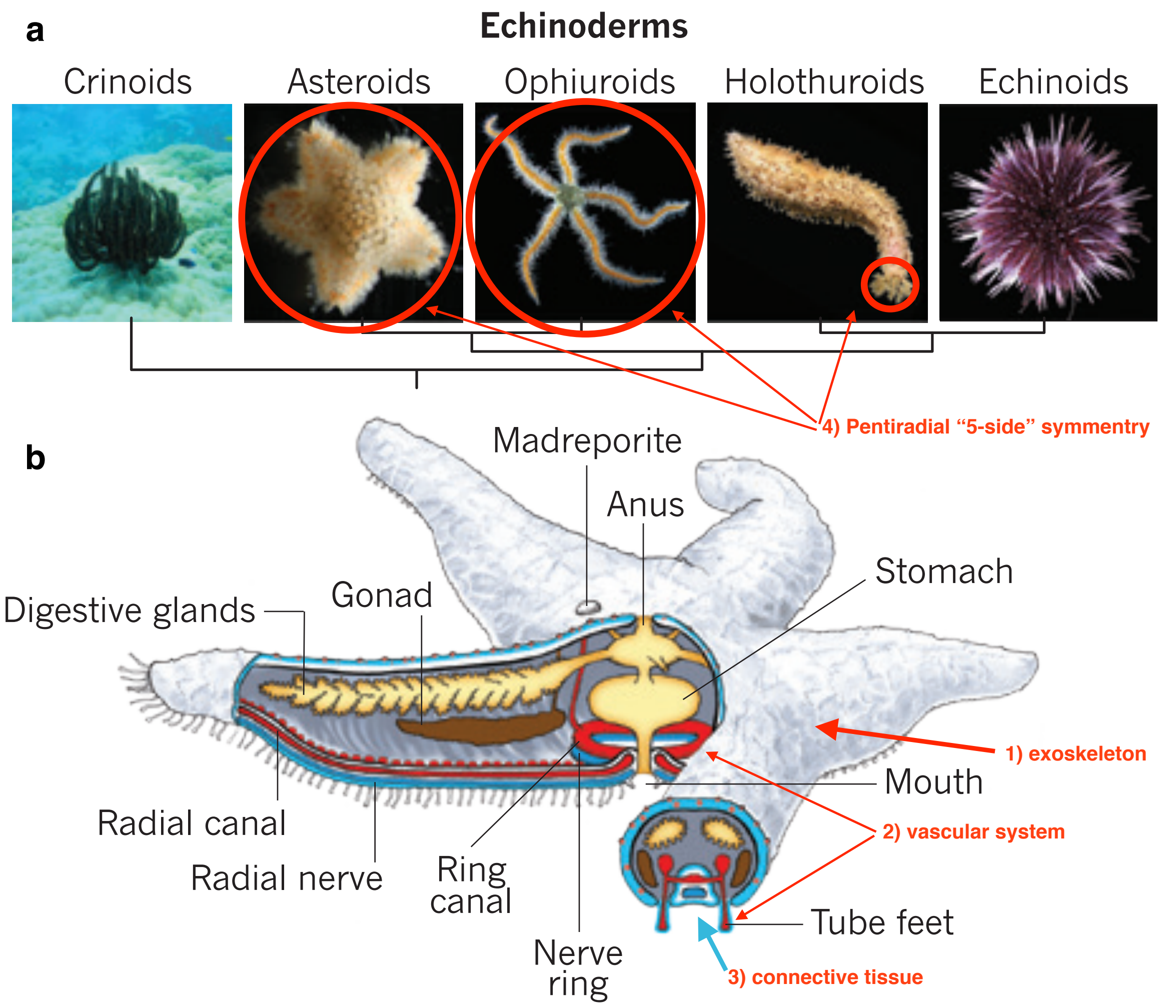


Figure 1.1. The 5 extant echinoderms and 4 echinoderm-defining synapomorphies.
(a) Examples from each extant class of echinoderm, with a current phylogenetic relation highlighted on the bottom. (b) Cross section of starfish, red text highlights the 4 main synapomorphies of echinoderms. Adapted from (Lowe et al. 2015).

Echinoderms share the first steps of embryological development with chordates in that their larva are bilateral, but after metamorphosis, adult echinoderms develop radial, generally pentameric, body plans. Any commonality between echinoderms, hemichordates and chordates, either with regard to genomic organization and synteny, or to developmental patterning, may have existed in the common ancestor of deuterostomes, and indeed bilateria itself (Lowe et al. 2015; Holland 2015). Classically, metazoans (e.g. kingdom Animalia) are divided between whether they have bilateral body plans (or not), and then by whether the first invagination (blastopore) of the developing embryo becomes a mouth (protostome) or anus (deuterostome) (Figure 1.2). Deuterostomes are then divided by those with a fish-like swimming larva (chordates), and those with bilateral larva that move using cilia (ambulacrarians) (Satoh et al. 2014). Though these divisions were initially observed and described by taxonomists and early embryologists more than a century ago, the advent of molecular techniques and genome sequencing has confirmed the genetic basis for these synapomorphies. Put more precisely, the functional significance of various developmental features that historically denoted clades, can now be interrogated at the genomic level (Satoh 2016).

At this early point, I wish to highlight a notion that will be revisited in the final chapter. Classical taxonomy and embryology have provided tremendous insights by describing morphology, and categorizing correspondingly. In our current molecular era, we have been able to further explore these relationships by identifying and comparing single genes, gene families, or gene clusters. Molecular phylogeny now allows us to make definitive statements about those classically described relationships between and across taxa. The notion is perhaps it is not these protein coding genes themselves that underlie the distinct morphology taxonomists have described. More specifically, the notion is that maybe the genetic control systems, the toolkit genes, the ‘programming’ layer (in contrast to the ‘data’ layer of protein coding genes), are what drives evolutionary divergence, speciation, and more simply, body plan divergence (Kirschner et al. 2006; Peter & Davidson 2015).

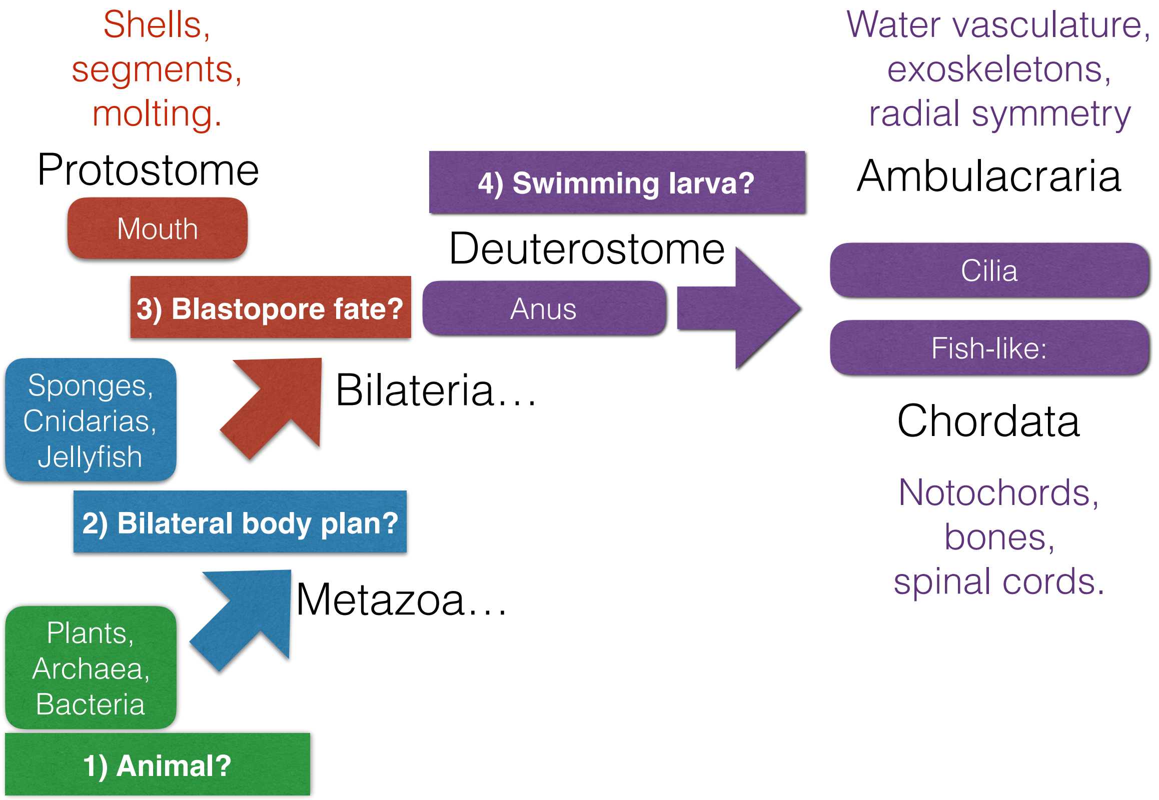


Figure 1.2. Major Phylogenetic Clades of Kingdom Animalia
A systematics-based representation for the definitions of Metazoa, Bilateria, Protostomia, Deuterostomia, Ambulacraria, and Chordata. The colored text highlights notable synapomorphies, from each clade.

1.2 The crown-of-thorns-starfish (COTS)

In the present study, I selected *Acanthaster planci,* commonly known as the Crown-of-Thorns-Starfish ("COTS") as the experimental system. COTS are one of the primary causes of coral reef devastation in the Indo-Pacific Oceans, largely due to population density fluctuations or aggregations, termed 'COTS outbreaks' (Sapp 1999). Over the past 50 years, *A. planci* have been the focus of more reef management efforts than any other marine species (Birkeland & Lucas 1990).

COTS are in the phylum Echinodermata, the class Asterodea, the order Valvitida, and the family Acanthasteridae (Mah & Blake 2012), shared with one sister species, *A. brevispinus,* to which COTS can hybridize (Lucas & Jones 1976). The name *Acanthaster planci* was given by Linnaeus in 1758 (Haszpruner & Spies 2014). "*Acanth*-" can be translated as "thorn", “*aster*” as "star", and “*planci*” is derived from the same root word as plankton, presumably a reference to the slow motility of the starfish. The common name for *A. planci* “Crown-of-Thorns” is a reference to their venomous spines and the crown placed on Jesus’ head during his crucifixion. Notably, the Japanese name for COTS is ‘onihitode’, which roughly translates as ‘demon starfish.’

COTS are one of the largest starfish species, with adults reaching up to 1.2 meters in diameter, over 70 kilograms in mass, and having up to 23 arms (Moran 1988). Adult starfish are generally 25-40 cm in diameter, have 10-15 arms covered with 2-4 cm spines that can range in color from orange or red, to yellow (Figure 1.3). The main body color is muted, generally brown, grey-green, or in some cases bluish or purplish. The spines are toxic to humans, and spine puncture results in rapid tissue inflammation, pain, and up to a week of nausea and vomiting. Saponins have been suggested to play a role in COTS toxicity, though this is an active research area (Komori 1997; Lee et al. 2013; Maoka et al. 2010; Lee et al. 2014). In 2012, the first recorded COTS-related fatality occurred when an Okinawa diver, who had previously been exposed 5 or 6 times, went into anaphylactic shock following a finger prick during a 20 meter dive (Ihama et al. 2014).

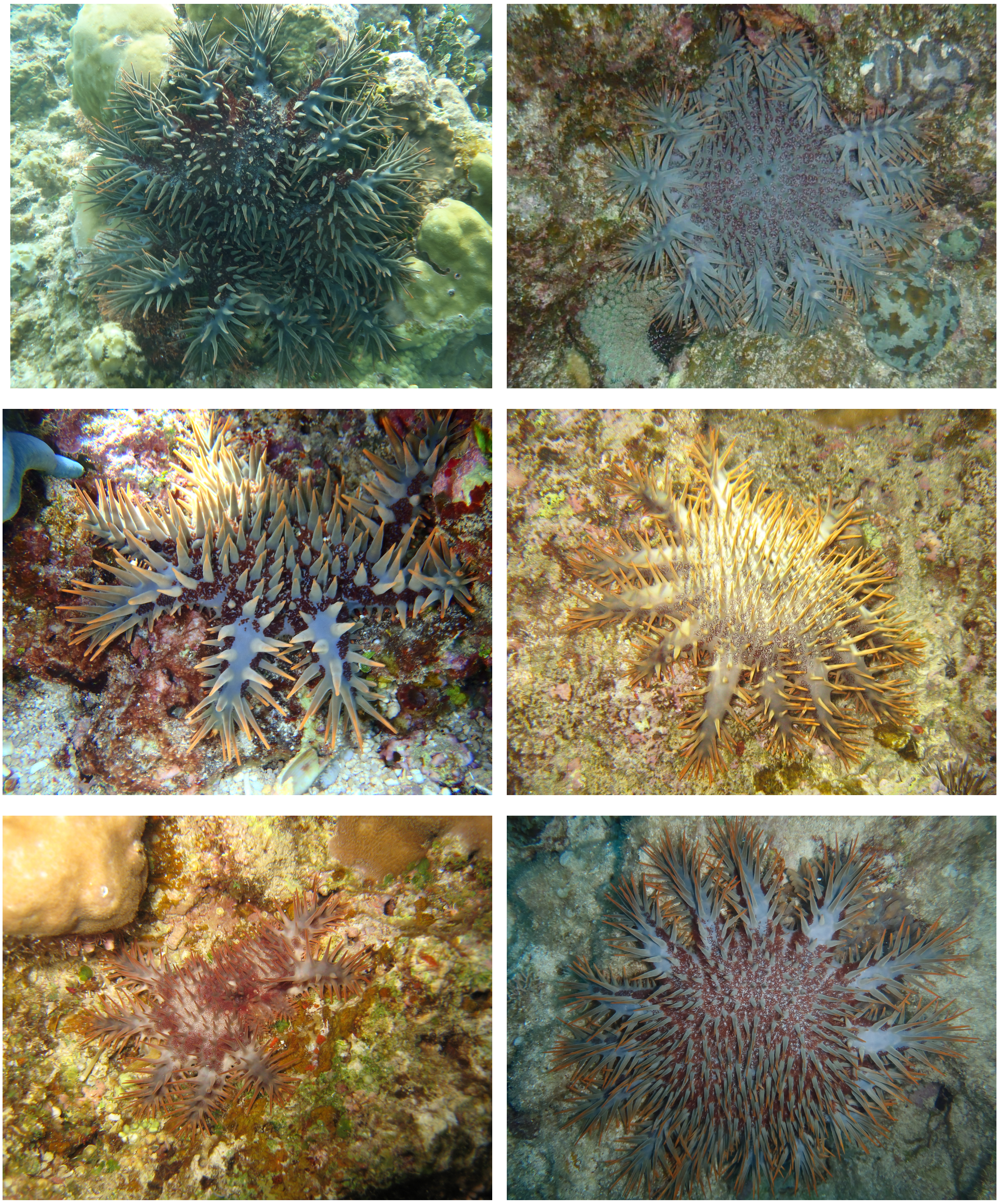


Figure 1.3. Six Crown-of-Thorns Starfish
Taken at 2-30 meter depth by the author, in Okinawa, Japan.

COTS generally spawn once per year, during 1-2 months in mid-summer, though in some locations they may remain fecund for longer periods of time (Moran 1988). The entire life cycle takes 2-4 years (Figure 1.4). Spawning occurs in early summer, which in Okinawa occurs in June and July, while on the Great Barrier Reef in the southern hemisphere, occurs in December and January (Moran 1988). Males and females release large quantities of sperm and egg into the water column, presumably in response to a spawning factor (Beach et al. 1975). Fertilized eggs then develop into blastula, gastrula, bipinnaria, and finally brachiolaria larva (Moran 1988). These larval stages are bilateral confirming evolutionary proximity of echinoderms to chordates, presumably in the form of a last common ancestor termed "urbilateria" (De Robertis 2008; Martindale & Hejnol 2009). After several weeks to months, free floating bilateral larva then settle to the sea bottom, and metamorphose into penta-radial juvenile starfish, which then transition into mature starfish that grow additional arms through an unknown mechanism. After 2 years, these juveniles become adult gamete-produce starfish that feed on corals (Moran & De'ath 1992). Progression through the *A. planci* life cycle may have temperature dependencies (Birkeland & Lucas 1990), and recent studies suggest that high levels of phytoplankton in the water column correlate with increased percentage of larval survival, which has been termed the "COTS Larval Hypothesis" (Fabricius et al. 2010; Wolfe et al. 2015; Uthicke et al. 2015).

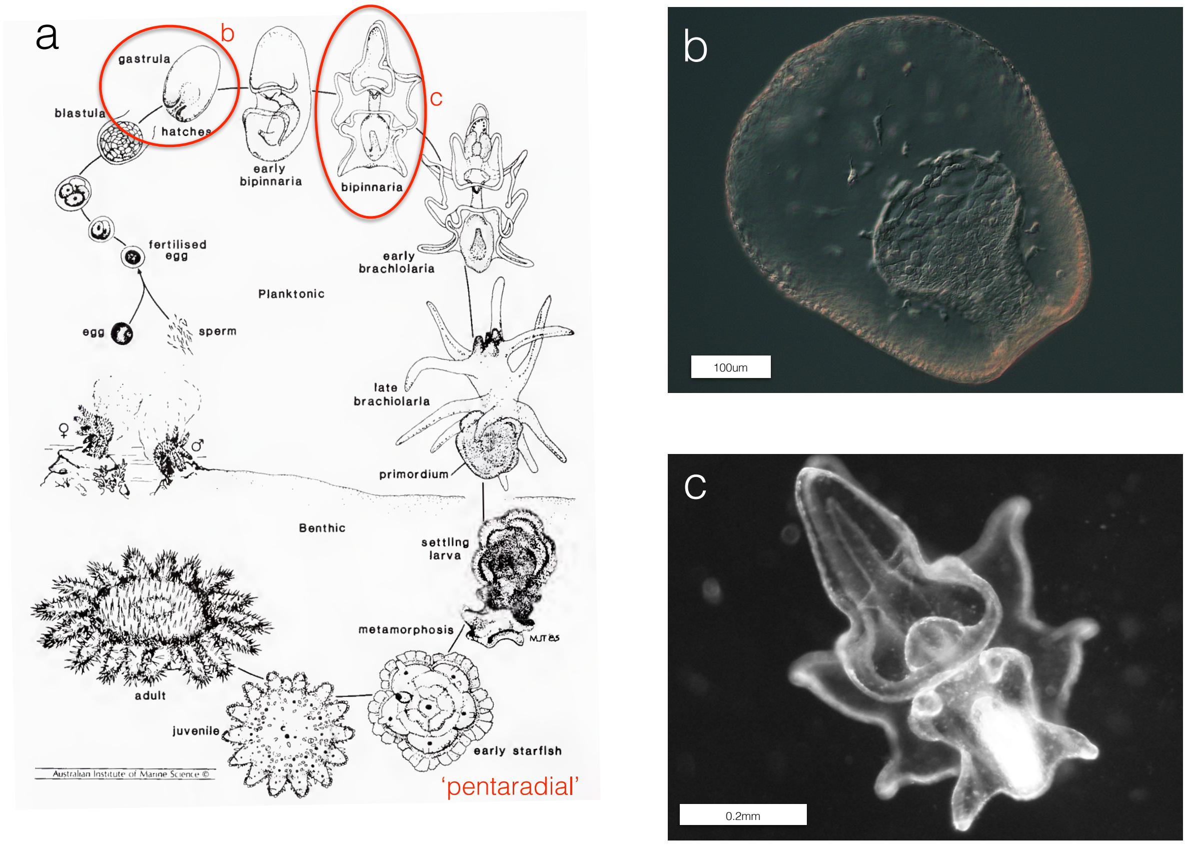


Figure 1.4. The COTS lifecycle.
a, Diagrammatic representation of the life cycle of COTS. Adapted from (Moran 1988). b, Photomicrograph of a live COTS early gastrula (day 2 after fertilization). **c,** Photomicrograph a live COTS bipinnaria (day 7 after fertilization).

1.3 A brief history of starfish and COTS research.

Two major discoveries from two different fields, both made in late 1960s, have greatly influenced the past 50 years of starfish research. The first discovery was made in the field of embryology, and determined that a single hormone triggered oocyte maturation (Ikegami et al. 1967). This discovery led to starfish becoming a model system for embryology, and subsequent discoveries in developmental and cell biology (McClay 2011). The second discovery was made in the field of ecology, and involved the discovery of large aggregations of COTS decimating reefs across the south pacific region(Chesher 1969). This discovery corresponded with the period in which the idea that humans could have lasting impacts on ecological systems first arose, and ushered in an era in which active management and regulation of the environment became commonly accepted and practiced (Sapp 1999).

The discovery that 1-methlyadenine was the hormone responsible for inducing starfish oocytes to prepare for fertilization by resuming meiosis (Kanatani 1964), led to the use of starfish, and in particular the bat star or *Patiria pectinifera (*previously *Asterina pectinifera*), as model system for developmental biology and embryogenesis. Starfish eggs are naturally stored by females at prophase of meiosis. The ectopic application of 1-methlyadenine allows for timed induction of complete meiotic maturation, synchronously en mass. The discovery of methods for inducing meiotic resumption in controlled manner in volumes of oocytes large enough to do biochemistry on, resulted in a number of findings related to the basic cell biology of meiosis (Ikegami et al. 1967; Shirai et al. 1972; Kishimoto & Kanatani 1976), and led to the concept of molecular control of the cell cycle(Draetta et al. 1989).

The first reports of COTS aggregations, or large groups of starfish decimating local reefs (Figure 1.5) were made in the late 1950's, in Okinawa, Japan (Yamaguchi 1986). Reports across the Indo-pacific region from the 1960s to the current day have since established COTS as the most notorious controllable cause of coral reefs devastation (Sapp 1999; Birkeland & Lucas 1990; Moran & De'ath 1992). Research of COTS aggregations can be broken into roughly three phases. The first phase (1960s-1970s) began with the initial discovery of the COTS aggregations, where research largely focused on characterizing the extent of the aggregations. The second phase (1970s-1980s) was initiated by the observation that aggregations had subsided, which opened the discussion as to whether cyclicality in COTS population density may be a natural phenomenon. Finally, the last phase (1990s-current) begins with the recurrence of aggregations in the 1990s and the advent of molecular approaches to population genetics. The key insight of the modern era was the observation that the frequency of observed, cyclical COTS aggregations was much higher than the recovery rate of coral can sustain (De'ath & Moran 1998).

The publication of several reports in high profile science journals on the COTS infestations and biology in the late 1960s and early 1970s (Chesher 1969; Barnes 1970; Brauer et al. 1970; Branham et al. 1971; J. A. Henderson & Lucas 1971; Pearson 1972; Ormond et al. 1973) denote the first phase of COTS research efforts (Sapp 1999). This first phase of research coincided with the general acknowledgement that human impact on local environments could dramatically and irreversibly effect ecology. The degree to which the reef was being impacted by COTS was unknowable at this time, and was subsequently grossly overstated by popular news outlets (Sapp 1999). Moreover, this research occurred as, for the first time ever, SCUBA and snorkeling made reef surveys easily accessible. Though the damage to coral reefs was obvious to even the untrained eye, methods for appropriately measuring and quantifying reef cover simply did not yet exist, and took time to be implemented. The primary scientific discussion was around how to manage the infestations, and what methods could be used to reduce the starfish population size. The cause of the infestations was generally assumed to be the loss of starfish predators (Chesher 1969).

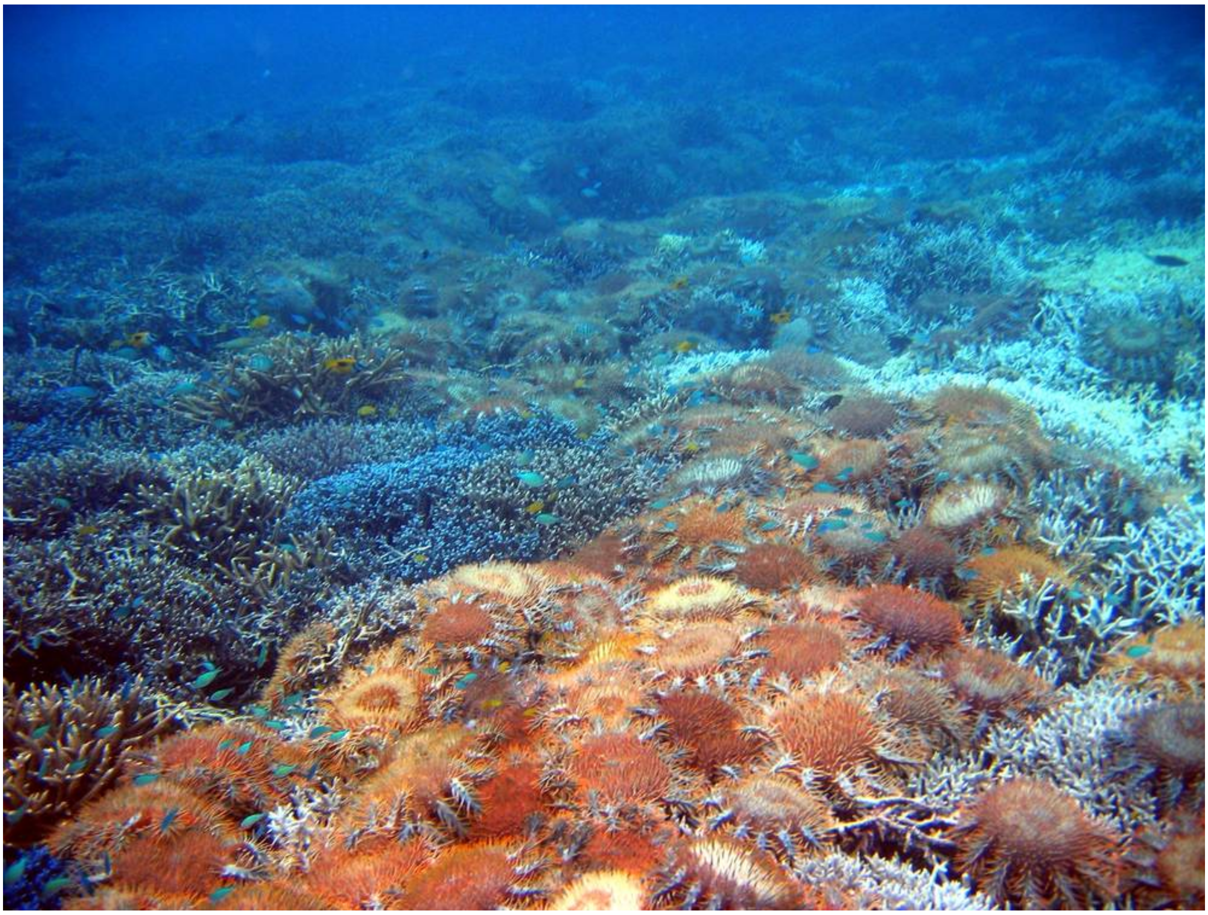


Figure 1.5. A COTS Aggregation.
A high-density COTS aggregation from Miyako Island, Okinawa. Note the dead, white, recently digested coral in the upper right corner. Courtesy of Dr. Kenji Kajiwara.

The second phase of COTS research began in the mid 1970s and continued into the early 1990s. In this period, definitive statements about COTS biology and ecology were made, albeit in a highly-polarized and politicized environment, with tremendous efforts put into determining the role of human activity. The initial wave of alarmist scientific reports had been seized upon by the popular press, leading to sensationalized public statements suggesting, for example, that the loss of the reefs could cause erosion of islands into the sea, which in turn would end human habitation of the South Pacific (Sapp 1999). Thus, scientific push back, as COTS outbreaks abated, was predictable. As more reef observation data were collected, the initial estimates of destruction were downgraded. Table 1.1 summarizes high profile research articles from these first two periods. Two main questions arose (Moran

1988); First, what caused the COTS outbreaks? Second, were the outbreaks an on-going natural cycle, or more specifically, what role did human environmental impact have on starfish populations? In the discussion section (Chapter 4), I will summarize the recent literature (e.g. the third and current phase) and discuss impact that the findings from this genome project have on these two critical questions.

Table 1-1. Summary of high profile COTS publications, 1969-1989.

Article	journal/date	Discipline
Chesher, R. H. **Destruction of Pacific corals by the sea star Acanthaster planci.**	Science 165, 280–283 (1969).	Ecology
Barnes, D. J. Field and Laboratory **Observations of the Crown-of-Thorns Starfish, Acanthaster planci: Locomotory Response of Acanthaster planci to Various Species of Coral.**	Nature 228, 342–344 (1970).	Behavior
Brauer, R. W., Jordan, M. R. & Barnes, D. J. **Triggering of the stomach eversion reflex of Acanthaster planci by coral extracts.**	Nature 228, 344–346 (1970).	Behavior
Branham, J. M., Reed, S. A., Bailey, J. H. & Caperon, J. **Coral-Eating Sea Stars Acanthaster planci in Hawaii.**	Science 172, 1155–1157 (1971).	Ecology
Henderson, J. A. & Lucas, J. S. **Larval development and metamorphosis of Acanthaster planci (Asteroidea).**	Nature 232, 655–657 (1971).	Rearing
Pearson, R. G. **Changes in distribution of Acanthaster planci populations on the Great Barrier Reef.**	Nature 237, 175–176 (1972).	Ecology
ORMOND, R. F. G. et al. **Formation and Breakdown of Aggregations of the Crown-of-Thorns Starfish, Acanthaster planci (L.).**	Nature 246, 167–169 (1973).	Behavior
Beach, D. H., Hanscomb, N. J. & Ormond, R. F. **Spawning pheromone in crown-of-thorns starfish.**	Nature 254, 135–136 (1975).	Behavior
Moore, R. J. & Huxley, C. J. **Aversive behaviour of crown-of-thorns starfish to coral evoked by food-related chemicals.**	Nature 263, 407–409 (1976).	Behavior
Walbran, P. D., Henderson, R. A., Jull, A. J. & Head, M. J. **Evidence from Sediments of Long-Term Acanthaster planci Predation on Corals of the Great Barrier Reef.**	Science 245, 847–850 (1989)	Geology/Ecology

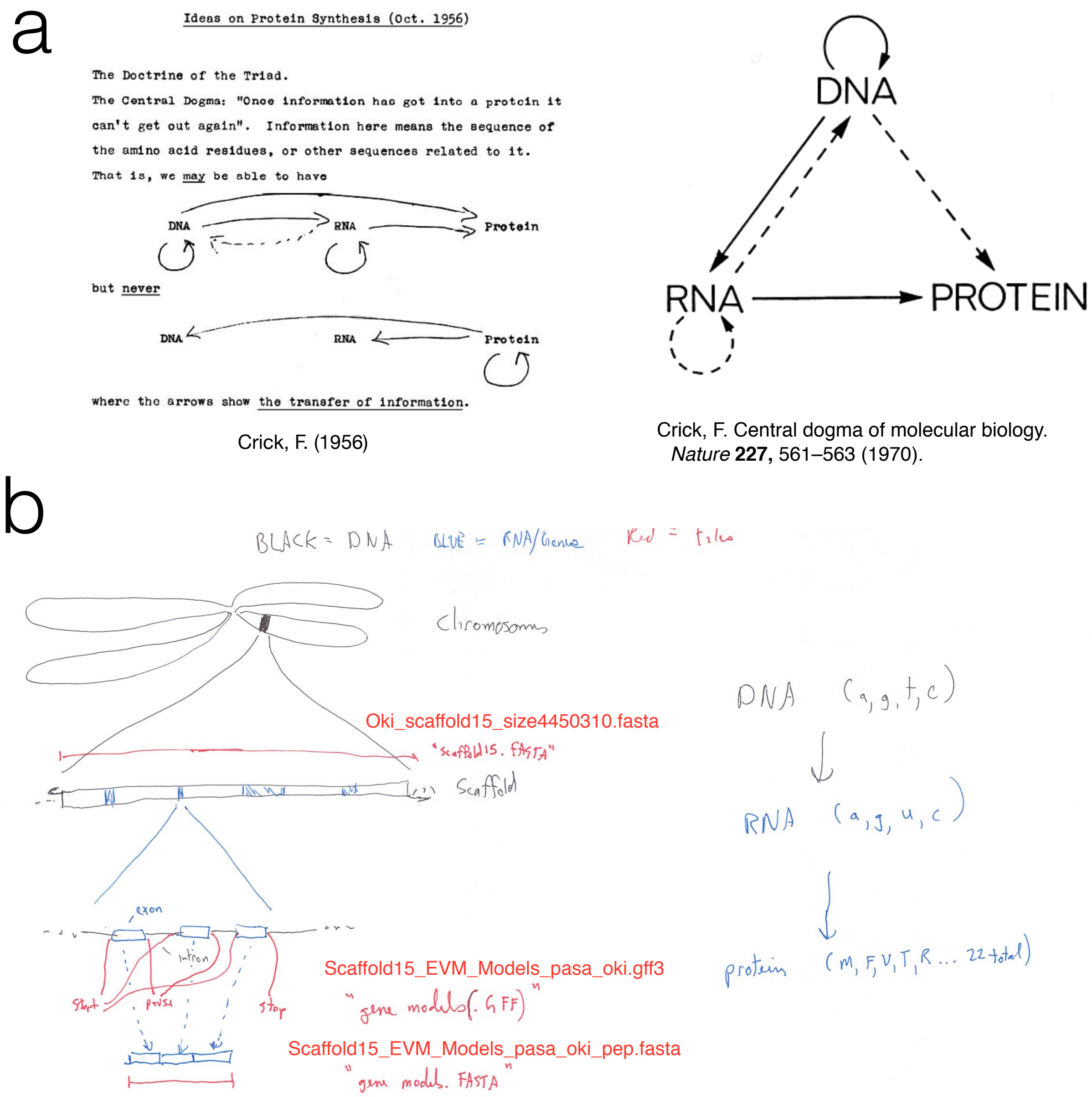
1.5 What is a genome?

What is a genome? A genome can be defined as “all the information required to make a living organism.” More pragmatically, a genome can be defined by the file types that constitute a modern genome-sequencing project. A genome may also be thought of as all the genetic information passed from one generation to the next. Additions, edits, and deletes in a genome occur during this hereditary process, and result in the variation upon which natural selection acts, and new species arise. In this section, I will briefly introduce basic concepts of molecular biology, and discuss how the advent of genomic sequencing has provided new avenues to hypothesize about the origins of species.

Biology is a uniquely challenging scientific discipline simply because the harnessing of nature’s bounty has been the bedrock of civilized society, predating even written language. In other words, Charles Darwin’s conclusions about the origins of species necessarily required integration into millennia of common knowledge about human reproduction, and perhaps more pointedly, the limitations of domesticating flora and fauna. Gregor Mendel’s experiments with peas provided a mechanism for heredity, which was also consistent with observations of any farmer or gardener. Yet, with regard to predictive mechanisms for how sexual selection and breeding result in favorable crops and livestock, our current genomic era remains opaque and fundamentally stuck at an observational perspective; specific, deterministic causal mechanisms are just beginning to be explored, as even the most basic data definitions, control mechanisms, and indeed linguistics are updated, almost annually (Brenner 2010).

1.6 The central dogma and the molecular mechanisms of heredity, by file type

The discovery that deoxyribonucleic acid (DNA) was on the one hand, periodic in structure (Watson & Crick 1953), and on the other, the primary chemical signature associated with heredity(Meselson & Stahl 1958), gave rise to the ‘central dogma of molecular biology’, as proposed by Crick (Figure 1.6a). The central dogma describes information flow within cells, as genomic information is transcribed into messenger RNA, and then translated from messenger RNA into proteins. Accordingly, the three main file types generated by a genomic sequencing project correspond to each of these three chemicals (Figure 1.6b).



Ideas on Protein Synthesis (Oct. 1956)

The Doctrine of the Triad.

The Central Dogma: "Once information has got into a protein it can't get out again". Information here means the sequence of the amino acid residues, or other sequences related to it.

That is, we may be able to have

but never

where the arrows show the transfer of information.

Crick, F. (1956)

Crick, F. Central dogma of molecular biology. *Nature* **227**, 561–563 (1970).

Figure 1.6. The Central Dogma of molecular biology, and genome file types.
a, The left side is a sketch of and early version of the central dogma, from The Francis Crick Papers (https://profiles.nlm.nih.gov/SC/B/B/F/T/). The right side was published more than 10 years later, virtually unchanged, as it remains so, today. **b,** A summary of the file types (in red) generated by a genome sequencing project, in the context of the central dogma (right side). In short, 'scaffolds.fasta' contains the long contiguous stretches of genomic DNA, as assembled following sequencing. The 'genes_models.gff3' file contains addresses from the genomic DNA file that determine which regions of the genomic DNA are expressed as genes. This file is similar to the role that RNA plays, in transcribing genomic information into functional protein. Finally, 'gene_models.fasta' is a file of the protein sequences as delineated by the gff3 file, and corresponds to proteins in the central dogma, which are the form by which the information encoded on the genome is translated into a functional mechanism that impacts cellular physiology.

The first file type of a genome sequencing project is the long genomic scaffolds that are assembled from reads of genomic DNA generated by sequencers; scaffolds.fasta. These scaffolds are bounded by the length of the chromosomes they come from; chromosomal resolution is the gold standard and limit for genomic assembly. Genomic sequences and their corresponding "scaffold" files are both made of DNA, and thus include linear sequences of adenines (A's), guanines (G's), thymines (T's), and cytosines (C's). Generally, diecious (as opposed to hermaphroditic) species that sexually reproduce contain duplicate (2n) copies of the genome in each somatic cell. This heterozygosity is collapsed into a single allele during sequencing, based on whichever allele is more prominent in the sequencing data (e.g. has more reads), and thus, the scaffolds.fasta file contains (1n) sequences of nucleotides.

The discovery that a degenerate triplicate code directly connects the hereditary information of a genome to cell physiology via proteins, and the decoding of this process in 1961 (CRICK et al. 1961), can be considered to be the single most influential consequence of understanding the molecular structure of genes, as this discovery launched our current era of molecular biology. This single discovery not only proved that biology works on a discrete system of information transfers, decoding the triplicate code also provided the tool kit and linguistic rules for manipulating those processes. For the information contained within a genome to affect the physical world, the nucleotides must be transcribed from mRNA into functional proteins. Proteins constitute the physical structure and chemistry of cells, which in turn controls the behavior of cells, and groups of cells.

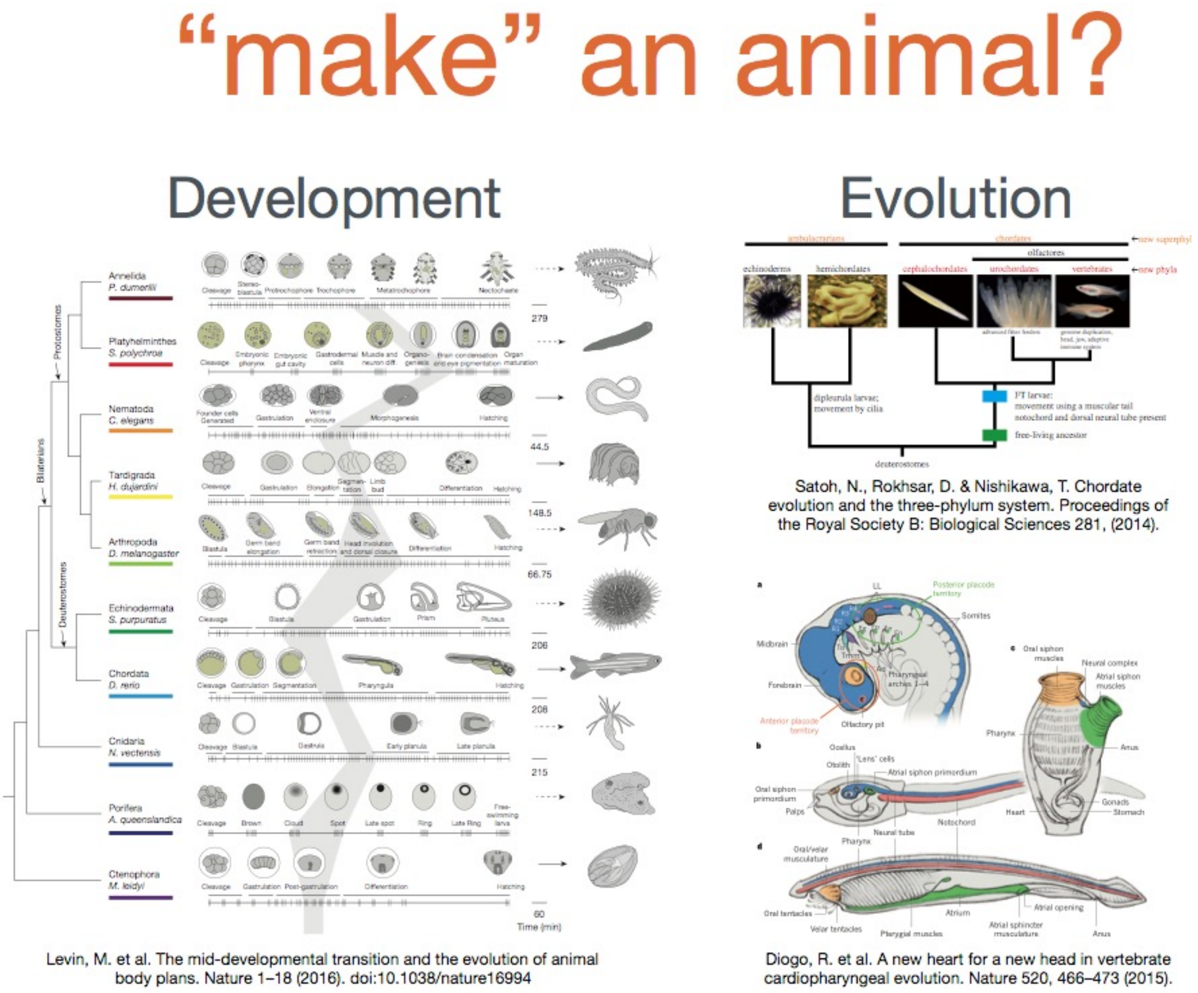
The second file type highlighted in Figure 1.6b corresponds to the regions of scaffolds that are transcribed into messenger ribonucleic acid (mRNA). General feature format or 'GFF' files assign regions of the genome scaffolds to either expressed (e.g. exon) or regulatory domains. GFF files most closely align with central dogmas' mRNA, as they record the genomic addresses for functional regions of a genome. Each entry in the GFF file describes the genomic address for a given mRNA transcript in a table outlining the order and genomic location of each respective component. GFF files themselves do not contain any sequence information.

The final file type in figure 1.6b is the gene_models.fasta, which contains protein sequences that exactly correspond to the addresses given by the GFF file, and the scaffold nucleotides assigned each of these gene models. Whereas scaffold and GFF files are based on nucleotides, gene_models are lists of amino acids; due to the degeneracy of the triplicate code, different nucleotide sequences can sometimes result in identical gene_models. Due to the informational flow described in the central dogma, gene_models cannot directly impact the organization or genomic nucleotide sequence. Recently it has been quite popular to highlight violations of the general informational flow of the central dogma, the vast majority of these mechanisms deal with control or feedback of gene expression.

Thus, the fundamental question driving this thesis is: how can the information encoded in a genome be used to make an organism? More specifically, what controls when and where proteins are expressed, which in turn result in defined cell types and their corresponding organs and tissues? Does the genome contain physical maps? Does the genome contain assembly instructions? The central dogma is a framework for the flow of information extraction from the genome. Clearly, the notion of a 'regulatory layer' within a given genome is one simplistic answer, but note that a standardized file type for encoding a regulator layer is not included here, nor has a rigorous, discrete linguistics for how this control layer might be translated been proposed. Surely, the GFF file and the gene_model file provide information that hints at the heuristics of such a regulatory layer, but to date, the data structure for the biological mechanisms constituting such a dataset, described as a genome project file type, simply do not exist.

1.7 What is EvoDevo?

If genomes include “all the information required to make an animal,” it is possible to consider the time scale over which the ‘making’ is happening (Figure 1.7). At the developmental timescale, a single life cycle requires that the information contained in a given genome be identical enough to that of another given genome from the same species, that the resulting sexually mature mates show up at the same time and place such that their gametes can interact, and more specifically, these two genomes be identical enough to hybridize with each other. Conversely, on the evolutionary time scale (e.g. long enough to allow for speciation), genomes must diverge enough to result in animals that are morphologically and behaviorally different enough to be considered separate species. This genome-oriented perspective on Evolutionary Developmental Biology, or EvoDevo highlights why toolkit genes, transcription factors, and systematic control of protein expression have been the primary dataset for comparing “all the information required to make an animal.” In other words, if EvoDevo is the study of how conserved developmental patterns can recapitulate ancient common ancestors, then it follows that the genetic control mechanisms driving those patterns may also be conserved, to some degree.



“make” an animal?

Development

Evolution

Satoh, N., Rokhsar, D. & Nishikawa, T. Chordate evolution and the three-phylum system. Proceedings of the Royal Society B: Biological Sciences 281, (2014).

Levin, M. et al. The mid-developmental transition and the evolution of animal body plans. Nature 1–18 (2016). doi:10.1038/nature16994

Diogo, R. et al. A new heart for a new head in vertebrate cardiopharyngeal evolution. Nature 520, 466–473 (2015).

Figure 1.7. Comparison between the developmental and evolutionary time scale.

Chapter 2 : Methods

There are seven steps to sequencing and annotating a genome: 1) Biological materials Sampling and Collection. 2) Isolation of high quality genomic DNA and RNA. 3) Preparation of DNA libraries for next generation sequencing. 4) Next-Generation Sequencing. 5) Assembly of genomic sequence data. 6) Annotation of genomic sequences: Gene modelling and annotation. 7) Genome Analysis. To minimize allelic variation and allow for comparative analysis of COTS specimens, one from the Great Barrier Reef ('GBR') and the other from Okinawa ('OKI'), the respective genomes were sequenced, assembled and annotated separately. The COTS genome pipeline is summarized in Figure 2.1, and related experiments are summarized in Table 2.1.

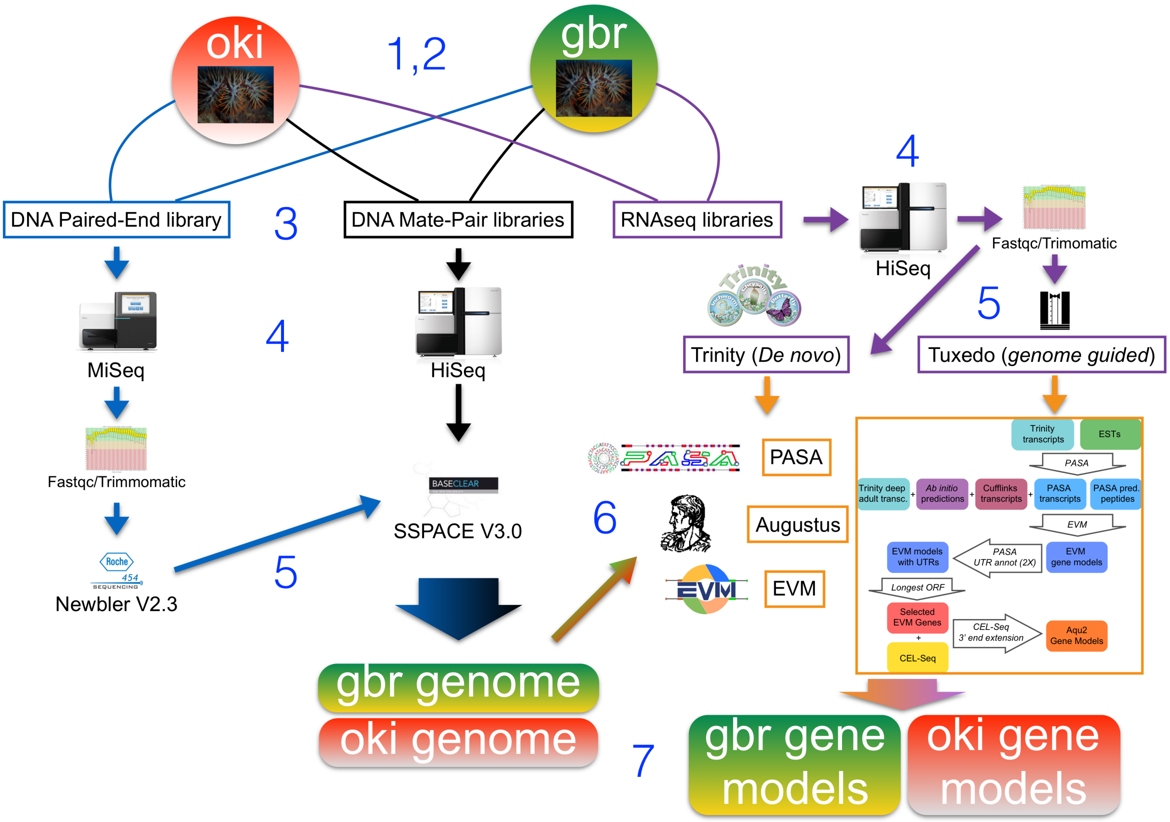


Figure 2.1. COTS Genome Assembly and Annotation Pipeline.
This figure summarizes the methods and pipeline used to sequence (in blue), assemble (in black), and annotate (purple and orange) two separate COTS genomes, "OKI (red/white)" and "GBR (green/yellow)", in parallel. The main steps in the pipeline: 1) Sample collection, 2) DNA and RNA extraction, 3) Library construction, 4) Sequencing, 5) Assembly, 6) Gene Model prediction, and 7) Genome analysis.

2.1 Biological materials: Sampling and collection

(a) Adult materials.

With the assistance of OIST technical diver Koichi Toda and the OIST Marine Resources section, mature male *Acanthaster planci* specimens were collected from reefs near Motobu, Okinawa (Figure 2.2; 26°40'46.1"N 127°52'46.1"E) on May 28th, 2013. Site#1 refers to the initial coral reef location where COTS were observed in their natural environment. Several starfish were collected at this site, and dissected to determine sex and confirm fecund gonads. During breeding season, in the days leading up to broadcast spawning which generally occurs around the full moon, both male and female COTS gonads become engorged, occupying 10-25% of the body cavity. Male gonads are yellowish in color, while female gonads have whitish egg sack structures, which are visible by eye. Site#2 was the location from which 2 male COTS (IDs #877 and #890) were collected for sequencing.

Individual COTS were detected by specific discoloration patterns found on top of corals (Figure 2.3). White, 'bleached' regions of coral result from COTS feeding on the coral. Generally individual COTS can be found within several centimeters of these bleached regions, often directly underneath the effected coral head. The generally accepted population density to denote a COTS outbreak is >15 COTS per hectare (Moran & De'ath 1992). Both regions highlighted in Figure 2.2, as well as reef location pictured in Figure 2.3 exceed this population density, though not large aggregations of COTS were observed.

Male gonad tissue from #877 and #890 were isolated, on site. In short, the dorsal skin was peeled back using forceps and dissection scissors. After dissection to remove pyloric cecum, male gonad samples were removed and transferred to 50 ml falcon tubes on dry ice for transport to the lab, where samples were snap frozen in liquid nitrogen and then stored at -80°C. Approximately 50 mls of male gonad were collected from each male. For RNAseq analysis and gene modeling, four tissue samples (testis, podia, spine, and mouth/stomach) were collected from each male. COTS sample#877 was chosen for genome sequencing.

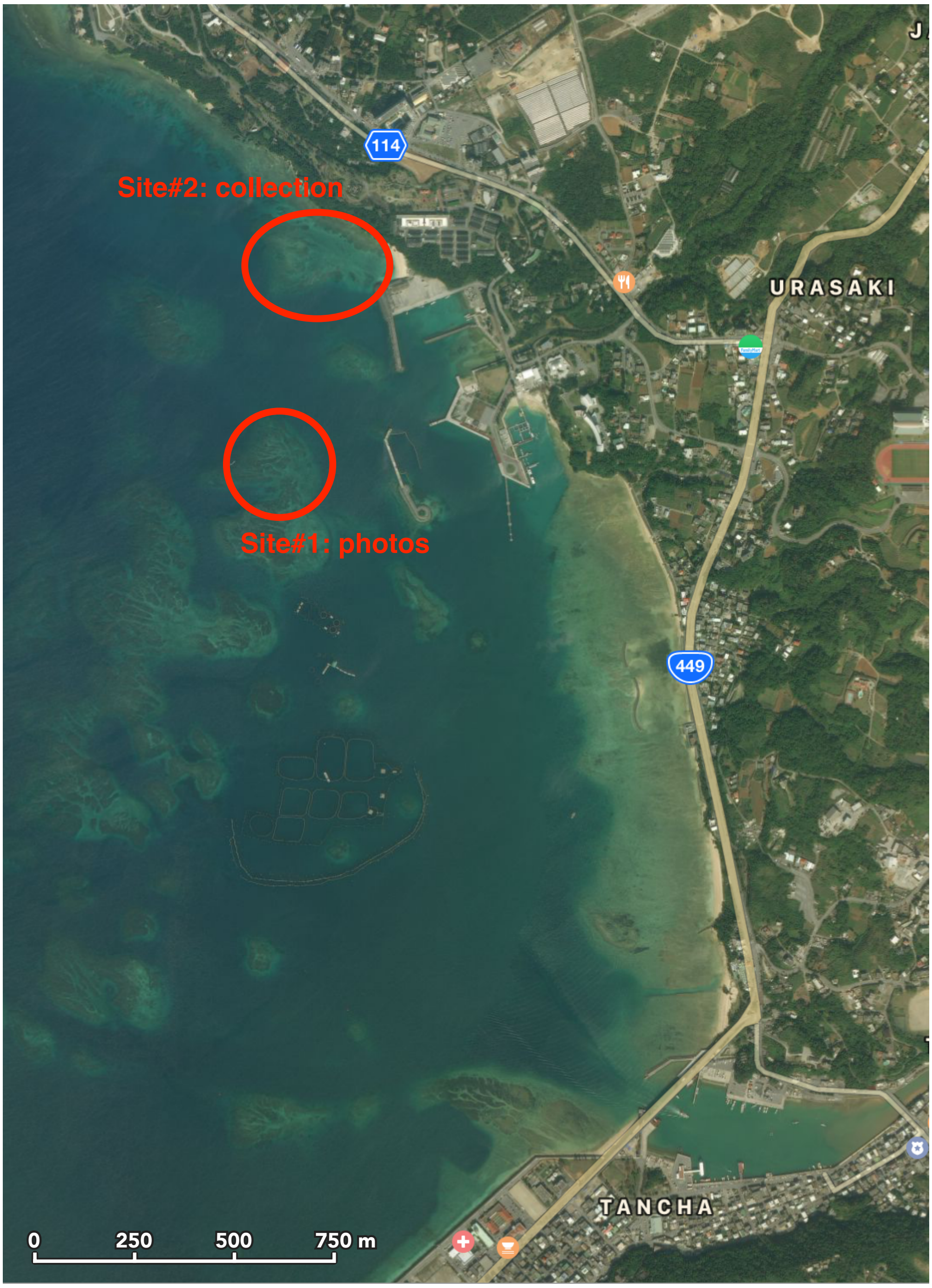


Figure 2.2. Map of 'OKI' sample collection site
Near Motobu, Okinawa, Japan. The inset in Figure 2.3 was taken at Site#1. The COTS that was sequenced was collected from site#2.

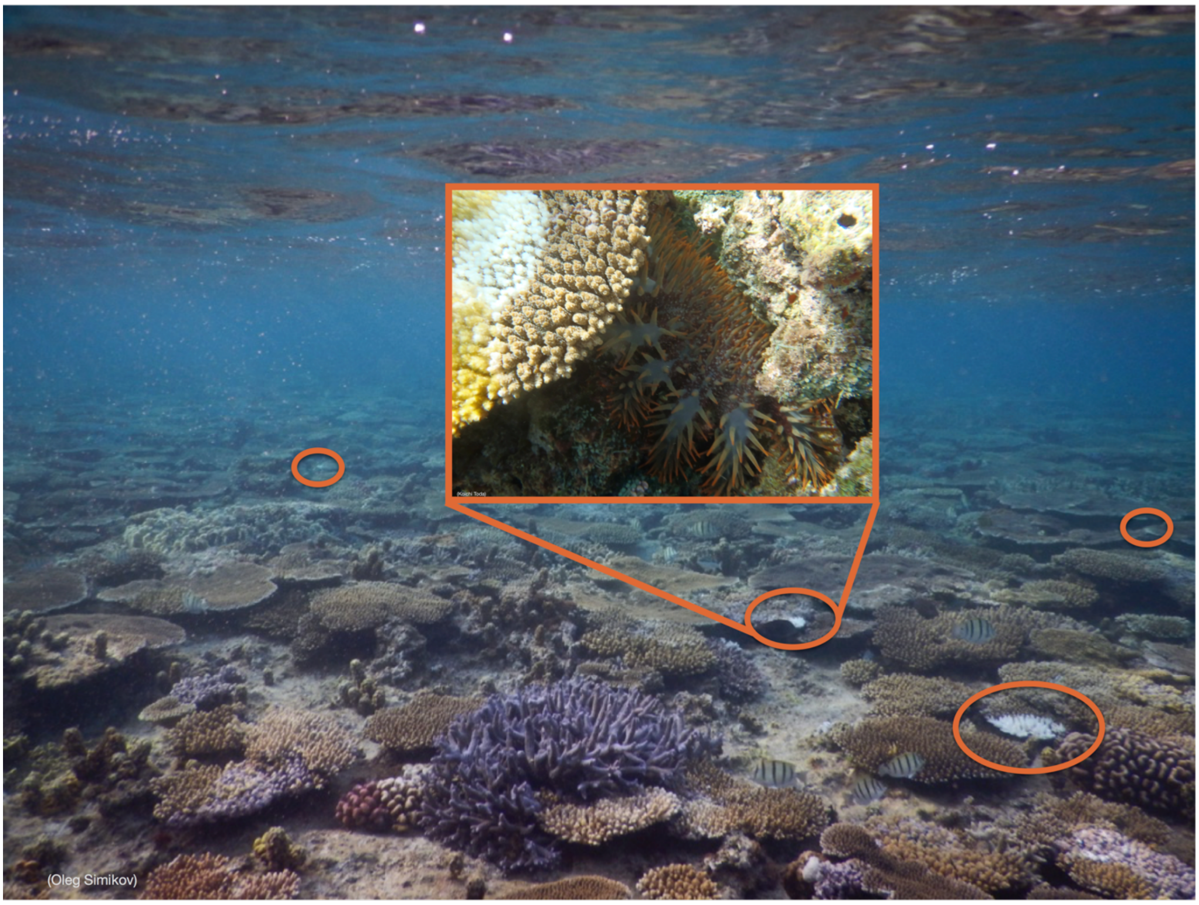


Figure 2.3. COTS feeding underneath *Acropora digitifera* coral heads.
Taken at 2-3 meters depth, Manzamo, Okinawa. Circles = ~20cm. The inset shows a representative COTS hiding under a coral head. Note, in the inset white (digested), brown (live coral), and yellow (algal transition) regions of the coral head are visible. (Photo by Oleg Simikov, inset by Koichi Toda)

An additional COTS specimen was collected from Rudder Reef on the northern Great Barrier Reef, Australia (16°11'46.4"S 145°41'48.7"E) on February 4th, 2013 by collaborators from the Australian Institute of Marine Science (AIMS), Cape Ferguson, Townsville, Queensland 4810, Australia. The male gonad sample was prepared using standard procedures and shipped to Okinawa for genomic library preparation and sequencing. Five additional tissue samples (testis, podia, spine, stomach, and body-wall) for GBR RNA were collected for RNAseq and gene modeling. An additional male COTS was collected on June 30th, 2016 and fresh sperm samples were provided to Dr. Ryo Koyanagi of the OIST DNA Sequencing Section (SQC) for BioNano and Dovetail genome 'polishing' methods.

For OKI RNA from COTS embryos, additional tissue samples were prepared from females and embryos generated from COTS collected by the Onna Fisheries Collective (Nakamura et al. 2014), during the summer of 2013. In collaboration with Dr. Eiichi Shoguchi, samples of nerve tissue, stomach tissue, and oocytes were collected, isolated in RNA-easy, and stored at -80°C.

(b) Larval Materials

Fecund, spawning adult male and female COTS were collected from the reefs in front of Onna village during the months of June, July, August, and September, by the local fisherman's association, by snorkeling. Although these animals are generally destroyed, several adult COTS were rehydrated and their gonads harvested for embryological studies. When the ambient water temperature exceeds 28°C, generally in cycle with the full moon, COTS become fecund. COTS exceeding 20 centimeters in diameter generated high quality eggs or sperm, and generally, the larger the starfish, the higher the number and quality of the gonads. In a petri dish, 2 µM 1-methyl-adenine (1-MA) was added to dissected egg cases, in order to trigger mitotic resumption in eggs at a specific time point (Figure 2.4). Natural spawning was observed on several occasions; spawning was induced several times by temperature change when adult COTS were shifted from 28°C natural sea water to 18°C tank water, by electric shock when a heating element fell into the tank, and during the sample collection when COTS spawning occurred in the field while starfish were being transferred by mesh net bag. In all cases of natural spawning, samples of fertilized embryos were reared alongside timed, 1-MA induced fertilizations. In addition to oocytes, two developmental time points, early gastrula (EG) and mid gastrula (MG) were collected and prepared for RNA sequencing.

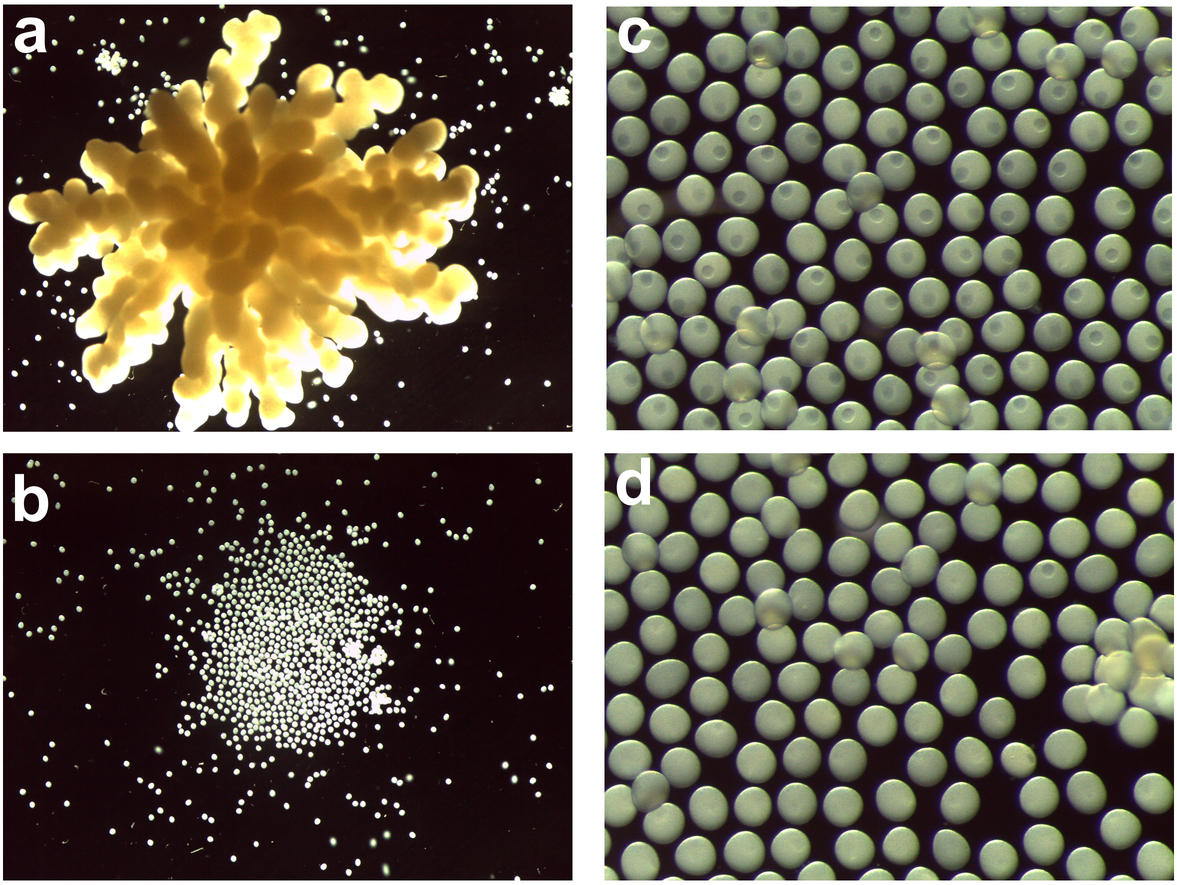
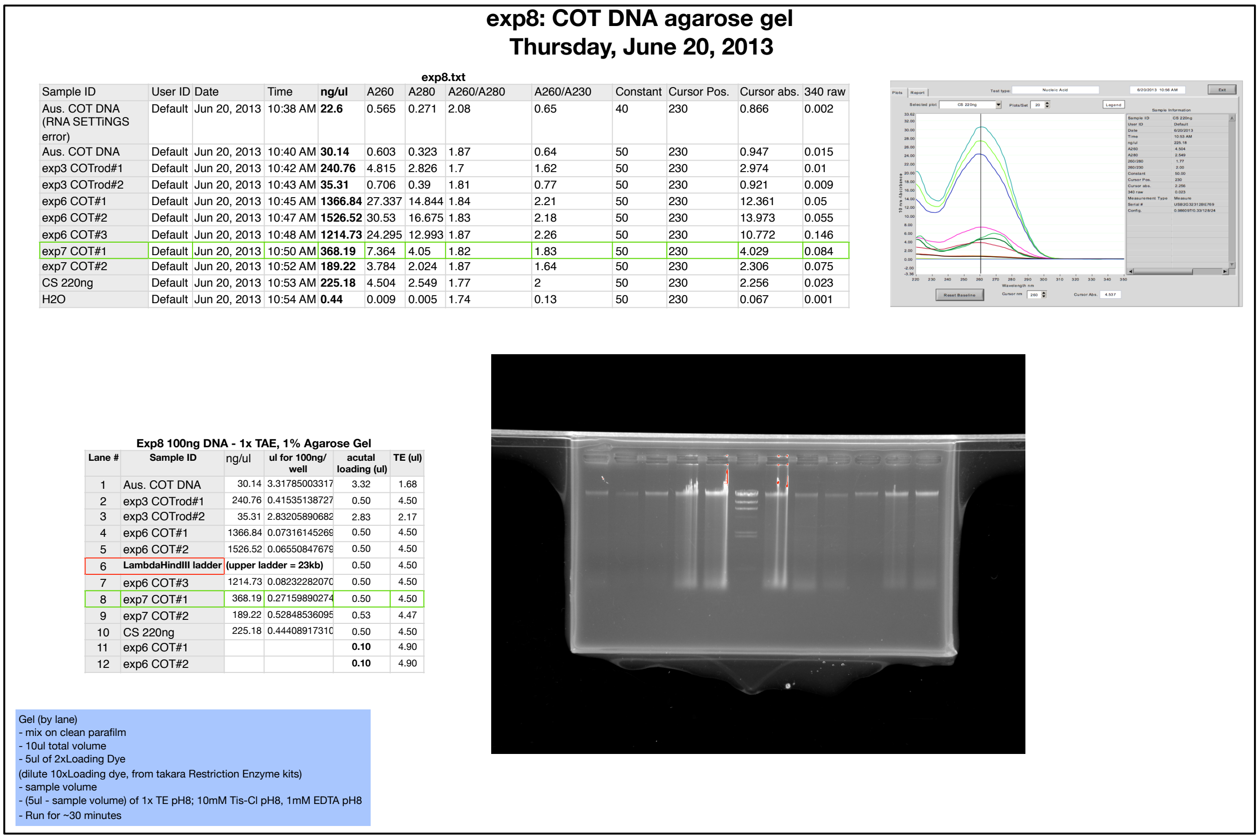


Figure 2.4. COTS eggs treated with 1-methyl-adenine (1-MA).
a, An untreated, dissected COTS egg case. **b,** After addition of 1-MA, eggs are ejected from the dissected egg case. **c,** Untreated eggs with visible germinal vesicles. **d,** After 15 minutes of 1uM 1-MA treatment, germinal vesicles have broken down (GVBD), and the eggs are ready for timed fertilization.

2.2 Isolation of high quality genomic DNA and RNA

Isolation of genomic DNA has traditionally been done from sperm samples, as sperm tissue is enriched in DNA, with lower levels of protein, lipid, or carbohydrate contaminants, as compared with other tissue sources. Although several kits or commercial products exist for isolating genomic DNA from tissue samples, we found that the classical ‘Phenol and Proteinase K” method (Green & Sambrook 2012)produced the highest quality gDNA. In short, frozen tissue samples are pulverized in liquid nitrogen by mortar and pestle. Cells are lysed in a Tris-EDTA RNAase A buffer for 30 minutes, then treated overnight with proteinase K. DNA is then extracted by phenol-chloroform, precipitated and washed in Ethanol, and re-suspended in TE buffer. The gDNA was then assayed by nanodrop and agarose gel electrophoresis (Figure 2.5). The primary difference between gDNA preparations was the thickness or ‘snottiness’ of the precipitated gDNA. The highest quality preparations resulted in large, gelatinous strands of visible DNA. RNA samples were transferred directly to RNA-easy solution and stored at -80°C until library preparation.



exp8: COT DNA agarose gel
Thursday, June 20, 2013

exp8.txt

Sample ID	User ID	Date	Time	ng/ul	A260	A280	A260/A280	A260/A230	Constant	Cursor Pos.	Cursor abs.	340 raw
Aus. COT DNA (RNA SETTiNGS error)	Default	Jun 20, 2013	10:38 AM	**22.6**	0.565	0.271	2.08	0.65	40	230	0.866	0.002
Aus. COT DNA	Default	Jun 20, 2013	10:40 AM	**30.14**	0.603	0.323	1.87	0.64	50	230	0.947	0.015
exp3 COTrod#1	Default	Jun 20, 2013	10:42 AM	**240.76**	4.815	2.826	1.7	1.62	50	230	2.974	0.01
exp3 COTrod#2	Default	Jun 20, 2013	10:43 AM	**35.31**	0.706	0.39	1.81	0.77	50	230	0.921	0.009
exp6 COT#1	Default	Jun 20, 2013	10:45 AM	**1366.84**	27.337	14.844	1.84	2.21	50	230	12.361	0.05
exp6 COT#2	Default	Jun 20, 2013	10:47 AM	**1526.52**	30.53	16.675	1.83	2.18	50	230	13.973	0.055
exp6 COT#3	Default	Jun 20, 2013	10:48 AM	**1214.73**	24.295	12.993	1.87	2.26	50	230	10.772	0.146
exp7 COT#1	Default	Jun 20, 2013	10:50 AM	**368.19**	7.364	4.05	1.82	1.83	50	230	4.029	0.084
exp7 COT#2	Default	Jun 20, 2013	10:52 AM	**189.22**	3.784	2.024	1.87	1.64	50	230	2.306	0.075
CS 220ng	Default	Jun 20, 2013	10:53 AM	**225.18**	4.504	2.549	1.77	2	50	230	2.256	0.023
H2O	Default	Jun 20, 2013	10:54 AM	**0.44**	0.009	0.005	1.74	0.13	50	230	0.067	0.001

Exp8 100ng DNA - 1x TAE, 1% Agarose Gel

Lane #	Sample ID	ng/ul	ul for 100ng/ well	acutal loading (ul)	TE (ul)
1	Aus. COT DNA	30.14	3.31785003317	3.32	1.68
2	exp3 COTrod#1	240.76	0.41535138727	0.50	4.50
3	exp3 COTrod#2	35.31	2.83205890682	2.83	2.17
4	exp6 COT#1	1366.84	0.07316145269	0.50	4.50
5	exp6 COT#2	1526.52	0.06550847679	0.50	4.50
6	**LambdaHindIII ladder**	**(upper ladder = 23kb)**		0.50	4.50
7	exp6 COT#3	1214.73	0.08232282070	0.50	4.50
8	exp7 COT#1	368.19	0.27159890274	0.50	4.50
9	exp7 COT#2	189.22	0.52848536095	0.53	4.47
10	CS 220ng	225.18	0.44408917310	0.50	4.50
11	exp6 COT#1			**0.10**	4.90
12	exp6 COT#2			**0.10**	4.90

Gel (by lane)
- mix on clean parafilm
- 10ul total volume
- 5ul of 2xLoading Dye
(dilute 10xLoading dye, from takara Restriction Enzyme kits)
- sample volume
- (5ul - sample volume) of 1x TE pH8; 10mM Tis-Cl pH8, 1mM EDTA pH8
- Run for ~30 minutes

Figure 2.5.Analysis of COTS genomic DNA.
COTS gDNA concentration by nanodrop (top) and Agarose gel electrophoresis. “exp7COT#1” was the 3rd DNA preparation from COTS#877, and was used for sequencing.

2.3 Preparation of DNA libraries for next generation sequencing

Libraries for sequencing were made by the OIST DNA Sequencing Section (SQC), with the help of member of the Marine Genomics Unit. Genomic DNA from each starfish (‘GBR’ and ‘OKI’) was used to make paired-end, mate-pair, and RNAseq libraries.

(a) Paired-end libararies (MiSeq)

Paired-end libraries for OKI and GBR were prepared in collaboration with Kanako Hisata, using standard methods supplied by Illumina (Baughman et al. 2014; Shoguchi et al. 2013). Paired-end libraries are made from fragmented genomic DNA. The genomic distance between the ends of each paired-end read should not exceed the length of that read. These reads are used to construct contiguous sequences (or contigs') that provide the bulk of the nucleotide specific information.

(b) Mate-Pair Libraries (HiSeq)

For the OKI genome, four mate-pair libraries with differing insert size were prepared in the OIST SQC by Drs. Miyuki Kanda and Manabu Fujie. For the GBR genome, three mate-pair libraries with differing insert size were generated and sequenced by Macrogen, inc. Mate-pair libraries are made by again fragmenting gDNA according to the Illumina protocol, but fragments of specific insert sizes are selected. The goal is to sequence two reads separated by an insert of a specific size. In the case of OKI COTS, four mate-pair libraries were made with insert size targets of 1.5-4, 4-6, 6-8 and 8-12 kb, while three mate-pair libraries with insert size targets of 3, 8, and 12 kb were made for GBR. These mate-pair reads are then used to align the contigs into ‘scaffolds’ of the final assembly.

(b) RNAseq libraries (HiSeq)

15 different RNA samples were collected from both the genome-sequenced COTS (e.g. OKI and GBR), as well as from several additional individuals. 15 RNAseq libraries were constructed by Saori Araki. RNA transcripts are used to determine tissue-specific gene expression patterns, and as a primary dataset for gene model prediction.

2.4 Next-generation sequencing

(a) MiSeq sequencing

In collaboration with Kanako Hisata, paired-end libraries of 40x coverage for GBR (3 sequencing runs) and 46x coverage for OKI (4 runs) were sequenced on an Illumina MiSeq sequencer. This generated 250-base overlapping reads. An ~800 bp paired-end library for GBR was sequenced in three MiSeq runs, and two paired-end libraries of ~600 bp and ~1000 bp for OKI were each sequenced in two MiSeq runs. This sequencing was done in the Marine Genomics Unit.

(b) HiSeq sequencing

For the OKI genome, 4 mate-pair libraries with target insert sizes of 1.5-4, 4-6, 6-8 and 8-12 kb were sequenced on Illumina HiSeq sequencer at 152x coverage in 2 lanes, by the OIST SQC. For the GBR genome, 3 mate-pair libraries with average insert sizes of 3, 8, and 12 kb were sequenced by Macrogen, Inc. resulting in 139x coverage.

Notably, the paired-end reads were sequenced by MiSeq, which resulted in lower coverage (~45x versus 150x), but longer reads (~250 bp versus 50 bp), as compared to the HiSeq. Conversely, high coverage sequencing of mate-pairs was done on the HiSeq. Notably, this methodology for genome assembly using only Illumina-based short read technology has been reported elsewhere, but the COTS assembly was an order of magnitude better in quality and length (Cameron et al. 2015).

(b) HiSeq RNAseq sequencing

The 15 RNA samples that were collected from both the sequenced individuals (e.g. OKI and GBR), as well as from several additional individuals for developmental stages were sequenced by HiSeq in the OIST SQC.

2.5 Assembly of genomic sequence data

Genome assembly involves filtering the raw reads generated by sequencers, assembling the filtered paired-end reads with software to build contigs, and scaffolding the contigs with mate-pair reads to generate the final assembly.

(a) Assembly of paired-end reads into contigs: Newbler

Paired-end raw reads were collected on the MiSeq sequencer and assembled into contigs. In collaboration with Kanako Hisata, raw paired-end read data was first trimmed and aligned the using Trimomatic (Bolger et al. 2014), and then filtered by read quality using FastQC (Gordon & Hannon 2010). The filtered, aligned reads were then assembled using GS De Novo Assembler version 2.3 (Newbler, Roche). The Newbler assembly software uses ‘overlap’ assembly algorithms, in which contigs are extended using paired-end reads that map to each growing contig(Nagarajan & Pop 2013).

I also attempted to assemble the paired-end reads with k-mer based methods using the velvet assembler (Hall et al. 2017; Zerbino & Birney 2008), in which reads are first designated to batches of k-mers, before these groupings are extended into contigs, based on read support. Initially, using only the paired-end reads, the velvet assembler was unable to improve upon the Newbler assembly. Therefore, published assemblies (V0.5, V1) were based on the Newbler contig data provided by Kanako Hisata.

(b) Scaffolding of contigs with mate-pair reads (SSPASE).

For all versions of genome assemblies, mate-pair sequencing data were used to scaffold the newbler contigs into the genomic scaffolds, using SSPACE (Boetzer et al. 2011). The raw reads were processed by two different methods: the first approach involved trimming mate-pair with PrinSeq (Schmieder & Edwards 2011), and then selecting high quality score mate-pairs with Trimmomatic (Bolger et al. 2014). The second approach involved removing low quality score reads with fastq_quality_trimmer, and then pairing and filtering reads by quality score with cmpfastq. The second pipeline resulted in more, higher quality reads, so this method was used to process the raw mate-pair read data. The initial attempt at scaffolding was done using the processed mate-pair reads by SSPASE_basic 2.0(Boetzer et al. 2011), which resulted in two assemblies (oki_V0.5 and gbr_V0.5) that were published in the Hox report(Baughman et al. 2014). The final, published scaffolding was done using the processed mate-pair reads by SSPASE3.0 (Boetzer et al. 2011), which resulted in two assemblies (oki_V1, gbr_V1) that were available on OIST Marine Genomics website (http://marinegenomics.oist.jp/cots/viewer/info?project_id=46). The main difference between SSPACE_basic2.0 and SSPASE3.0 is the software used to map the mate-pair read data to the scaffolds; SSPACE_basic2.0 uses bowtie2, while SSPASE3.0 uses BWA.

The gapfilling techniques described below (e.g. replacing N's in the scaffolded assemblies) are often employed during scaffolding, but were omitted in this study as they either reduced the quality of the scaffolds, or did not add significant improvement to the overall assembly. First mate-pair reads can be used to extend scaffold length during scaffolding, but we found that using the "EXTEND" option during both versions of SSPACE actually reduced the quality of the final assembly. Without a better understanding of how SSPACE functions, it is unclear why this happened. Second, gap closing software is often used with mate-pair data, to replace the N's, or 'unknown bases' in contigs. I tried both gapfiller (Boetzer & Pirovano 2012) and gapcloser (R. Luo et al. 2012) software packages. Gapfiller took much longer to run, but only replaced approximately 10% of the N's in both genomes. Gapcloser ran much more quickly, and replaced around 50% of the N's. Given that only 2.6% of both OKI and GBR genomes were unknown (N's), we concluded that neither approach improved the overall quality and thus did not use either method for final assemblies.

To confirm insert size and quality of mate-pair reads, mate-pair reads were mapped back to final scaffolds with BWA (Li & Durbin 2010), and analyzed with QualiMap (Garcia-Alcalde et al. 2012). Histograms for insert size distribution for each library (Figure 2.6) confirm that mate-pair reads used for scaffolding fall within predicted insert size, as targeted during library preparation.

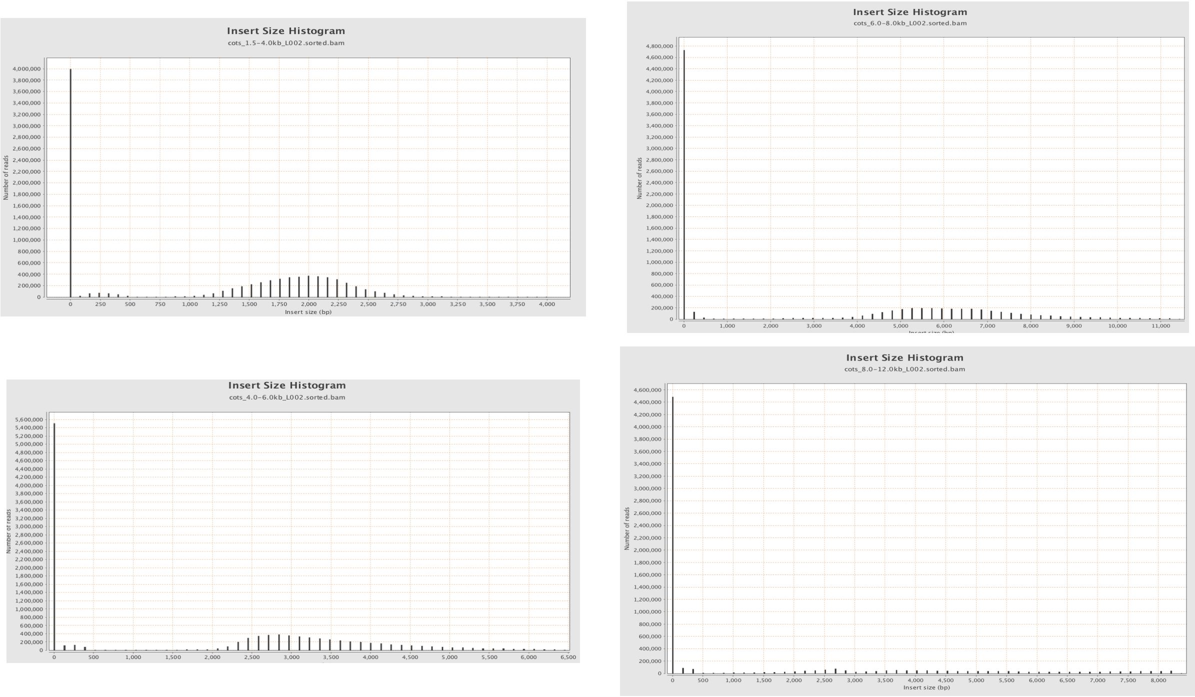


Figure 2.6. Insert size distribution for OKI mate-pair reads.
QualiMap histograms for insert size distribution for each OKI mate-pair library. The paired, processed OKI mate-pair reads were mapped back to the assembled OKI genome. The x-axis is the size of insert length between pairs. As these distributions match the length targeted during library construction for each of the 4 libraries, these results confirm the high quality of library construction and read processing.

(c) Genome polishing by Dovetail and BioNano

Recently, a k-mer-based assembly using the original paired-end, mate-pair, and a novel ‘chicago library’ reads has been developed by Dovetail, Inc. (https://dovetailgenomics.com). The method attempts to further increase scaffold length of genome assemblies, by synthesizing extremely long mate-pair (30,000+ kbs). Likewise, a restriction-digest and labeling approach created by BioNano, Inc. (http://bionanogenomics.com) increases scaffold length, and can provide insight into genome structure. Both methods require the OKI V1.0 COTS scaffolds as input.

Both methods also require additional, fresh genomic material. The genomic sample prepared from COTS#877 was determined to have a fragment size that was too small. Thus, 'high molecular weight' genomic DNA was prepared from fresh COTS male gonad tissue by the OIST SQC by an agar-based *in situ* digestion method (Zhang et al. 2012). This high molecular weight gDNA was used successfully in both DoveTail and BioNano protocols. The final Dovetail scaffolding was done in house by Dovetail, and represents one of their best improvements, to date. The final BioNano scaffolding was done by the OIST SQC.

(d) RNAseq transcriptome assembly.

RNAseq reads were assembled into transcriptomes by two different methods; the first did not use genome scaffolds (e.g. de novo assembly), while the second used the genome scaffolds as reference. Raw RNAseq reads from the 15 different tissues were filtered using Trimomatic (Bolger et al. 2014) and fastQC (Gordon & Hannon 2010). RNAseq transcriptomes were then assembled *de novo* (e.g. without using genome scaffolds as reference) using Trinity (version r20131110)(Haas et al. 2013). RNAseq reads were also concatenated into 'all oki', 'all gbr', and 'all COTS' reads, leading to 18 total RNAseq Trinity assemblies. Genome-guided Trinity assemblies were also generated, but not analyzed. The Tuxedo pipeline (Trapnell et al. 2012) was used to generate a separate set of 15 RNA transcriptomes, based on scaffolds from the respective GBR or OKI genome.

2.6 Annotation of genomic sequences: Gene modelling and annotation.

Once scaffolding is complete, gene models are predicted and annotated. In short, gene models are generated by aligning assembled RNAseq transcripts to the genome, and filtering those results to align to known constraints within the gene structure.

(a) *Ab initio* Gene model prediction for Hox and Parahox cluster analysis.

Initially, I generated COTS gene models for only V0.5 GBR scaffolds containing Hox or Parahox gene clusters. The initial clusters were identified by aligning scaffolds to the Hox and Parahox genes from sea urchin, starfish, hemichordate, and Drosophila, as well as the homeodomain, with TBLASTN (Baughman et al. 2014). The Hox and Parahox containing scaffolds from OKI and GBR were identified (4 total: grbV0.5#27, grbV0.5#59, okiV0.5#15, and okiV0.5#470). The Hox and Parahox-containing GBR v0.5 scaffolds (grbV0.5#27, grbV0.5#59) were submitted to FGENESH for *ab initio* gene prediction, using *S. purpuratus* as reference (Solovyev et al., 2006).

(b) Preliminary full genome gene model assembly using RNAseq transcripts.

I used RNA transcripts and V1 genome scaffolds in a basic pipeline to predict both OKI and GBR 'preliminary' gene models. Briefly, the Tuxedo de novo mRNA transcripts were mapped back to the genome scaffolds using PASA (Haas et al. 2011) to find gene model boundary support. Open Reading Frames (ORFs) were generated using Transdecoder (Haas & Papanicolaou 2012). These assemblies were used to train AUGUSTUS (Stanke et al. 2006) to generate parameters for gene model prediction. Finally, these AUGUSTUS training parameters were used to generate gene models for OKI and GBR, respectively. 26,135 protein sequences were predicted for the OKI genome, and 26,586 protein sequences were predicted for the GBR genome.

(c) Final/published gene model assembly by EVM pipeline.

The final gene models for publication were refined and generated by Australian collaborators (Hall et al. 2017). The initial gene models and Augustus parameters generated at OIST were provided to collaborators, along with all raw data. The primary difference with this final approach for gene modelling was the use of an iterative pipeline, based on the EVidence Based Modeler (EVM) software package (Haas et al. 2008), in addition to a custom pipeline for integrating developmental transcriptomes (Fernandez-Valverde et al. 2015). This method allows for new data sets, specifically new RNAseq data, to be incorporated into the final gene model sets in an iterative manner. Once these two sets of gene models were finalized and agreed to by all collaborators, ‘EVM2’ gene models were used for all subsequent analyses.

2.7 Genome analysis methods

In order to analyze assembled and annotated genomes, several standard methods were used. Additionally, several new methods were developed, particularly to compare the two genomes from the separate COTS that were sequenced and assembled, independently.

(a) CEGMA/BUSCO

In order to assess the overall quality of the assemblies, CEGMA (Core Eukaryotic Genes Mapping Approach)(Parra et al. 2007) which maps and scores a conserved set of 248 eukaryotic genes to draft genomes, was used. Support and development for CEGMA was discontinued in May 2015 (http://korflab.ucdavis.edu/Datasets/cegma/#SCT7). Although the authors recommended users switch to BUSCO (Benchmarking Universal Single-Copy Orthologs)(Simão et al. 2015), both methods are reported in the literature, so both methods were performed on the COTS assemblies. Importantly, while CEGMA was originally built for gene model prediction and later updated for use in genome comparisons, BUSCO was purpose built for scoring genome completeness. Both pipelines were followed using default values and protocols.

(b) Whole Genome Alignment: BLAST/LAST whole genomic alignment

In order to compare the overall alignment of the OKI COTS scaffolds to GBR COTS scaffolds, both BLAST+ (Camacho et al. 2009) and LAST (http://last.cbrc.jp) alignment software were used, using default values and protocols.

(c) Macrosynteny Analysis: Gene Liftover

In order to compare the order of gene models between OKI and GBR, we adopted the 'Liftover' pipeline. Liftover was originally designed find one-to-one gene model identity between different builds or releases of the same genome, for example between human genome versions (GRCh38.p7 and GRCh36). The Liftover pipeline works by splitting scaffolds or chromosomes up into 5000 basepair fragments, and then aligning those fragments between the two genomes, providing a coordinate system between the two genomes. Finally, gene models from each assembly can be aligned to each other, based on the coordinate system. I performed an initial attempt with the liftover pipeline at OIST, and found that only around 16,000 gene models (out of 25,000) lifted over. Our Australian collaborators then modified the liftover pipeline, and were able to liftover 20,000 gene models, as was reported in the final COTS genome manuscript (Kent et al. 2002).

(d) SNP calling and analysis

Overall genome heterozygosity was estimated by single-nucleotide polymorphism (SNP) analysis. Because pooled sperm samples were used for short read sequencing, all heterozygous loci were captured in the short-read sequencing. During assembly, these SNPs are lost, as a single haplotype is selected at each loci, based on read quality and read coverage. After genomic scaffolds have been generated, the original reads can be mapped back to the scaffolds. Loci that do not align to the scaffolds, but are supported by multiple over lapping reads are defined as SNPs. SNPs were found by mapping pair-end reads back to the assembled scaffolds using BWA (Li & Durbin 2010), and were called and analyzed using samtools (Li et al. 2009). Additional SNP analysis was also done using vcfstats (Danecek et al. 2011).

(e) Repeats/transposable elements

Both GBR and OKI genomes were masked (e.g. repeats were called) using RepeatMasker version 4.0.3 (Smit et al. 2016)with the following parameters (-qq -pa 8 -gff -species 'fungi/metazoa group' -no_is). This masking process (e.g. ‘hiding’ repetitive regions of the genome) allows for the categorization and analysis of those repetitive regions, based on previously annotated sequences known to function as Transposable Elements. Unknown repeats were identified by blasting RepeatMasker-generated sequences against a manually annotated repeat library, as reported previously (Simakov et al. 2012).

(f) Molecular phylogeny

Short domain-specific regions of the gene models were manually selected (e.g. for Hox genes, 56 amino acids of the homeodomain were selected) and aligned using clustalX (Sievers et al. 2011). These regions were then analyzed and phylogenetic trees were generated using Mega6.06 software (Tamura et al. 2013).

(g) Local alignments (LASTD)

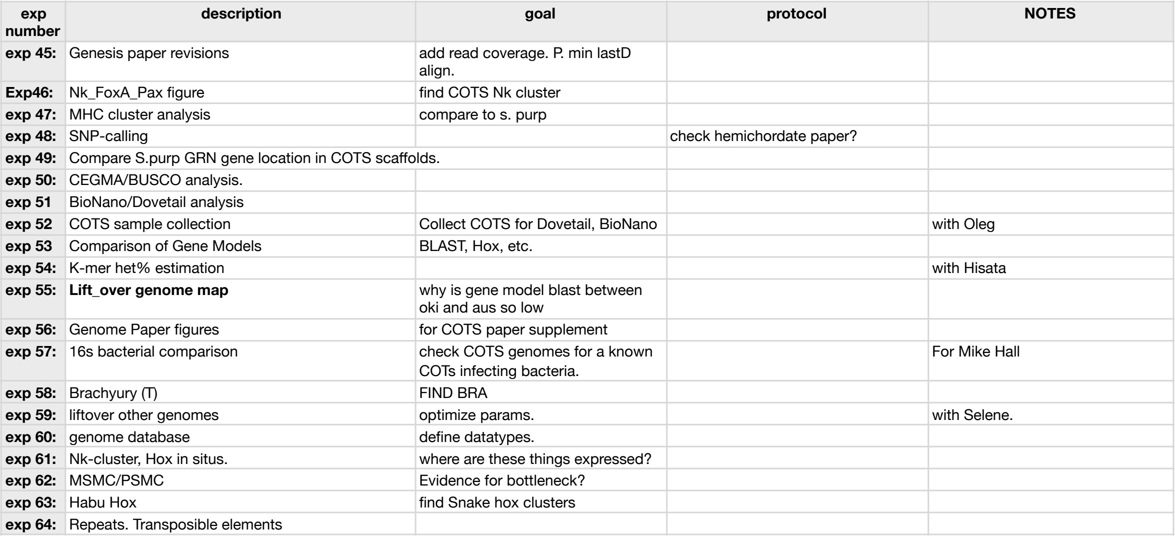
Local alignments between individual genomic scaffolds, both between the two COTS genomes, but also between COTS and the sea urchin (*S. purpuratus*), COTS and the starfish (*P. Miniata*) were performed and visualized using the LASTD software package (http://last.cbrc.jp).

(e) Genome Size estimation by K-mer frequency

In the context of genomics, K-mers are any unique string of nucleotides, where the ‘K’ referrers to the number of nucleotides, or ‘word length’ that exists in a dataset. The distribution of K-mers found for either a genomic assembly, or raw read data can be used to greatly simplify assembly calculations. The distribution of K-mers for a gievn length can be used to estimate genome size. The ‘kmergenie’ tool was used to to select the optimal K-mer length for estimating COTS genome size (Chikhi & Medvedev 2014), which was 35. Jellyfish (Marcais & Kingsford 2011) and R (https://www.r-project.org) were used then graph a histogram of 35-mer frequencies, and estimate genome size, following the methods of: (http://koke.asrc.kanazawa-u.ac.jp/HOWTO/kmer-genomesize.html).

Table 2-1. List of experiments/analyses

exp number	description	goal	protocol	NOTES
exp 1	isolation of genomic DNA from Alga	develop COT gen. DNA protocol	DNA preparation0515.doc	w/Eichi Shoguchi
exp 2	isolation of genomic DNA from Alga	develop COT gen. DNA protocol	COT DNA Preparation V1	redo. increase DNA yeild. culturable dinoflagellate, Symbiodinium minutum
exp 3	COT tissue collection	collect COT sperm from live samples		
exp 4	COT DNA isolation	1st try, used #877 sample	COT DNA Preparation V1.1	lots of DNA, but not clean
exp 5	DNA quantification	nanodrop, agarose gel.		
exp 6	COT DNA isolation#2	took 1week to prep. not much DNA by eye	COT DNA Preparation V1.1	not much DNA?
exp 7	COT DNA isolation#3	“the last one”	COT DNA Preparation V1.1	Looks ok.
exp 8	DNA quantification	confirm exp 6, check Aus DNA, exp 7 DNA		loaded too much DNA...
exp 9	comparison of velvet assembly	try different k-mers to see if assembly is better		with Takeshi Takeuchi
exp 10	COT embryology	observe COT fertilization, development	starfish inesemination.doc	used 1-MA
exp 11	COT embryo imaging Day 2	image COT day 2 embryos.		
exp 12	COT bipinnaria imaging Day 9	image COT day 9 bipinnaria		
exp 13	COT bipinnaria imaging Day 18 lightsheet	Demo Zeiss Z1.lightsheet		
exp 14	COT RNA collection	collect COT RNA from different tissues and developmental stages	ARAKI-san and Eiichi.	
exp 15	COT bipinnaria imaging Day 18	image COT day 18 bipinnaria		
exp 16	MiSeq run OKI03	Run 3 of 5 NGS sequencing of oki COT.	Koyanagi MiSeq Protocol	W/Hisata and Shoguchi
exp 17	sbgn	Run 4 of 5 NGS sequencing of oki COT.	Koyanagi MiSeq Protocol	W/Hisata and Shoguchi
exp 18	COT embryology#2	COT fert./dev, collect RNA and possibly tissue, vary embryo temp. try Yi-jyun’s alga feeding protocol	starfish inesemination.doc	2nd time around. Isolate nervous tis
exp 19	Hox Region Assembly (tblastX:Flava, Kowalevski)		do both OKI and AUS...	with Takeshi Takeuchi, oki_HOX_fromFlava_tblastx.sh.o485880
exp 20	Hox Region Assembly (tblastN:floridae)			aus_B_floridae_Hox_tblastN
exp 21	Hox Region Assembly (tblastN:Mouse, human, fly ascidian)			oki_Hox_Homo_Mus_TblastN
exp 22	Hox Region Assembly (tblastN: Spu_ Sea Urchin)			aus_SPU_Hox_tblastN, oki_SPU_Hox_tblastN
exp 23	Blast Patiria genome for HOX			
exp 24	COT HOX Scaffold alignment.	BLAST aus assembly with oki HOX scaffolds to see if there is an overlap	output: oki_HOX_Scaff_aus_blastN	aus and oki scaffolds are the same...
exp 25	Trinity RNAseq (demo data set)	learn how to use Trinity	1. Haas, B. J. et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc 8, 1494–1512 (2013).	be very careful with typos...
exp 26	FastQC of COT RNAseq raw reads	check RNAseq raw reads	(FROM HTSA)	
exp 27	trimmomatic of RNAseq data, 2nd fastQC	trim RNAseq reads		
exp 28	tophat demo			
exp 29	tophot and cuff links: RNAseq genome refence	check RNAseq trimmed reads		
exp 30	run_trinity on tissues	run trinity on all samples		
exp 31	GenomeGuidedTrinity	run GG trinity on RNA samples from Genome seq.		meh. same number of transcripts as tuxedo.
exp 32	Trinity comparisons, Analysis	compare the RNA assemblies from Triniity		
exp 33	Scaffolding with SSPASE	Extent the PE sequences with MP data.		
exp 34	1-MA blast of Aus_SSPACE scaffolds.	find the 1-MA genes.		
exp 35	BLAT: aus.SSPACE with aus.Trinity.all data	What % of trinity trascripts map to genome.		
exp 36	TOPHAT: align aus.SSPACE with aus.Trinity.all data			
exp 37	HOX: phylogen (CLUSTAL, with Eiichi)	phylogenetics		with eiichi
exp 38	Determination of Nerve tissue in RNAseq data	check for neuronal specific genes.		
exp 39	Augustus Gene Prediction	MAKE GENE MODELS	see Eiichi’s paper, also coral paper.	with Hisata, Kostya
exp 40	P. min Hox analysis	Confirm Hox4 versus Hox6		
exp41	compare Aus and Oki scaffolds (MUMMER, LAST, BLAT)	compare raw scaffolded sequence.		
exp42	gapFiller/gapCloser	fill in “N’s”		NOT NEEDED.
exp43:	**Mox.blast**	Do COTS have MOX?		
exp 44:	phoronid HOX	check phoronid assembly.		



exp number	description	goal	protocol	NOTES
exp 45:	Genesis paper revisions	add read coverage. P. min lastD align.		
Exp46:	Nk_FoxA_Pax figure	find COTS Nk cluster		
exp 47:	MHC cluster analysis	compare to s. purp		
exp 48:	SNP-calling		check hemichordate paper?	
exp 49:	Compare S.purp GRN gene location in COTS scaffolds.			
exp 50:	CEGMA/BUSCO analysis.			
exp 51	BioNano/Dovetail analysis			
exp 52	COTS sample collection	Collect COTS for Dovetail, BioNano		with Oleg
exp 53	Comparison of Gene Models	BLAST, Hox, etc.		
exp 54:	K-mer het% estimation			with Hisata
exp 55:	**Lift_over genome map**	why is gene model blast between oki and aus so low		
exp 56:	Genome Paper figures	for COTS paper supplement		
exp 57:	16s bacterial comparison	check COTS genomes for a known COTs infecting bacteria.		For Mike Hall
exp 58:	Brachyury (T)	FIND BRA		
exp 59:	liftover other genomes	optimize params.		with Selene.
exp 60:	genome database	define datatypes.		
exp 61:	Nk-cluster, Hox in situs.	where are these things expressed?		
exp 62:	MSMC/PSMC	Evidence for bottleneck?		
exp 63:	Habu Hox	find Snake hox clusters		
exp 64:	Repeats. Transposible elements			

Chapter 3 : Results

3.1 Decoding and general analysis of COTS genome

3.2 Analysis of the COTS Hox and ParaHox clusters

3.3 Analysis of the COTS Nkx pharyngeal-gill-slit-related gene cluster

3.4 Systems Biology analysis of COTS 1-MA-dependent oocyte maturation

3.1 Decoding and general analysis of COTS Genome

3.1 Decoding and General Analysis of COTS Genome

The following section is adapted and updated from: (Hall et al. 2017)

The final assembly (V1.0) of the GBR genome was 383,525,304 base pairs long across 3274 scaffolds with an N50 of 917kb, while the OKI genome was 383,843,944 base pairs long across 1765 scaffolds with an N50 of 1,521kb (Table 3.1). N50 refers to the length of the scaffold that covers 50% of the total genomic length, when all scaffolds are sorted, aligned by length, and added together (Nagarajan & Pop 2013). Thus, a longer N50 implies that longer scaffolds cover a higher percentage of a genome assembly, and generally indicates a higher quality assembly. In addition, a low number of scaffolds in total indicates that relative to other published genomes (see Table 3.3), both of the COTS assemblies do not suffer from excessive fragmentation. Finally, the number of gene models validated for both respective genomes indicates that in addition to having a non-fragmented assembly, the assembled sequence is biologically meaningful.

Table 3-1. COTS Genome summary.
Summary of COTS genomic sequencing: Great Barrier Reef, Australia ("GBR") and Okinawa, Japan ("OKI").

• Specimen name	*A. planci* "gbr"	*A. planci* "oki"
• Scaffold total length (bp)	383,525,304	383,843,944
• Scaffold N50 (bp)	916,880	1,521,119
• No. of scaffolds	3274	1765
• No. of genes	24,747	24,323

(a) Genome size estimation

COTS genome size was estimated by three methods. By k-mer analysis of the raw read data, I calculate the size of GBR genome to be 441 Mb, and that of OKI to be 421 Mb. Flow cytometry estimated the OKI genome to be 480 Mb. Finally, the total length of both scaffolded assemblies was 384mb. Thus, we estimate the COTS genome to be 400-450 Mb in total length.

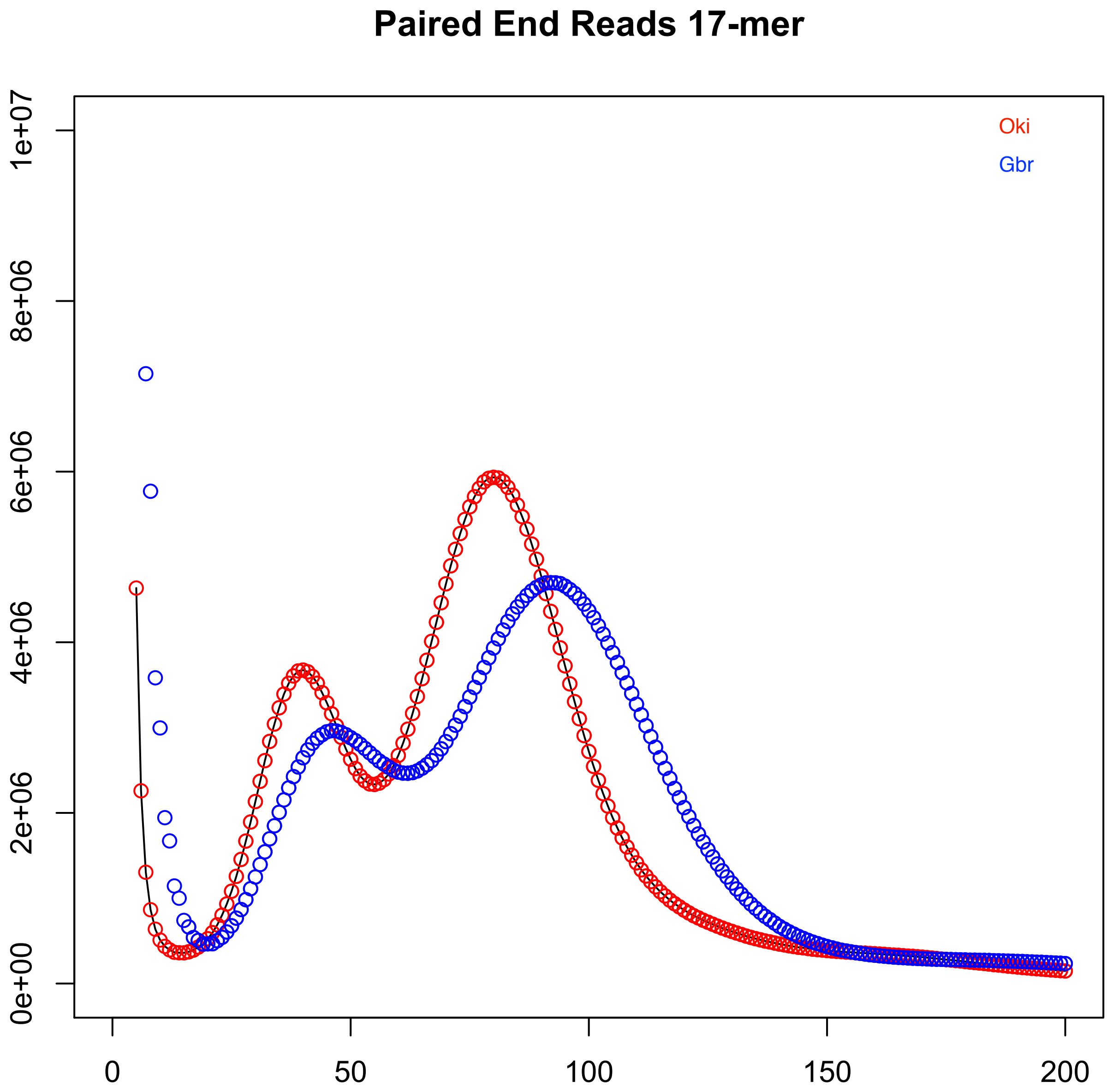


Figure 3.1. K-mer (17-mer) plot.
The GBR genome was estimated to be 441mb long, while the OKI genome was estimated to be 421mb.

(b) Genome assembly completeness by CEGMA and BUSCO methods

Two standard methods have been developed for analyzing the quality of a genomic assembly, and were used to analyze the COTS assemblies, specifically CEGMA (Parra et al. 2007)and BUSCO (Simão et al. 2015). Both methods search for a predetermined set of genes that are common to all clades or metazoans, and score the presence or absence of these sequences. Although CEGMA is no longer supported (see methods section 2.4A), both CEGMA and BUSCO scores are reported in Table 3.2. The CEGMA results are in line with other genomes sequenced and annotated in the OIST Marine Genomics Unit. Although the BUSCO genomic scores for both COTS genomes are approximately 20% lower than that of sea urchin (*S. purpuratus*), the COTS gene models score 10% higher than the sea urchin gene models. Unfortunately, the meaning of a low CEGMA or BUSCO is very much up to debate, as lower scores could be caused by poor assembly, or by evolutionary divergence (e.g. the assembled genome is correct, but divergent from the selected 248 or 429 reference sequences, for CEGMA and BUSCO, respectively). Nevertheless, in comparison to genomic assemblies with extensive sequencing and annotation efforts, the BUSCO scores from both COTS assemblies confirming the high quality of the assemblies.

Table 3-2. COTS CEGMA, BUSCO scores
Comparison of Genome Sequencing Completeness: BUSCO, CEGMA
(Data: https://www.ncbi.nlm.nih.gov/genome/browse/, http://busco.ezlab.org/v1/files/BUSCO-SOM.pdf)

• Specimen name	COTS“gbr”	COTS “oki”	*S. purpuratus*	*H.sapiens*	*C. elegans*	*D.melangaster*
• Scaffold total length (mb)	384	384	1,032	3238	100	144
• Scaffold N50 (kb)	917	1,521	0.431	59,364	N/A	23,011
• No. of scaffolds	3274	1765	31,879	831	7	1,870
• No. of genes	24,747	24,323	31,871	59,911	46,728	17,682
CEGMA (248 genes)						
• Complete genes (#)	178	184	**(N/A)**	**(N/A)**	**(N/A)**	**(N/A)**
• Completeness (%)	**71.77**	**74.19**				
• Partial genes (#)	236	236				
• Completeness (%)	95.16	95.16				
BUSCO Genome Score (eukaryota)			GCA_000002235.2	(GCA_0000 01405.15)	(GCA_00000 2985.3)	Dmel_r5.55
• C:complete	**C:66%**	**C:64%**	**C:87%**	**C:89%**	**C:85%**	**C:98%**
• [D:duplicated]	[D:2.0%]	[D:1.6%]	[D:6.5%]	[D:1.5%]	[D:6.9%]	[D:6.4%]
• F:fragmented	F:4.4%	F:6.0%	F:7.8%	F:6.0%	F:2.8%	F:0.6%
• M:missed	M:28%	M:29%	M:4.9%	M:4.5%	M:11%	M:0.3%
• n:genes	n:429	n:429	n:843	n:3023	n:843	n:2675
BUSCO Gene Models Score (eukaryota)		Oki.EVM2	GCA_000002235.2.22	GRCh37.75	WBcel235.22	Dmel_r5.55
• C:complete	**(N/A)**	**C:96%**	**C:83%**	**C:99%**	**C:90%**	**C:99%**
• [D:duplicated]		[D:30%]	[D:19%]	[D:1.7%]	[D:11%]	[D:9.1%]
• F:fragmented		F:2.3%	F:15%	F:0.0%	F:1.7%	F:0.2%
• M:missed		M:0.6%	M:0.7%	M:0.0%	M:7.5%	M:0.0%
• n:genes		n:429	n:843	n:3023	n:843	n:2675

(c) COTS genomes compared to other marine invertebrate genome assemblies

A comparison of the two COTS genomes to previously published marine invertebrate deuterostome genomes is summarized in Table 3.3, which was updated from an previously published table(Cameron et al. 2015). Table 3.3 summarizes marine invertebrate deuterostome genomes assembled to date, and highlights that the COTS genomes have higher scaffold and contig N50 values, as well as lower scaffold and contig counts, as compared to all other non-congenic species. Notably, COTS have the highest values of any echinoderm sequenced to date. The final assemblies for both COT samples are of remarkable high quality with respect to genomic scaffolds, which is likely due to low genomic homozygosity within each individual genome. These data are inconclusive with regard to addressing whether recently population density dynamics have resulted from anthropomorphic causes, but low heterozygosity both within each genome and between the two assemblies (OKI versus GBR) is consistent with the notion that population density has increased dramatically in the recent past.

Table 3-3. Comparison of marine genome assemblies.
Genomic assembly statistics for marine invertebrate deuterostomes. Updated from (Cameron et al. 2015).

Species Name, genome version	phylum	common name	GenBank access.#	Total length (Mb)	Scaffold number	Scaffold N50 (kb)	Contig number	Contig N50 (kb)	GC (%)	Genes (#)	Reference
Acanthaster planci (COTS), gbr-v1.0	Echinodermata	COTS, Australia	DRA004862	383	3,274	916	17,868	54.9	41.31	24,747	Hall et al (sub. 2016)
Acanthaster planci (COTS), oki-v1.0	Echinodermata	COTS, Japan	DRA004863	383	1,765	1,521	17,265	54.7	41.30	24,323	Hall et al (sub. 2016)
Patiria miniata, v1.0*	Echinodermata	bat star	GCA_000285935.1	811	60,183	53	179,756	9.4	40.20	29,697	Cameron et al. (2015)
Strongylocentrotus purpuratus, v4.8*	Echinodermata	purple sea urchin	GCA_000002235.3	1032	31,879	431	140,454	17.6	38.30	31,871	Cameron et al. (2015)
Lytechinus variegatus, v2.0*	Echinodermata	green sea urchin	GCA_000239495.2	1061	322,936	46	481,804	9.7	36.40	28,204	Cameron et al. (2015)
Saccoglossus kowalevskii, v1.1	Hemichordata	acorn worm, direct dev.	GCA_000003605.1	758	7,282	552	20,913	89	38	34,239	Simakov et al. (2015)
Ptychodera flava, v0.6	Hemichordata	acorn worm, indirect dev.	GCA_001465055.1	1,229	218,255	196	322,077	7.6	37	34,647	Simakov et al. (2015)
Ciona intestinalis, vKH*	Chordata	tunicate, sea squirt	GCA_000224145.2	115	1,280	3,102	6,381	37	36.02	14,983	Satou et al. (2008)
Ciona savignyi*	Chordata	transparent sea squirt	GCA_000149265.1	587	34,009	601	74,923	23	37.10	-	Small et al. (2007)
Botryllus schlosseri *	Chordata	golden star tunicate	GCA_000444245.1	580	120,139	7	130,124	7	40.60	-	Voskoboynik et al. (2013)
Oikopleura dioica*	Chordata	pelagic tunicate	GCA_000209535.1	70	4,196	22	6,678	11	39.90	13,505	Denoeud et al. (2010)
Branchiostoma floridae, v2.0*	Chordata	Amphioxus, Lancet	GCA_000003815.1	522	398	2,587	41,927	28	41.20	28,627	Putnam et al. (2008)

* Updated from: Cameron, R. A., Kudtarkar, P., Gordon, S. M., Worley, K. C. & Gibbs, R. A. "Do echinoderm genomes measure up?" Marine Genomics (2015). doi:10.1016/j.margen.2015.02.004

(d) Overall comparison between OKI and GBR genome assemblies

Comparison of the general assembly characteristics from OKI and GBR genomes indicate that both genomes are very similar to each other (Table 3.1). The overall scaffold lengths are within 0.1% of each other (~300 Kb difference between 383 Mb genomes), which is remarkable, considering that the source genomic DNA was collected from two different biological samples collected over 5000 kilometers apart, and the degree to which the final assembly characteristics differed. The total number of scaffolds and the N50 were about twice as complete for the OKI genome (1765 versus 3274 scaffolds, and 1.5 Mb versus 0.9 Mb for N50). Lastly, the total gene model number is also within 0.1% (~300 gene models difference over ~24,500 gene models), suggesting that the overall similarity between the assembled sequences was maintained after integrating RNAseq data, which were more divergent.

The similarity between the assembled genomes indicates that both individual starfish (OKI and GBR) are from the same species complex. Previous COTS population genetics studies have suggested that the overall COTS species contains at least 4 subspecies, but that the pacific clade represents a single grouping (Yasuda et al. 2014). Our genome assembly data is consistent with this result. Initially, I hypothesized that differing ecological constraints between the largely continuous Great Barrier Reef and the smaller, more punctate Okinawan reefs, in addition to differing biocontrol policies may lead to speciation or divergence. However, the overall genome assembly statistics do not provide evidence to support any significant differences between the two genomes, or cryptic speciation.

(e) Scaffold alignments between OKI and GBR genome assemblies

In order to determine how diverged the two genomes (OKI and GBR) were from each other, each assembly was aligned to the other using either BLAST and LASTAL software packages. Using this approach, it is possible to estimate the overall heterozygosity between the two assemblies. Because of the high similarity between the two assemblies, the unfiltered output files were very large. For example, both genomes files were 373 Mb, but BLAST and LASTAL output files, unfiltered, were over 30 gigabytes, or around 85 times larger. The reason for these large output files sizes is that both alignment software packages include all possible alignments, and thus include both all possible shortened versions for each alignment region, as well as low quality alignments for regions that may resemble each other, but are otherwise not corrected assigned. Both types of errors, redundancy versus incorrect alignment, lead to a dramatic increase in meaningless or incorrect alignments.

Graphing the BLAST output files in excel, using pivot tables to batch and sort by scaffold length, led to two main observations (Figure 3.2): The first was that majority of alignments were 100%. Second, there was an increased number of alignments between 98.5% and 95.5%, this increase density is circled in red, in Figure 3.2. By filtering the alignments either by length (e.g. only alignments longer than 10 kb), or by % identity (great than 95% identity), both genomes had 98.8% nucleotide identity between GBR and OKI genomes (Figure 3.3).

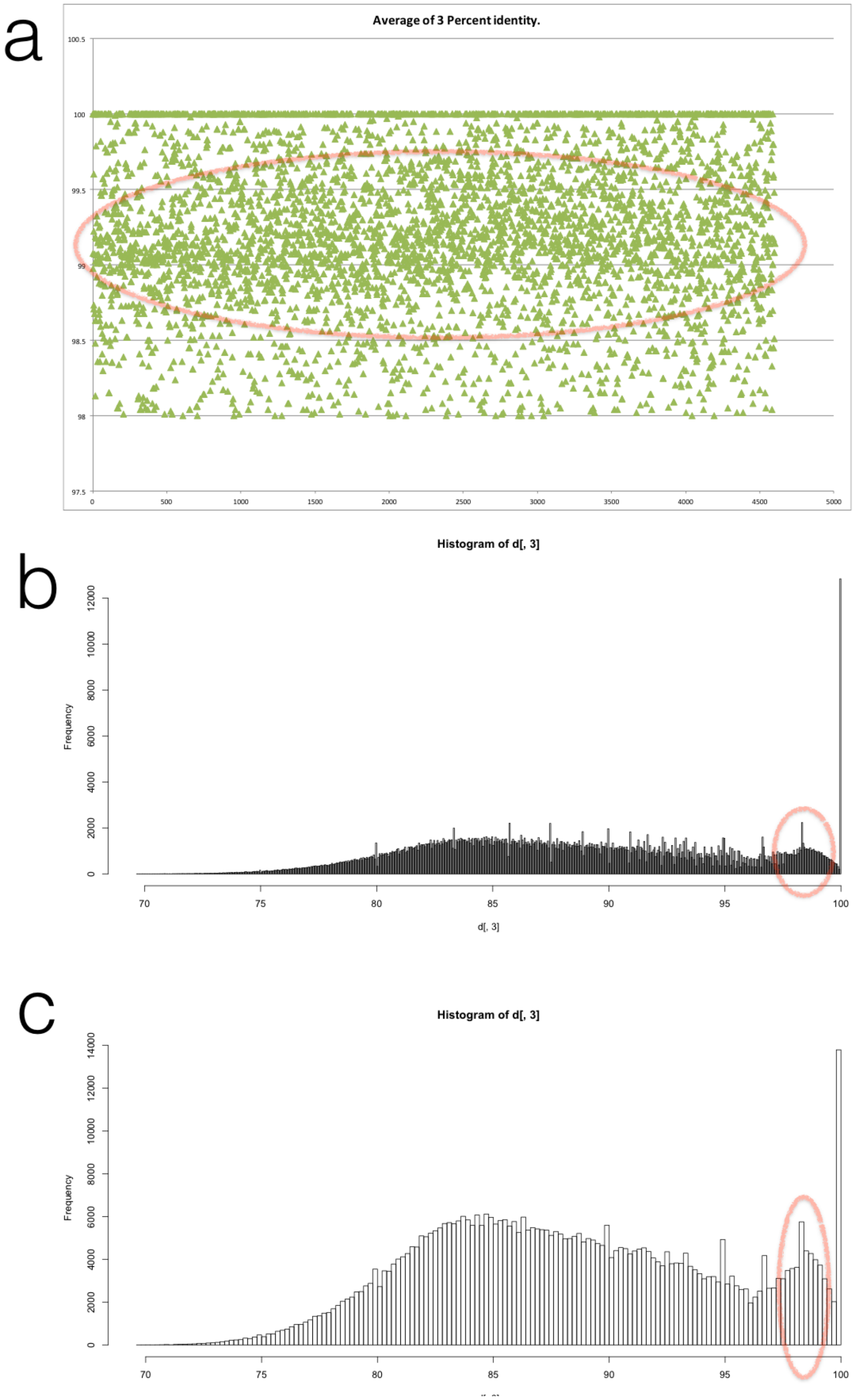
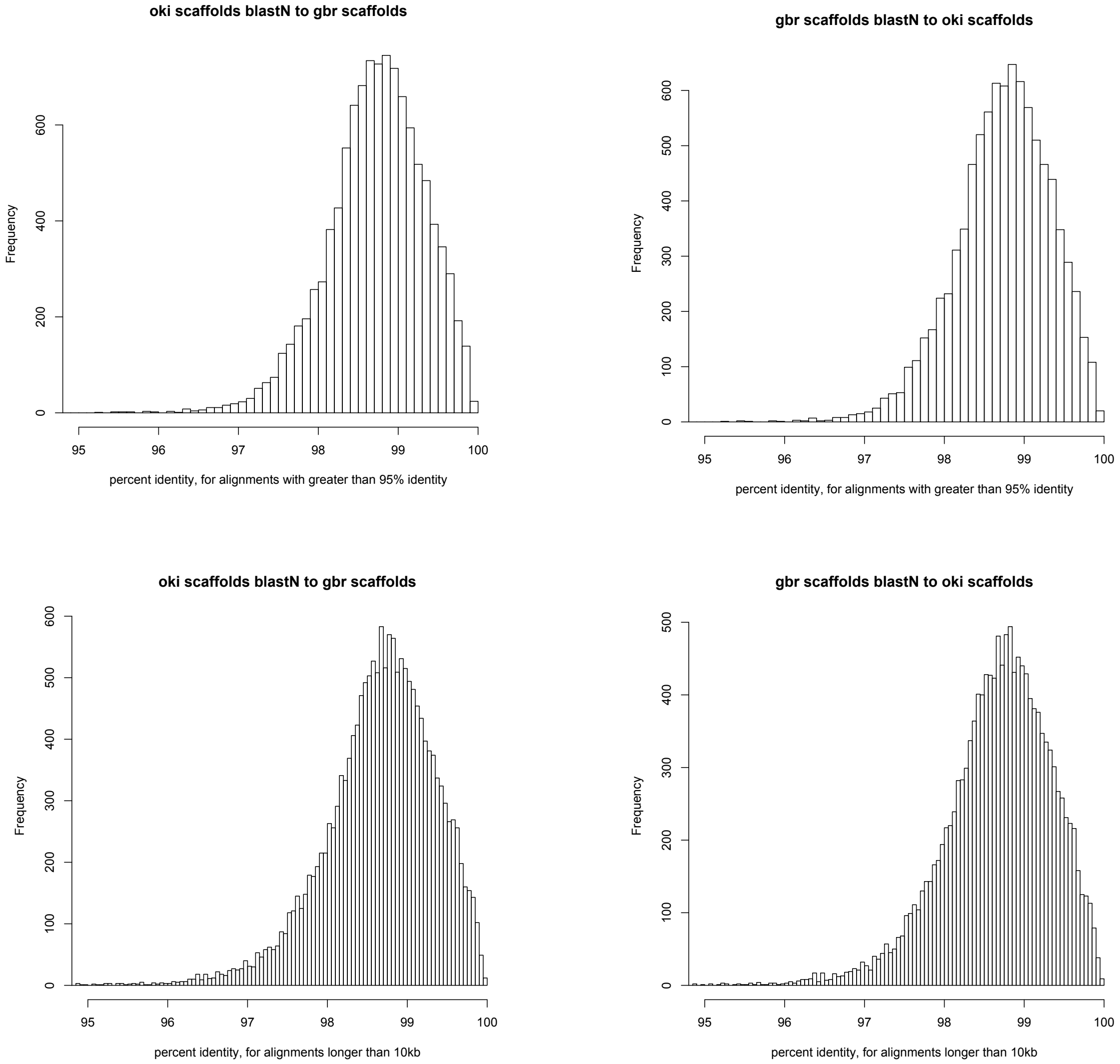


Figure 3.2. Calculating COTS genomic heterozygosity.
Determination of % heterozygosity based on % alignment between OKI scaffolds-aligned by BLAST onto GBR scaffolds. **a,** a MS excel scatterplot where the x-axis is %identity, and the y-axis is scaffold number. The red circle highlights a region between 98.5% and 99.5% identity. Note that the data are arbitrarily cut off at 96%, to prevent excel from crashing. **b,** Histogram of % identity (same data from figure 3.2a), batched into 500 groups. **c,** Histogram of % identity (same data from figure 3.2a), batched into 100 groups.



oki scaffolds blastN to gbr scaffolds

gbr scaffolds blastN to oki scaffolds

oki scaffolds blastN to gbr scaffolds

gbr scaffolds blastN to oki scaffolds

Figure 3.3. Inter-genomic heterozygosity by BLASTN alignment.

BLASTN was used to align OKI and GBR scaffolds, and to generate histograms for alignments with greater than 95% identity, or longer than 10,000 base pairs. The mean value is 98.721% and the median is 98.77% for GBR scaffolds aligned to OKI scaffolds longer than 10kb, and the mean value is 98.670% with a median of 98.74% for OKI reads aligned to GBR scaffolds, longer than 10kb.

(f) Single Nucleotide Polymorphism (SNP) analysis

Overall genome heterozygosity was also estimated by single-nucleotide polymorphism (SNP) analysis. During scaffold assembly, single nucleotide heterozygosity is collapsed into a single genotype, which can be recovered by mapping the the processed pair-end reads back to the reference genome. Additionally, because two genomes were sequenced in parallel, OKI reads were mapped to OKI scaffolds, AUS reads were mapped to AUS scaffolds, AUS reads were mapped to OKI scaffolds, and OKI reads were mapped to AUS scaffolds.

The internal SNP rate was 0.91722% for OKI and 0.87526% for GBR, while overall SNP rate from mapping OKI reads to GBR scaffolds was 1.42184%, and from mapping GBR reads to OKI scaffolds was 1.36604% (Figure 3.4a). These SNP rates matched the hetereozygocity rate as measured by blast alignments. Of the common SNPs, 64.5% of GBR SNPs and 64.2% of OKI SNPs were common to both sets of reads, which was slightly below the expected rate of 66.7%, consistent with reduced overall heterozygosity.

In order to determine the likely origin of SNPs, I counted the number of SNPs per a 100 basepair window, taken at 50 bp increments along the respective alignments (Simakov et al. 2015). The resulting histograms of SNP count (Figure 3.4c) can be best fit to either a geometric or Poisson distribution. COTS genomes show a geometric distribution of SNPs, which suggests that SNPs are caused by recombination and not random mutation, consistent with overall low genomic heterozygosity.

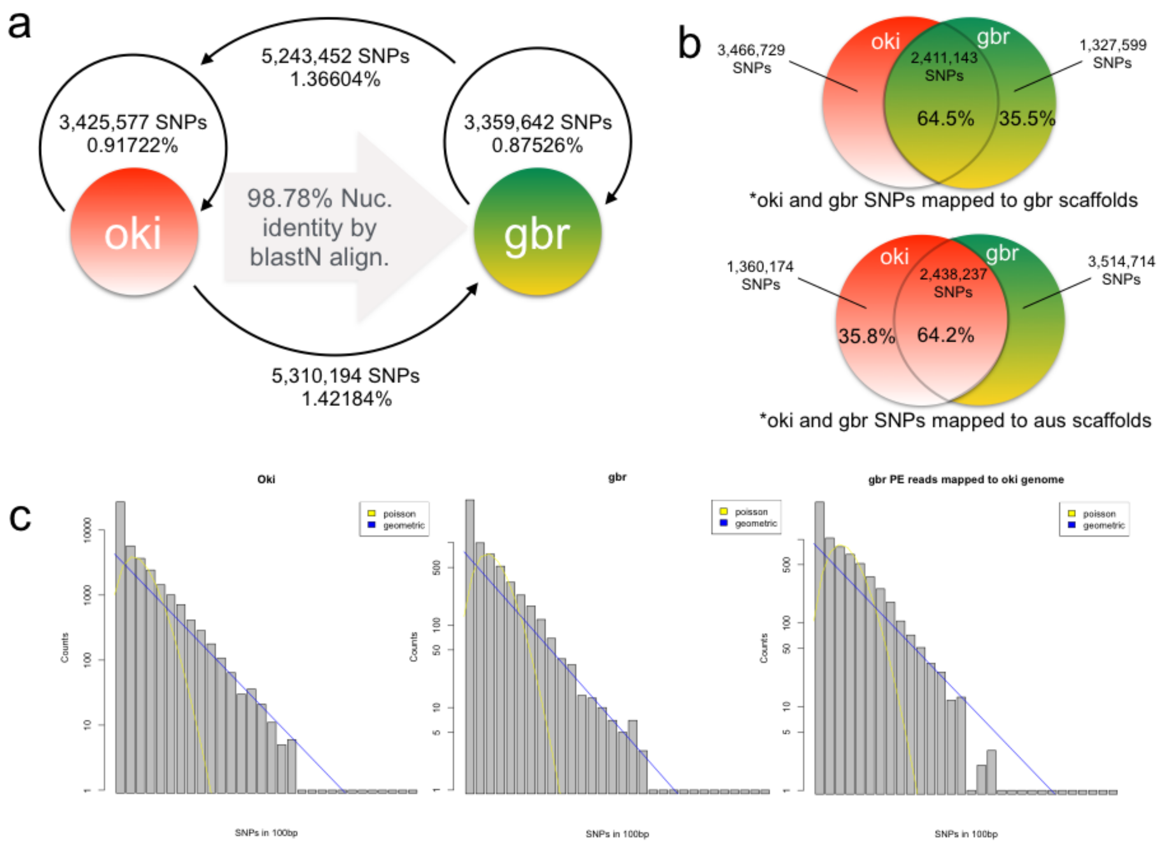


Figure 3.4. Single Nucleotide Polymorphism (SNP) Analysis.
a, Overall SNP rates, by genome. **b,** Unique and shared SNPs for a given genome assembly, based on complimentary reads (e.g. OKI reads mapped to GBR scaffolds). **c,** Histograms of SNPs counts in a sliding 100 bp window, to confirm that heterozygosity likely arose from point mutations (geometric) versus recombination (Poisson).

(g) Repeats/transposable elements

Overall, 23.36% of the GBR and 23.38% of the OKI genomes were masked. We noted that “unclassified” masking covered 17.56% of the Gbr and 18.43% of the Oki genomes, respectively (Table 3.4). Overall, the type and distribution of annotated repeats was not markedly different from either sea urchin or bat star, nor were there any significant differences between the OKI and GBR genomes (Figure 3.5). Initially, the percentage of total masking (~23.4%) for both COTS genomes appeared to be low, but subsequent analysis confirmed that COTS repeats were not significantly different from other genomes, when those genomes were masked and annotated correctly.

Although the majority of repeats remain 'Unknown,' several subtypes were found in one genome and not the other (Table 3.5), though the total coverage of these annotated repeats represent around 1% of the respective genomes, and re-masking with annotated repeats only masked less that 2% of the respective genomes.

Table 3-4. . RepeatMasker Output

	GBR			**OKI**		
Name	Count	bp masked	% genome	Count	bp masked	% genome
SINEs:	10753	1629226	0.42	16134	2737298	0.71
ALUs	0	0	0	0	0	0
MIRs	2138	424482	0.11	2634	468590	0.12
LINEs:	21211	4278930	1.12	19331	3846292	1
LINE1	0	0	0	0	0	0
LINE2	14863	2034773	0.53	15036	2267302	0.59
L3/CR1	1221	368776	0.1	851	238042	0.06
LTR elements:	10391	4440077	1.16	6210	3296512	0.86
ERVL	0	0	0	0	0	0
ERVL-MaLRs	0	0	0	0	0	0
ERV_classI	0	0	0	0	0	0
ERV_classII	0	0	0	0	0	0
DNA elements:	24817	8419905	2.2	16774	6285024	1.64
hAT-Charlie	0	0	0	0	0	0
TcMar-Tigger	3870	679696	0.18	0	0	0
Unclassified:	305759	67341409	17.56	311960	70753089	18.43
Total interspersed repeats:	N/A	86109547	22.45	N/A	86918215	22.64
Small RNA:	1311	283802	0.07	1908	353680	0.09
Satellites:	1310	820594	0.21	0	0	0
Simple repeats:	43265	2273675	0.59	42464	2410686	0.63
Low complexity:	6764	328833	0.09	6963	336733	0.09

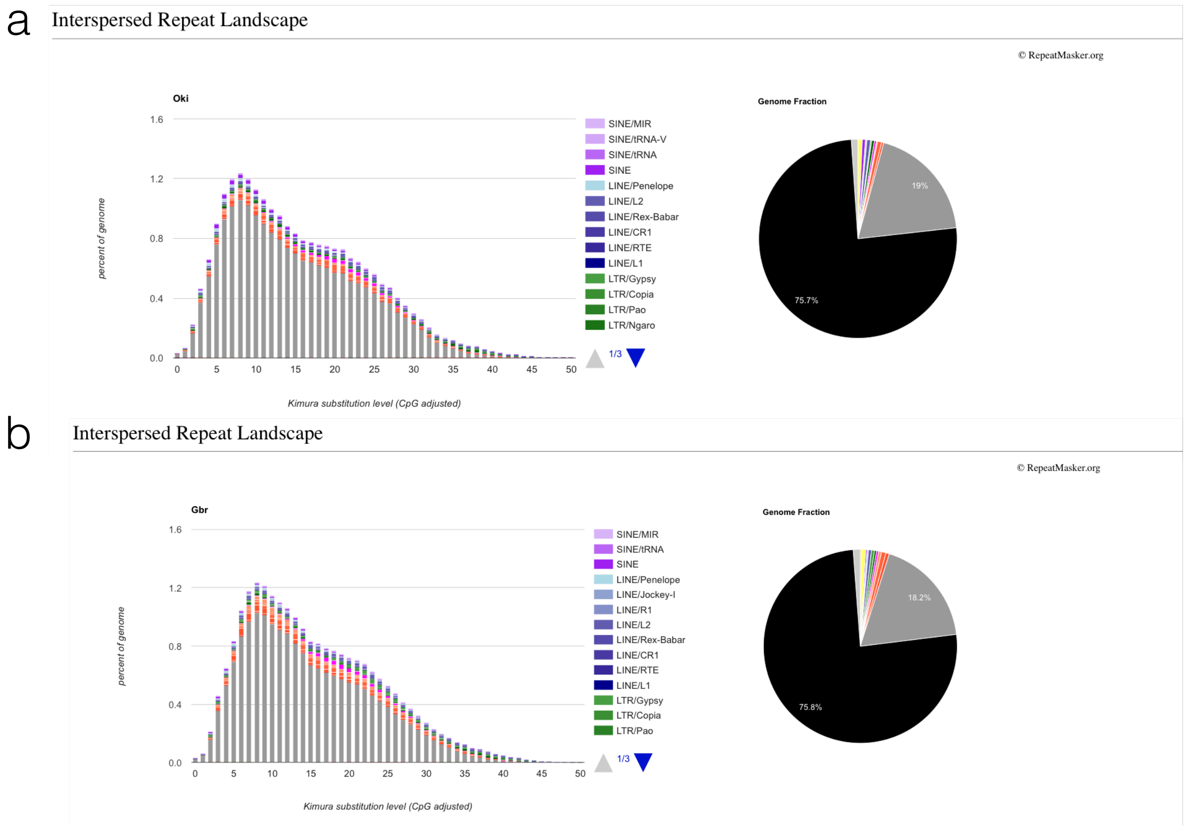


Figure 3.5. COTS Repeat types.

Top 14 Transposable Elements Repeat types**. a,** OKI **b,** GBR

Table 3-5. COTS repeats, by alignment length.
Alignment length for repeats based on a manually annotated library from (Simakov et al. 2015).

Repeat Name	Gbr alignment length (bp)	Oki alignment length (bp)
Unknown	26223	23405
LINE/L2	4126	3776
LINE/CR1	2573	4879
LINE/Penelope	2067	2588
DNA/PiggyBac	1522	1423
LTR/Gypsy	1109	2795
LTR/Pao	1026	2900
LTR/Gypsy-Gmr1	942	#N/A
LINE/L1-Tx1	932	474
LTR/DIRS	928	606
LINE/Rex-Babar	798	813
LTR/Gypsy-Cigr	763	2301
DNA/TcMar-Tc1	704	993
DNA/Maverick	559	594
LINE/RTE-BovB	519	1445
LTR/Ngaro	450	285
LINE/I-Nimb	200	622
RC/Helitron	178	66
SINE?	175	124
DNA/P	160	#N/A
DNA/IS4EU	132	83
DNA/Chapaev	132	#N/A
DNA	130	309
DNA/hAT-Tip100	121	124
DNA/Crypton	118	#N/A
DNA/hAT-Blackjack	105	305
DNA/Ginger	94	49
DNA/PIF-Harbinger	84	240
Simple_repeat	81	412
DNA/hAT-Ac	78	105
SINE/tRNA	70	143
SINE	54	27
DNA/hAT-hAT5	54	126
LINE/RTE-X	53	#N/A
LTR/Copia	42	#N/A
SINE/V	40	#N/A
DNA/MULE-MuDR	33	36
SINE/MIR	26	42
DNA/MULE-F	25	#N/A
DNA/hAT	24	67
DNA/Zator	12	49
LTR/DIRS?	#N/A	31
DNA/Sola	#N/A	42
Satellite	#N/A	43
DNA/TcMar-Tigger	#N/A	54
LINE/I	#N/A	54
SINE/B2	#N/A	64
DNA/Academ	#N/A	66
DNA/TcMar-ISRm11	#N/A	222

(f) Gene model Liftover (mapping OKI genes to GBR genes)

16,004 OKI gene models were lifted over to GBR scaffolds, and 16,370 GBR genes lifted over to OKI scaffolds. This compares to 20,055 GBR gene models that blast align to OKI scaffolds with greater 95% ID and e-value less than E-10. In other words, the Liftover pipeline, which involves splitting up scaffolds into 3 kb chucks and aligning them to each other, and then assigning gene models on a one-to-one basis based on those coordinates, recovered fewer gene models than simply blasting gene models against scaffolds. Subsequent optimization was done by Australian collaborators, with the final, optimized gene liftover resulting in around 22,000 gene models being assigned between OKI and GBR (Hall et al. 2017).

(f) Genomic polishing (Dovetail and BioNano)

In order to extend the genome scaffold assembly, both Dovetail and BioNano genome polishing protocols were used, incorporating fresh genomic DNA with the OKI V1.0 genome assembly scaffolds. The final Dovetail assembly was 384 Mb long, across 730 scaffolds, with an N50 of 4.44 Mb, representing one of the best improvements from the Dovetail ‘Chicago library’ method. The BioNano method integrated the Dovetail assembly with BioNano scaffolds, resulting in a 385 Kb assembly over 718 scaffolds, with an N50 of 4.9 Mb.

(h) RNAseq transcriptomes

RNA Transcriptomes were collected from testes, podia, spines, and stomach/mouth tissue from the individual specimens used for genomic DNA isolation, as well as from nerve and developmental tissues from other specimens, collected at later dates (Table 3.6). Comparison of RNA transcript expression level by tissue, between OKI and GBR confirm the high quality of transcript assembly (Figure 3.6).

Table 3-6. COTS RNAseq assembly.
Comparison of Trinity (*de novo*) versus Tuxedo (*Genome guided*) RNA transcriptome assembly.

Location	Tissue	Trinity (*de novo*)				Tuxedo (*Genome guided*)		
	* - Gbr genome sequenced, º - Oki genome sequenced	Genes (#)	Isoforms (#)	Contig N50	GC (%)	Genes (#)	Isoforms (#)	Aligned/paired reads (%)
• Gbr	Testis*	103915	193591	3440	44.22	27819	35469	78.3
• Gbr	Podia*	96841	153629	3043	43.64	23083	30145	78.7
• Gbr	Spine*	70975	97780	1949	40.97	21105	24780	76.7
• Gbr	Stomach*	91997	154134	3132	44.16	23104	29842	78.7
• Gbr	Body Wall*	74119	103046	1774	40.55	23833	27789	78.5
• Gbr	(All Gbr reads)	93094	153191	3255	43.72	29635	52365	N/A
• Oki	Testisº	40482	35852	811	42.32	18857	22387	73.2
• Oki	Podiaº	85307	56760	2642	42.94	22215	28768	74.1
• Oki	Spineº	104055	64509	2833	43.16	24289	31576	73.2
• Oki	Mouthº	25147	22322	801	38.47	13065	13681	72.7
• Oki	Nerve-Female#1	73842	53860	3006	43.41	21244	27173	66.4
• Oki	Nerve-Female#2	67649	50909	3006	43.32	22211	26848	65.5
• Oki	Nerve-Male#1	78054	56489	2352	43.15	25124	31221	65.0
• Oki	Oocyte	164663	118728	1425	41.80	51470	55967	62.6
• Oki	Early Gastrula	75552	49745	2306	43.29	21244	27173	64.0
• Oki	Middle Gastrula	147017	82413	2772	43.38	29068	36306	63.1
• Oki	(All Oki reads)	186200	110737	2853	43.11	33036	69261	N/A
• Oki/Gbr	All RNAseq reads	259329	147429	3171	43.44	N/A	N/A	N/A

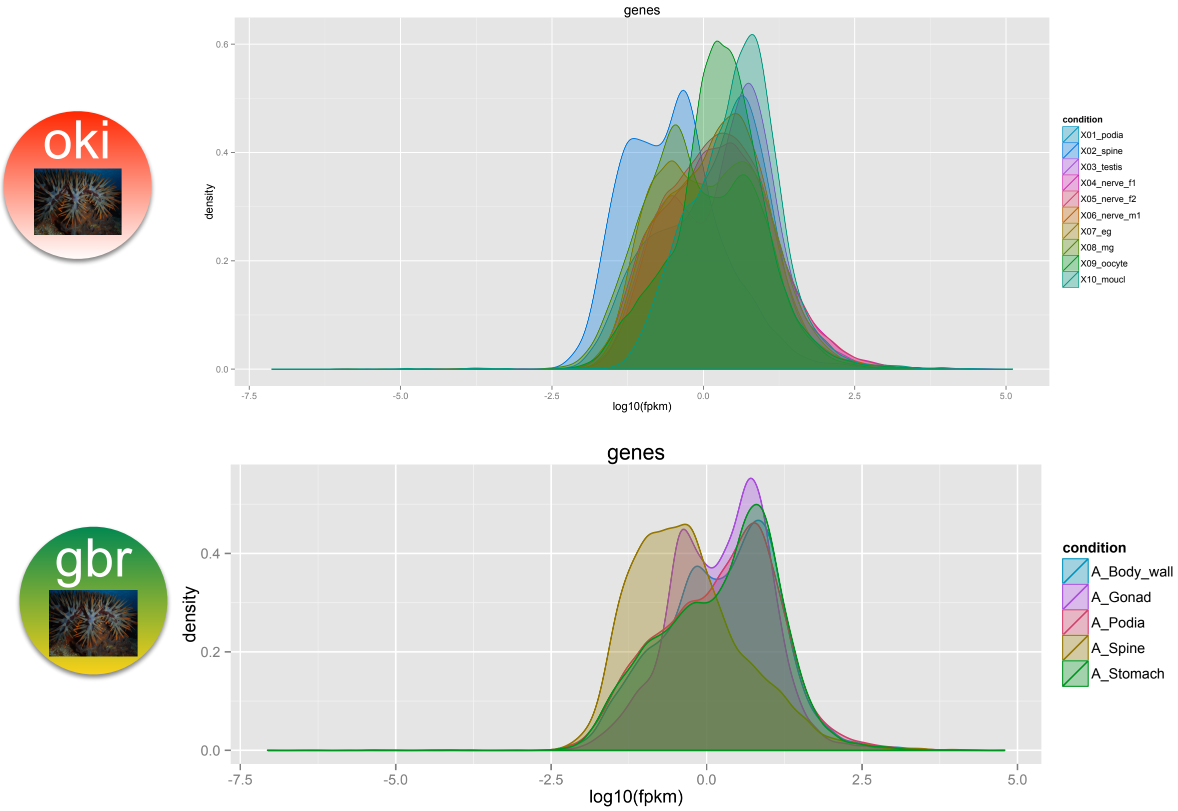


Figure 3.6. Histograms of Tuxedo RNA transcript expression level.
Histograms of Tuxedo genome-guided transcript express, by tissue/sample type confirm a general overlap of expression patterns between oki and gbr. For example, compare X02_spine (top, OKI, blue) versus A_spine (bottom, GBR, brown), or X03_testis (OKI, purple) versus A_Gonad (GBR, purple).

(i) COTS Mitochondrial and Bacterial 16S Alignments.

In order to determine the quality of the V1.0 COTS genome assemblies, the published COTS mitochondrial genome (gi|86476000|dbj|AB231475.1| Acanthaster planci mitochondrial DNA, complete genome) (Yasuda et al. 2006) was aligned to both OKI and GBR genomes using BLASTN. OKI scaffold570 and GBR scaffold845 aligned almost perfectly, as visualized by the LAST alignment (Figure 3.7). Interestingly, this result contrasts with a recent report of divergent mitochondrial genomes between lingual specimens, presumably from the same species (Y.-J. Luo et al. 2015).

Additionally, a bacterial 16S rRNA tag sequence was provided by Prof. Lone Høj (AIMS) for a that is dominant (97% relative abundance) in male gonads of COTS, likely to be an intracellular bacterium. This tag was aligned to both V1.0 genomes using BLASTN, Althought alignments to OKI scaffold#215 were found, but no alignments to GBR were apparent. OKI scaffold#215 was 450 Kb in length, contains no gene models, and has a GC% of 30.9%. In contrast, OKI scaffold#214 has 46 gene models and scaffold#216 has 92 gene models, and both had GC% of 41.3%, which is the GC% for the COTS genome. Thus, OKI scaffold#215 may be from pathogenic bacteria. Importantly, these bacteria are known to present during the spawning season when male gonads are engorged, while the GBR sample was collected outside of the spawning season. Moreover, the presence of a pathogen in the OKI genome suggests that COTS may be amenable to bacterial or viral control approaches.

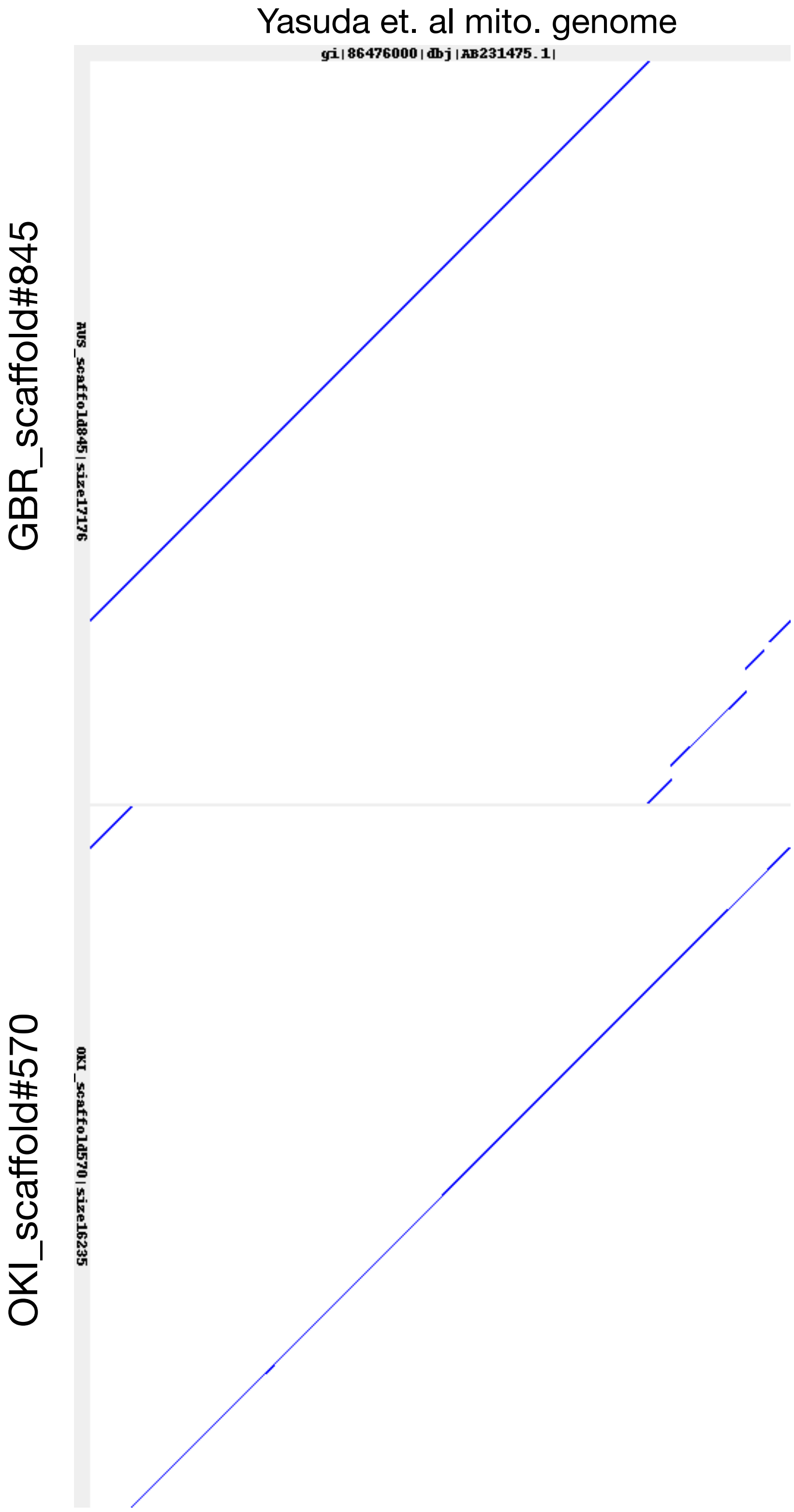


Figure 3.7. Alignments of COTS mitochondrial genomes.
LAST Alignment of GBR scaffold #845 (top) OKI scaffold#570 (bot) to the published COTS mitochondrial genome(gi|86476000)(Yasuda et al. 2006)

3.2 Analysis of the COTS Hox and ParaHox clusters

The following text is adapted and updated from: (Baughman et al. 2014)

(a) Introduction

The Hox cluster was first identified in *Drosophila melanogaster* (Lewis, 1978) and is comprised of a set of homeobox genes that encode a subfamily of homeodomain transcription factors, which are critical to the formation of bilaterian body plans (Pearson et al., 2005). Hox genes display developmental expression 'colinearity' in many animals in which the relative genomic position of a Hox gene correlates with its temporal expression and/or spatial expression along the anterior/posterior axis (Carroll, 1995). The diversification of the eumetazoan body plan has been attributed to the expansion and regulation of the Hox cluster.

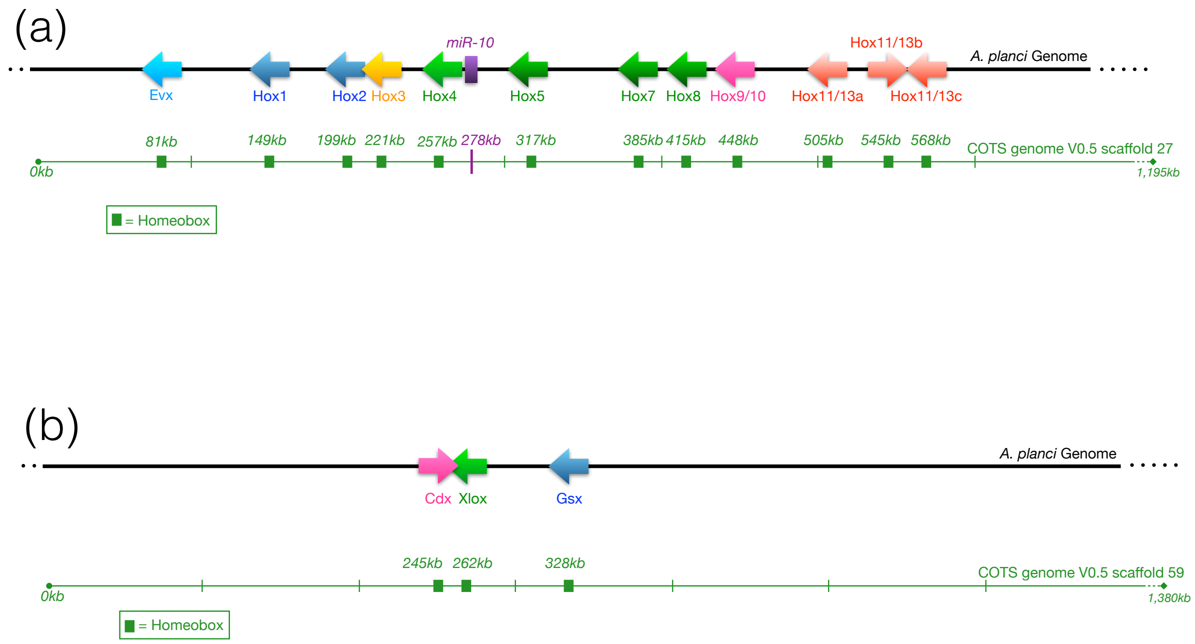


Figure 3.8. The Acanthaster planci Hox and ParaHox Clusters.
a, *Acanthaster planci* genomic scaffold #27 contains 12 regions that align with the homeobox sequence, denoted by green boxes. Phylogenetic analysis assigned these regions to specific Hox paralogy groups. Identification of *mir-10* is consistent with the proposed orientation and identity of the *A. planci* Hox cluster. Arrows denote predicted Hox genes by color; Anterior – light blue, Group 3 – yellow, Central – green, and Posterior – pink/red. **b,** A similar technique was used to identify the ParaHox cluster, which aligns with the previously published *P. miniata* ParaHox cluster (Annunziata et al., 2013), on *A. planci* genomic scaffold #59.

Echinoderms and chordates diverged from a common bilaterially-symmetrical ancestor with deuterostomous development 480 to 520 million years ago (Wada & Satoh 1994; Pisani et al. 2012; Satoh 2016). Adult echinoderms have three phyletic innovations that differ from other bilaterians: 1) pentaradial symmetry; 2) calcium carbonate endoskeletons; and 3) an ambulacral or internal water vascular system (Mooi and David, 2008). To date, the sea urchin *Strongylocentrotus purpuratus* is the only echinoderm for which genomic organization of its Hox cluster has been characterized. In this species, Hox genes 1-3 have been translocated to the 5' end of the cluster, as shown in Figure 3.11 (Martinez et al., 1999; Cameron et al., 2006). Despite this, sea urchins show aspects of Hox spatial colinearity, as posterior Hox genes participate in A/P patterning in larvae (Arenas-Mena et al., 2000). Although a full Hox cluster based on genomic data has yet to be published for sea stars, *Hox4* expression has been characterized in *Parvulastra exigua* (Byrne et al., 2005; Cisternas and Byrne, 2010) and a number of Hox genes has been characterized during starfish arm regeneration (Ben Khadra et al., 2013). Based on the available data, it has been proposed that the derived development of a pentaradially symmetrical adult was facilitated by the disruption of the Hox cluster observed in *S. purpuratus*, and that similar disruptions can be expected in the Hox clusters of other echinoderm lineages (Mooi and David, 2008).

The ParaHox cluster consists of 3 genes, *Gsx, Xlox*, and *Cdx*, and is considered the "evolutionary sister" of the Hox cluster (Brooke et al., 1998). The ParaHox cluster is involved in the development of the central nervous system and gut in bilaterians (Pearson et al., 2005; Garstang and Ferrier, 2013). An analysis of the ParaHox gene confirmed Hox-like genomic clustering in the echinoderm/asteroid ancestor, and a degree of spatial and temporal colinearity (Annunziata et al., 2013). The ParaHox cluster likely arose from a duplication of the Proto-Hox cluster, though the nature of this duplication and its relationship to the Hox cluster remain unclear (Brooke et al., 1998; Hui et al., 2011).

An understanding of *A. planci* developmental biology may offer insights into key embryonic and larval processes and reveal avenues by which this species may be manipulated to help mitigate the damage these starfish causes to the coral reef (Brodie et al., 2005; Fabricius et al., 2010). As a first step towards linking ecological data and echinoderm developmental biology, I sequenced the *A. planci* genome and identified a Hox and ParaHox cluster (Fig 3.8).

(b) Methods

Acanthaster planci genomic DNA from sperm of a mature male specimen was isolated using standard procedures (Shoguchi et al., 2013). Collaborators provided additional Australian *A. planci* DNA and RNA samples, which were processed using the same protocols outlined in chapter 2: Methods, regarding the assembly of the GBR V0.5 genome. Briefly, genomic paired-end, mate-pair, and cDNA (mRNA) libraries were prepared by standard protocols (Shoguchi et al., 2013), and sequenced on Illumina Miseq and HiSeq instruments, respectively. Initial genomic assembly was done with GS De Novo Assembler version 2.3 (Newbler, Roche), and scaffolding was done with SSPACE-BASIC-2.0. RNAseq raw reads were assembled *de novo* using Trinity (Haas et al., 2013). Raw paired end reads were mapped back to GBR V0.5 Scaffold #27 and *A. planci* Scaffold #59, in order to confirm read coverage (Figure 3.11). The Tuxedo pipeline was used to generate a separate set of RNA transcripts (Trapnell et al., 2012). NCBI blast+ was used to identify *A. planci* scaffolds containing Hox genes (Camacho et al., 2009). LAST was used to compare and visualize local synteny (Kielbasa et al., 2011). Molecular phylogenetics analysis was performed using an alignment of 56 amino acids of the homeodomain (Carroll, 1995; Gyoja, 2014). FGENESH was used for *ab initio* gene prediction on GBR V0.5 Scaffold #27 and A. planci Scaffold #59

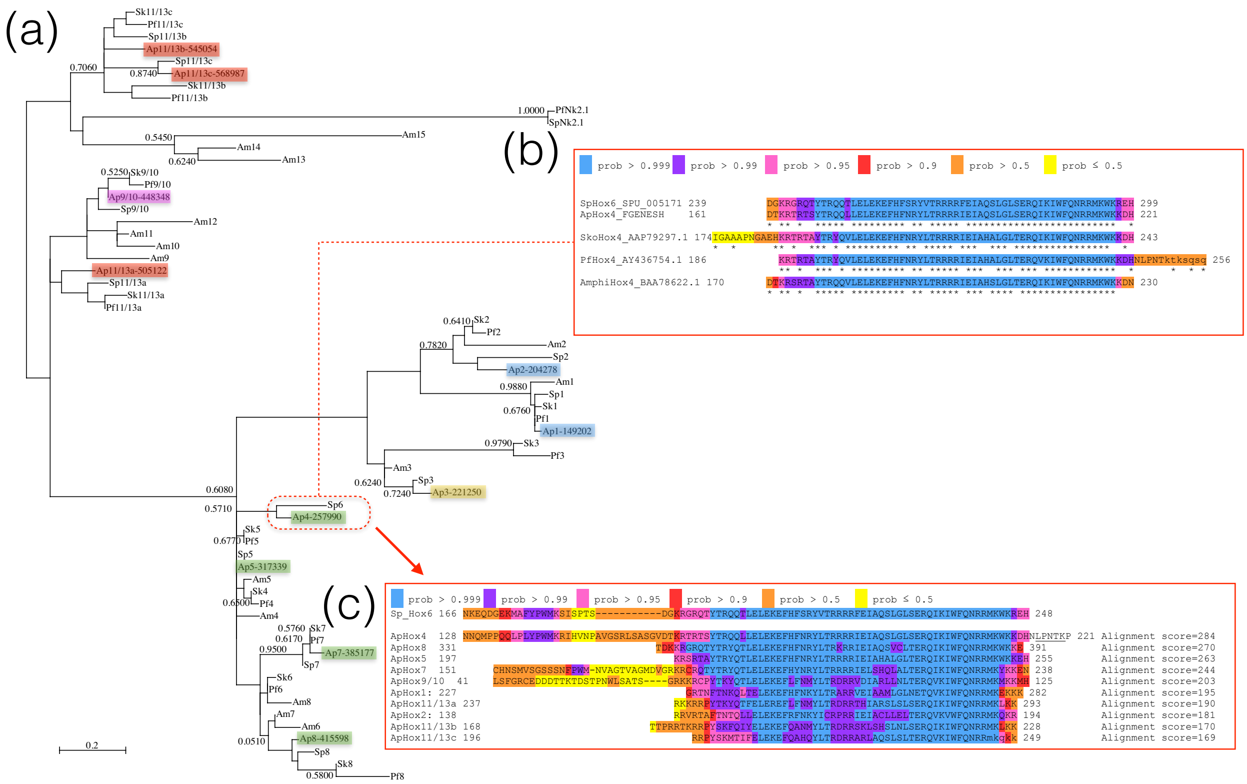
(Solovyev et al., 2006). Scaffold sequences have been deposited with DDBJ/EMBL/GenBank as accession numbers DF933567 (A_planci_scaf27_V0.5) and DF933568 (A_planci_scaf59_V0.5). The 39 contigs for Scaffold #27 (A_planci_scaf27_V0.5_contig1 to A_planci_scaf27_V0.5_contig39) have accession numbers (BBNW01000001 to BBNW01000039), and the 44 contigs for Scaffold #59 (A_planci_scaf59_V0.5_contig1 to A_planci_scaf59_V0.5_contig44) have accession numbers (BBNW01000040 to BBNW01000083).

(c) Results and Discussion

We identified the Hox and ParaHox clusters by comparison of the *A. planci* genome with Hox sequences from the starfish *Patria miniata,* two hemichordates, *Saccoglossus kowalevskii* and *Ptychodera flava,* the amphioxus *Branchiostoma floridae*, and the sea urchin *Strongylocentrotus purpuratus.* We found a single *A. planci* genomic scaffold (Scaffold #27: 149 - 568 kb) that contains a cluster of 12 homeobox genes within a 420kb region (Figure 3.8a). Six of these homeobox-containing genes are expressed in adult tissues. The highly conserved Hox cluster-associated microRNA, *mir-10*, is also present in *A. planci* (Scaffold #27: 278 kb) (Figure 3.8a).

In contrast to the Hox cluster of *S. purpuratus*, the *A. planci* Hox gene order is conserved with both chordate and hemichordate Hox clusters, and thus likely represents the order present in the last common ancestor to extant deuterostomes. The orientation of *A. planci Hox11/13b* is inverted with respect to the rest of the cluster, as found in *S. purpuratus,* and the proximity of *Evx* to *Hox1* (versus *Hox14* in chordates) implies conservation within the echinoderm clade. *Hox4* is present in *A. planci*, and consistent with previous results from sea stars, contains the 'LPNTK' motif found 3' to the homeodomain (Byrne et al., 2005; Cisternas and Byrne, 2010). In contrast, *Hox6* is absent from the *A. planci* Hox cluster. The loss of *Hox6* in *A. planci* is supported by the lack of a homeobox sequence between *Hox5* and *Hox7* (Figure 3.9a), and the lack of synteny between *A. planci Hox5* and *Hox7* and any regions of *S. purpuratus* (Figure 3.8).

To predict the orthologous relationships among the Ambulacraria Hox genes, molecular phylogenetic analysis of the homeodomains from *A. planci*, *S. purpuratus*, *P. flava*, *S. kowalvskii* and *B. floridae* was performed (Figure 3.9). Note that full length *S.purpuratus Hox6* is most closely phylogenetically linked to the *A. planci Hox4* of all other *A. planci* Hox genes (Figure 3.8c), suggesting that *A. planci Hox4* and *S. purpuratus Hox6* may have the same ancestral origin.



```
SpHox6_SPU_005171 239        DGKRGRQTYTRQQTLELEKEFHFSRYVTRRRRFEIAQSLGLSERQIKIWFQNRRMKWKREH 299
ApHox4_FGENESH    161        DTKRTRTSYTRQQLLELEKEFHFNRYLTRRRRIEIAQSLGLSERQIKIWFQNRRMKWKKDH 221
                             * ** *  ***** ********* ** ***** ************************  *

SkoHox4_AAP79297.1 174IGAAAPNGAEHKRTRTAYTRYQVLELEKEFHFNRYLTRRRRIEIAHALGLTERQIKIWFQNRRMKWKKDH 243
                      *  *      ** *  *** * ********* ** ***** ***  *** ***************  *

PfHox4_AY436754.1 186          KRTRTAYTRYQVLELEKEFHFNRYLTRRRRIEIAHALGLTERQVKIWFQNRRMKWKKDHNLPNTktksqsq 256
                               ** *  *** * ********* ** ***** ***  *** *** ***********  *       *  * *

AmphiHox4_BAA78622.1 170     DTKRSRTAYTRQQVLELEKEFHFNRYLTRRRRIEIAHSLGLTERQIKIWFQNRRMKWKKDN 230
                             * ** *  ***** ********* ** ***** ***  **** ***************
```

```
Sp_Hox6 166 NKEQDGEKMAFYPWMKSISPTS-----------DGKRGRQTYTRQQTLELEKEFHFSRYVTRRRRFEIAQSLGLSERQIKIWFQNRRMKWKREH 248

ApHox4   128 NNQMPPQQLPLYPWMKRIHVNPAVGSRLSASGVDTKRTRTSYTRQQLLELEKEFHFNRYLTRRRRIEIAQSLGLSERQIKIWFQNRRMKWKKDHNLPNTKP 221 Alignment score=284
ApHox8   331                                TDKKRGRQTYTRYQTLELEKEFHFNRYLTRKRRIEIAQSVCLTERQIKIWFQNRRMKWKKE 391       Alignment score=270
ApHox5   197                                   KRSRTAYTRYQTLELEKEFHFNRYLTRRRRIEIAHALGLTERQIKIWFQNRRMKWKKEH 255       Alignment score=263
ApHox7   151      CHNSMVSGSSSNFPWM-NVAGTVAGMDVGRKRCRQTYTRYQTLELEKEFHYNRYLTRRRRIELSHQLALTERQIKIWFQNRRMKYKKEN 238       Alignment score=244
ApHox9/10  41     LSFGRCEDDDTTKTDSTPNWLSATS----GRKKRCPYTKYQTLELEKEFLFNMYLTRDRRVDIARLLNLTERQVKIWFQNRRMKMKKMH 125       Alignment score=203
ApHox1: 227                                      GRTNFTNKQLTELEKEFHFNKYLTRARRVEIAAMLGLNETQVKIWFQNRRMKEKKK 282       Alignment score=195
ApHox11/13a 237                                RKKRRPYTKYQTFELEREFLFNMYLTRDRRTHIARSLSLSERQIKIWFQNRRMKLKK 293       Alignment score=190
ApHox2: 138                                    RRVRTAFTNTQLLELEKEFRFNKYICRPRRIEIACLLELTERQVKVWFQNRRMKQKR 194       Alignment score=181
ApHox11/13b 168                            TTPRRTKRRPYSKFQIYELEKEFQANMYLTRDRRSKLSHSLNLSERQIKIWFQNRRMKLKK 228       Alignment score=170
ApHox11/13c 196                                   RRPYSKMTIFELEKEFQAHQYLTRDRRARLAQSLSLTERQVKIWFQNRRmkqkk 249       Alignment score=169
```

Figure 3.9. COTS Hox gene Phylogenetic Analysis.
a, Molecular phylogenetic analysis of the echinoderm Hox genes by maximum-likelihood method. The six digit number following the *A. planci* (Ap) proteins corresponds to the gene location on scaffold #27. Molecular phylogenetic analysis is based on comparison of 56 amino acid positions from the homeodomains of Hox genes from *Acanthaster planci (Ap), Saccoglossus kowalevskii (Sk), Ptychodera flava (Pf), Branchiostoma floridae (Bf)*, and *Stronglocentrotus purpuratus (Sp)*. Bootstrap values of more than 0.5 are shown. The bar shows branch length for a 0.2 amino acid substitution. **b,** Local homeodomain sequence alignment using LAST of *Hox6* from *S. purpuratus*, against *Hox4* from *A. planci, S. kowalevskii, P. flava,* and *B. floridae*. Asterisks denote the conserved sites from *S. purpuratus Hox6* for each *Hox4*, respectively. The colors denote alignment probability. **c,** Full length LAST alignments of all *A. planci* Hox genes to *Hox6* from *S. purpuratus.* 'LPNTK' motif confirming identity of Ap*Hox4* is underlined, which is absent in *SpHox6*. Note that *ApHox4* has the highest alignment score of 284.

In order to confirm the orientation and organization of the *A. planci* Hox cluster, a local synteny analysis of *A. planci* genomic scaffold #27 and *S. purpuratus* genomic scaffold #636 was performed (Figure 3.10). *A. planci Hox11/13b* is the only Hox gene consistently inverted with regard to every other *A. planci* Hox gene. The conservation of the orientation of the posterior Hox genes in *A. planci* and *S. purpuratus* strengthens the model in which these posterior Hox genes may be involved in the divergence of adult echinoderm body plans (Cameron et al., 2006; Freeman et al., 2012). *Hox11/13b* is expressed in embryos of *S.*

purpuratus, which is likely when divergence from the bilaterian body begins (Martinez et al., 1999; Cameron et al., 2006).

Additionally, in Figure 3.10 there are two regions of synteny outside the Hox clusters. Although neither of these two sequences mapped to coding regions in *S. purpuratus*, the second sequence at 1025 kb of *A. planci* scaffold #27, aligned with *Asterina pectinifera* inositol 1,4,5-trisphosphate receptor mRNA (ApIP3R, GenBank accession #: AB071372.1), which was also aligned to two related *A. planci* RNAseq transcripts (Figure 3.10). The first region of synteny outside the *A. planci* Hox cluster, at 778 kb of *A. planci* scaffold #27, did not align to any RNAseq transcripts.

As a whole, the *A. planci* Hox cluster resembles that of two hemichordates (*S. kowalevski*, and *P. flava*), and a cephalochordate (*B. floridae*); displaying the ancestral arrangement of the anterior, medial and posterior Hox genes (Figure 3.10). This result, while somewhat unexpected, supports the notion that the Hox cluster has been evolutionarily conserved amongst all deuterostome groups. *Hox4* is conserved in Asteroidea (Byrne et al., 2005; Cisternas and Byrne, 2010) in contrast to the loss of *Hox4* in *S. purpuratus* (Cameron et al., 2006).

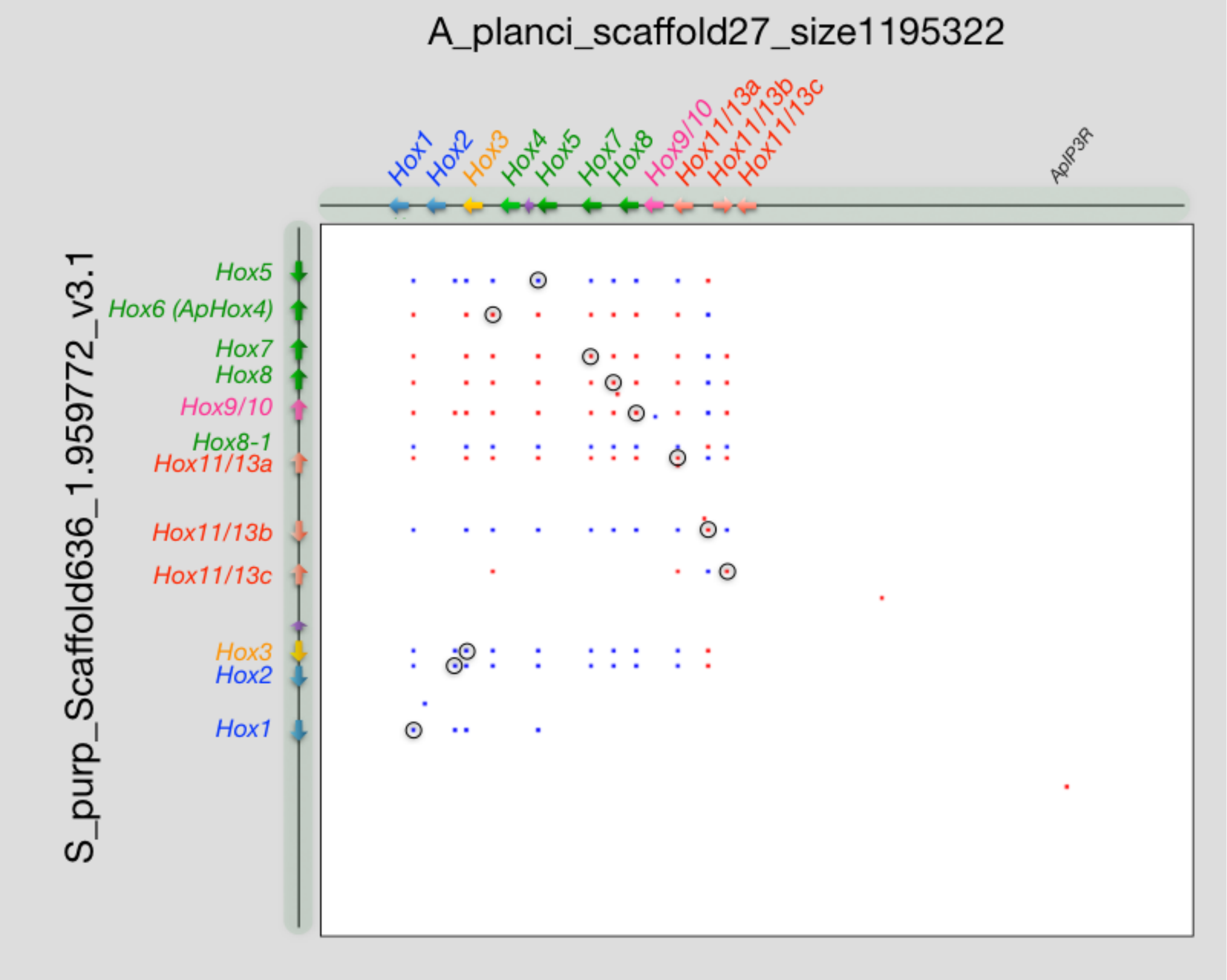


Figure 3.10. COTS and sea unchin Hox clusters, based on Scaffold synteny.
The *A. planci* scaffold #27 runs along the horizontal axis, and the *S. purpuratus* scaffold #636 runs from the upper along the vertical axis. Areas of synteny are noted in blue (same orientation) or red (reverse orientation). The regions of synteny within the Hox cluster are mostly restricted to the homeobox, though a small region adjacent to *Hox8* outside of the homeodomains was also syntenic. For a given Hox gene on the *A. planci* scaffold, all other Hox genes are blue except for the loci associated with *Hox11/13b,* which is red, or vice versa. Orthologous predicted genes, supported by molecular phylogenetic analysis, are circled. Several regions of synteny outside of the cluster are noted, including the inositol 1,4,5-trisphosphate receptor.

Our results support a two-phase model for echinoderm Hox cluster evolution in which the transition from a hypothetical ambulacrarian ancestor to sea urchin may have proceeded through at least four steps (Figure 3.11). In the first phase, the ancestral echinoderm Hox cluster evolved in two steps; inversion of *Hox11/13b,* and loss of *Hox6.* The sequence of these steps cannot be determined from data presented here. Previous models assumed that the asteroidea Hox cluster would mirror the disorganization of the *S. purpuratus* Hox cluster, and proposed that the anterior Hox cluster translocated prior to the inversion of *Hox11/13b* and preceded the loss of *Hox4* (Cameron et al., 2006; Freeman et al., 2012). We note the similarity of *S. purpuratus Hox6* to *A. planci Hox4*, based on phylogenetic analysis (Figure 3.9b). Thus, if *S. purpuratus Hox6* is reclassified as *Hox4*, we can propose a simplified 2-step process for the second phase of our model. First, a segment containing *Hox4* and *mir-10* is inverted locally (step 4 in Figure 3.11). This is followed by the translocation and inversion of a large region containing *Hox7* through *Hox11/13c* into the inverted segment, between *Hox4* and *mir-10* (step 5 in Figure 3.11).

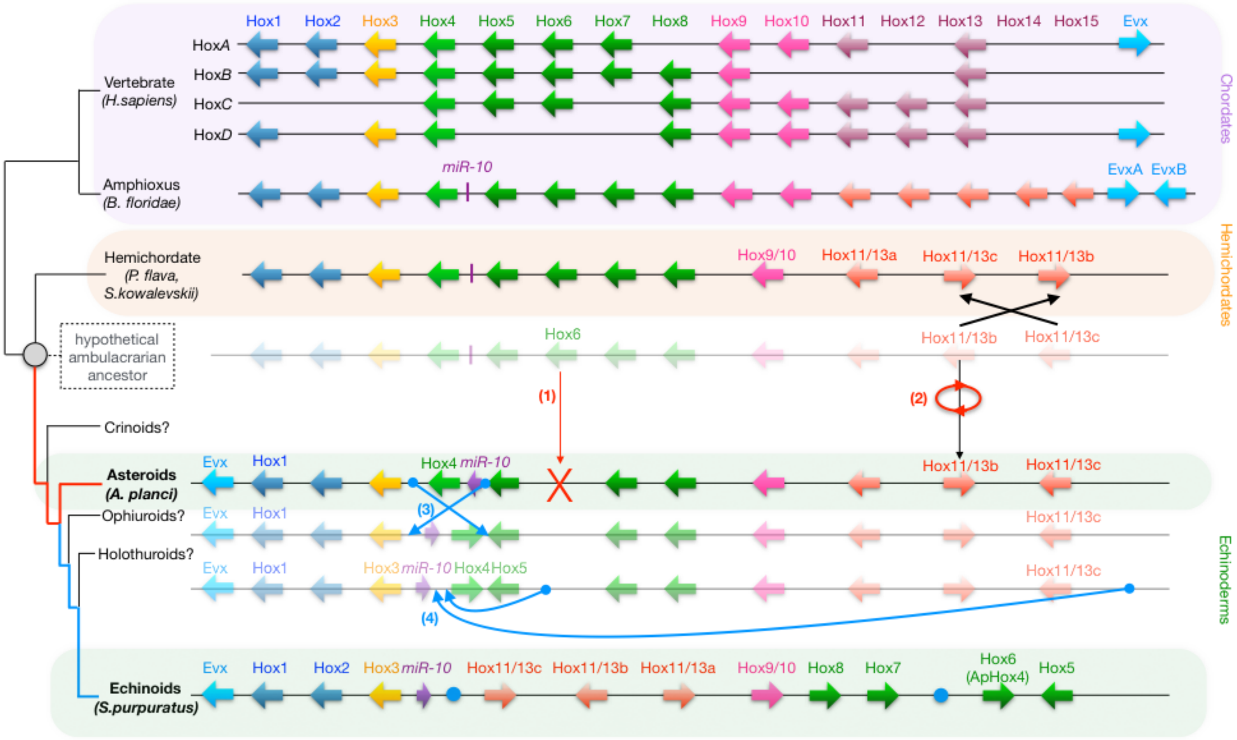


Figure 3.11. A model for Echinoderm Hox Cluster Evolution.
The evolution of the echinoderm Hox cluster from a hypothetical ambulacrarian ancestor to sea urchin proceeds through two phases in five steps. Phase one, in red: (1) Deletion of *Hox6.* (2) Inversion of *Hox11/13b*. Phase two, in blue: (3) Inversion of a segment containing *Hox6* (*ApHox4*) and *mir-10*. (4) Translocation and inversion of a segment containing *Hox7* through *Hox11/13c* to between *Hox4* and *mir-10*.

In this scenario, an explanation for both the loss of *Hox4* and recovery of *Hox6* in *S. purpuratus* is no longer required, and the proximity of *mir-10* to *Hox3,* which is unique to sea urchin, is explained. Additionally, *mir-10* is inverted in *A. planci*, relative to *S. purpuratus*. The medial Hox genes are difficult to classify, due to a high degree of similarity of the homeodomain. Is it possible that the asteroid *Hox4* is actually *Hox6,* and thus *Hox6* is not missing in *A. planci*? Based on the location of *Hox4* and *Hox6* within the collinear *A. planci* Hox cluster (Figure 3.8a), and the presence of the 'LPNTK' motif in the *Hox4*, this is highly unlikely (Figure 3.9c). Thus, our model for echinoderm Hox cluster evolution proposes that the major rearrangement mechanism might be local inversion.

Lastly, the alignment and orientation of the *A. planci* ParaHox cluster to *P. miniata* (Figure 3.12) (Annunziata et al., 2013) generated blast+ alignments scores of the same magnitude as the alignments for the Hox cluster, confirming our Hox gene cluster methodology. In short, we found a genomic scaffold that aligns to the three *P. miniata* ParaHox genes, and these regions each contained complete homeobox sequences (Scaffold #59: 245 - 328 kb) (Figure 3.9b).

Despite having a chordate-like Hox organization, sea stars adults have pentaradial symmetry, possess a unique water vascular system and a calcium carbonate endoskeleton. Thus, it seems unlikely that these departures from the body plan of the hypothetical ambulacrarian ancestor are due to the disorganization of the anterior and medial Hox cluster. Our results support the notion that *A. planci* research based on understanding the ecological devastation *A. planci* causes to coral reefs are useful for exploring questions in development and evolution.

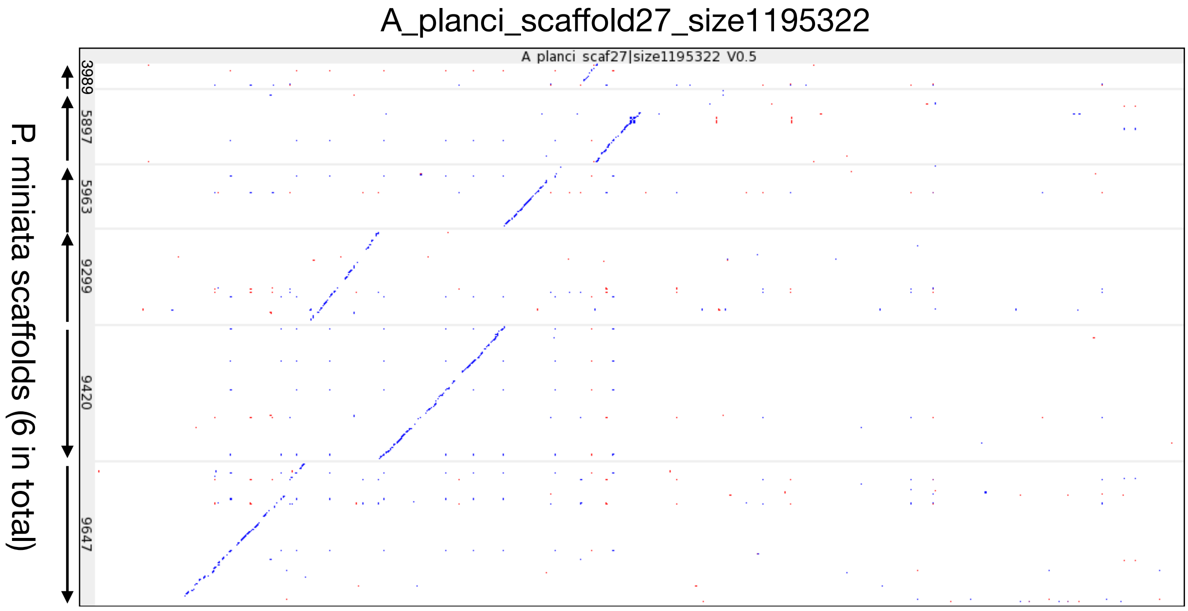


Figure 3.12. Alignment of COTS Hox scaffold to bat star genome.
LASTAL alignment of COTS Scaffold #27 (top) to *P. miniata* Scaffolds (left side, 5 in total). Areas of synteny are noted in blue (same orientation) or red (reverse orientation).

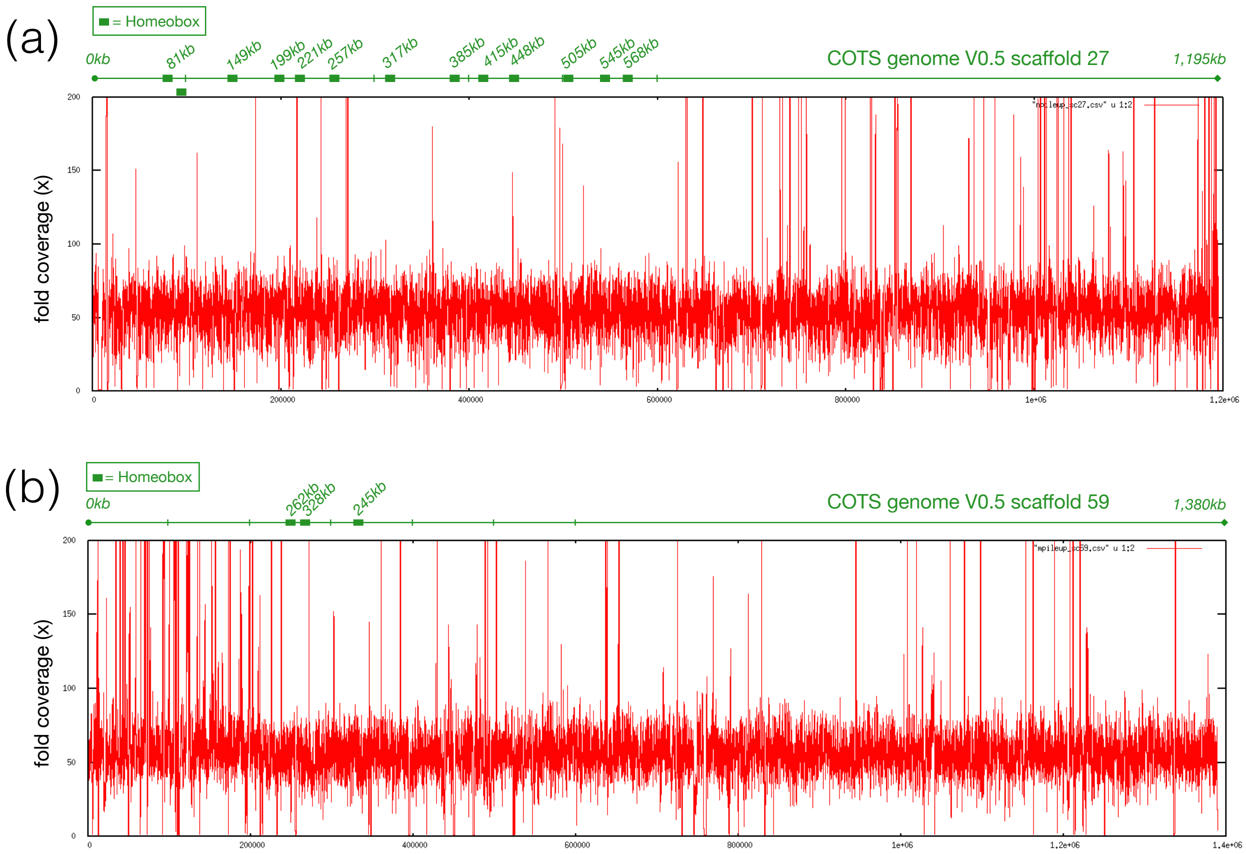


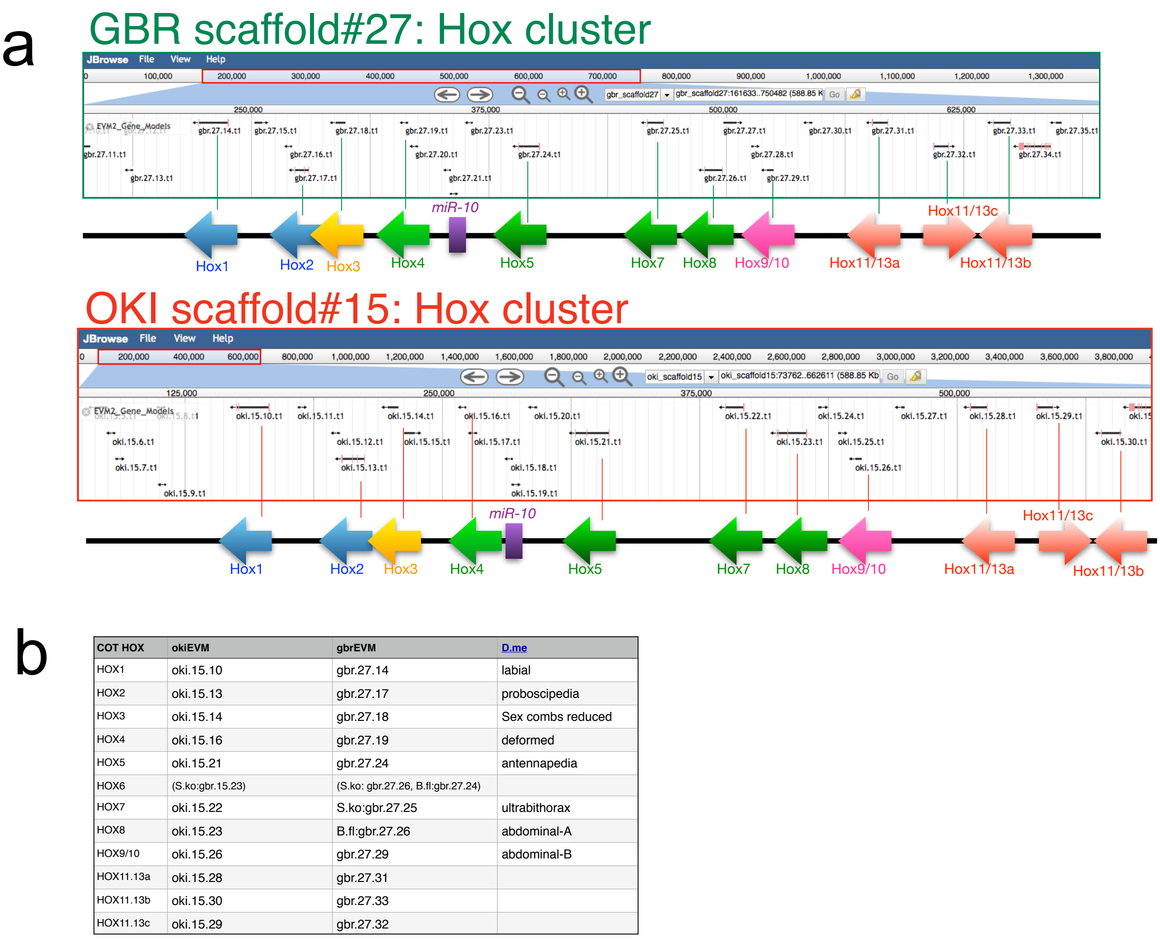
Figure 3.13. Paired-end (PE) read coverage.

a, Scaffold #27 and **b,** Scaffold #59. The high read coverage (~46x) across both scaffolds confirms that inappropriate scaffold joining is unlikely, and that the collinearity of the Hox cluster and parahox cluster are biologically relevant.

(d) Further analysis of Hox and ParaHox clusters, based on final genome assemblies. Figures adapted from (Hall et al. 2017)

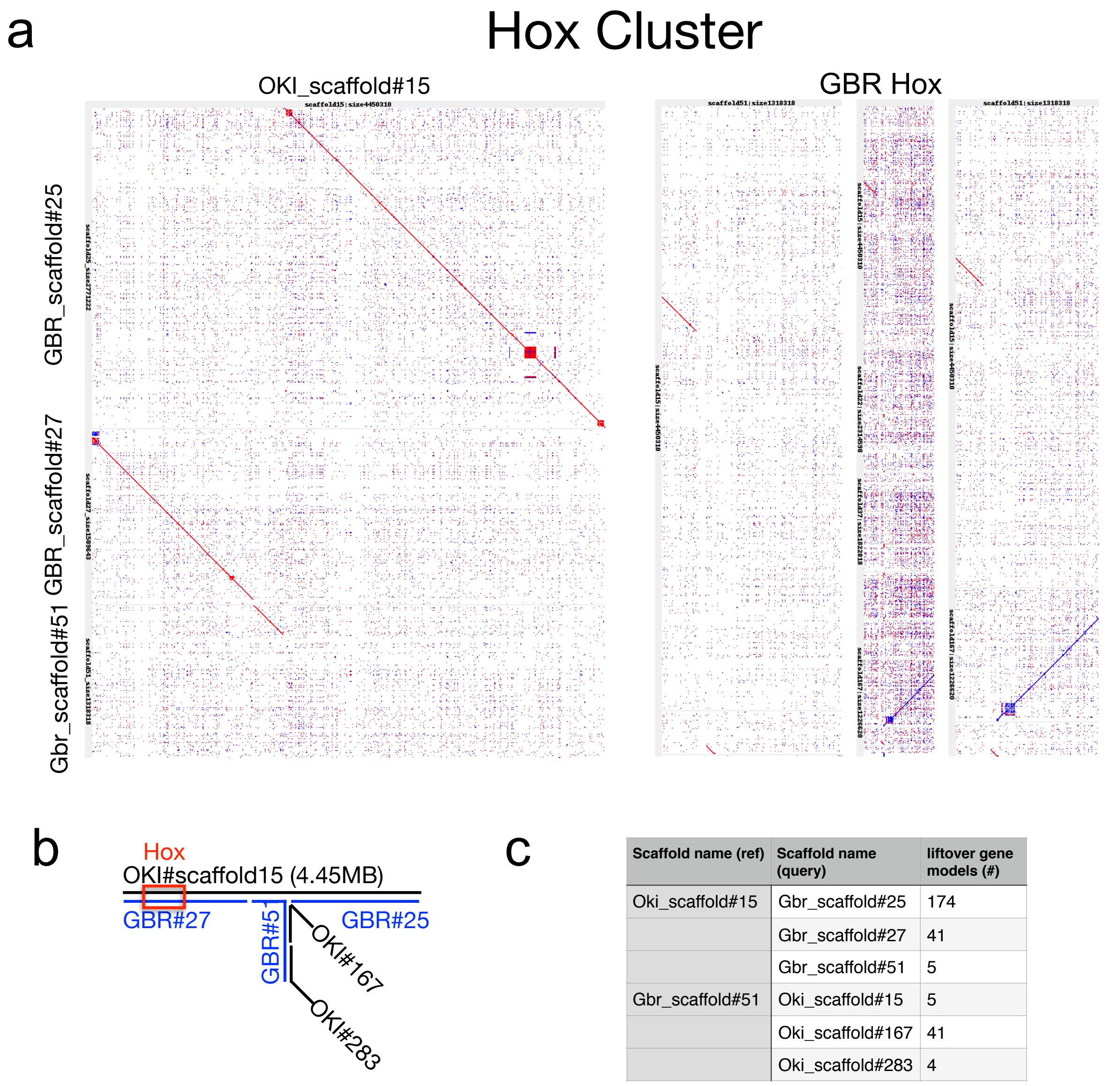
The initial discovery of the Hox and ParaHox clusters in COTS were made based on an earlier scaffolding of the GBR genome (GBR V0.5). Importantly, gene modeling was done only for these two Hox and ParaHox containing scaffolds (#27 and #59). Finally, RNA transcripts were not available at the time of publication.

Hox clusters were identified in both GBR and OKI V1.0 assemblies, and gene models for all Hox genes were present (Figure 3.14). By aligning the OKI and GBR Hox containing scaffolds to each other, it was possible to identify additional OKI and GBR scaffolds (Figure 3.15a), and create a map of scaffold joins (Figure 3.15b), each of which may represent either a biologically significant polymorphism, or an assembly error. Gene liftover analyses confirms corresponding scaffolds, but cannot confirm polymorphism versus assembly error. Presumable, Polymerase Chain Reaction (PCR) products generated across the regions in question would confirm polymorphism. Phylogenetic analysis of the 56 amino acid homeobox region from all OKI Hox gene models confirms the previously published Hox gene identities (Figure 3.16).



COT HOX	okiEVM	gbrEVM	D.me
HOX1	oki.15.10	gbr.27.14	labial
HOX2	oki.15.13	gbr.27.17	proboscipedia
HOX3	oki.15.14	gbr.27.18	Sex combs reduced
HOX4	oki.15.16	gbr.27.19	deformed
HOX5	oki.15.21	gbr.27.24	antennapedia
HOX6	(S.ko:gbr.15.23)	(S.ko: gbr.27.26, B.fl:gbr.27.24)	
HOX7	oki.15.22	S.ko:gbr.27.25	ultrabithorax
HOX8	oki.15.23	B.fl:gbr.27.26	abdominal-A
HOX9/10	oki.15.26	gbr.27.29	abdominal-B
HOX11.13a	oki.15.28	gbr.27.31	
HOX11.13b	oki.15.30	gbr.27.33	
HOX11.13c	oki.15.29	gbr.27.32	

Figure 3.14. Confirmation of Hox clusters in OKI and GBR V1.0 Genomes.
a, Screen shots from the COTS genome browser (http://marinegenomics.oist.jp/cots/viewer/info?project_id=46), with gene models for all Hox genes highlighted in OKI and GBR V1.0 genome assemblies. **b,** OKI and GBR Hox gene model Hox gene correspondence.



Scaffold name (ref)	Scaffold name (query)	liftover gene models (#)
Oki_scaffold#15	Gbr_scaffold#25	174
	Gbr_scaffold#27	41
	Gbr_scaffold#51	5
Gbr_scaffold#51	Oki_scaffold#15	5
	Oki_scaffold#167	41
	Oki_scaffold#283	4

Figure 3.15. Alignment of OKI and GBR Hox containing scaffolds.

a, LAST alignment of OKI scaffold#15 against GBR scaffolds#25, #27 (Hox containing), and #51. **b,** GBR Scaffold alignments. GBR#25 and #27 only align to OKI#15, while sections of GBR#51 align to OKI#167 and #283. **c,** Gene lift over counts, confirming the orientation of the alignments.

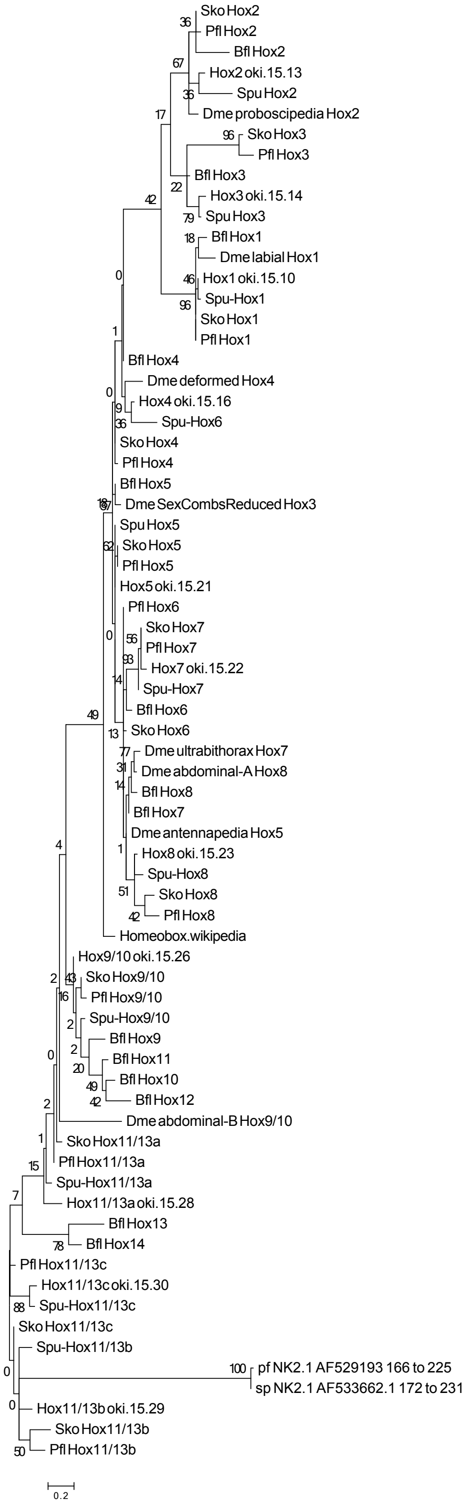


Figure 3.16. Phylogentic Analysis of OKI V1.0 Hox Gene Models.
Molecular phylogenetic analysis is based on comparison of 56 amino acid positions from the homeodomains of Hox genes from *Acanthaster planci (Ap), Saccoglossus kowalevskii (Sk), Ptychodera flava (Pf), Branchiostoma floridae (Bf)*, and *Stronglocentrotus purpuratus (Sp)*. The bar shows branch length for a 0.2 amino acid substitution**.** These data confirm that the V1.0 gene models match the Hox Genes predicted in (Baughman et al. 2014).

3.3 Analysis of the COTS Nkx pharyngeal-gill-slit-related gene cluster

The following text is adapted from: (Simakov et al. 2015)

(a) Introduction

Among the various deuterostome-defining synapomorphies, which notably include radial cleavage, development of the anus from the blastopore, and triploblastic composition of adult tissue, pharyngeal gill slits have been proposed as a clade-defining feature of early, filter-feeding deuterostome ancestors (Satoh 2016). While hemichordates and basal chordates such as Amphioxus maintain functioning pharyngeal slits currently, all chordates have pharyngeal-slit-like features which often appear transiently during development. Although some echinoderm fossils appear to have pharyngeal-gill-slit-like structures, extent echinoderms are not known to have pharyngeal gill slits nor pharyngeal-like tissues, at any point during development (Satoh 2016). A recently published comparison of two Hemichordate genomes identified a cluster of genes expressed in the pharyngeal slits and surrounding pharyngeal endoderm. The cluster is conserved in several deuterostome genomes (Simakov et al. 2015), which, surprisingly, I found intact in the COTS genome (Figure 3.17).

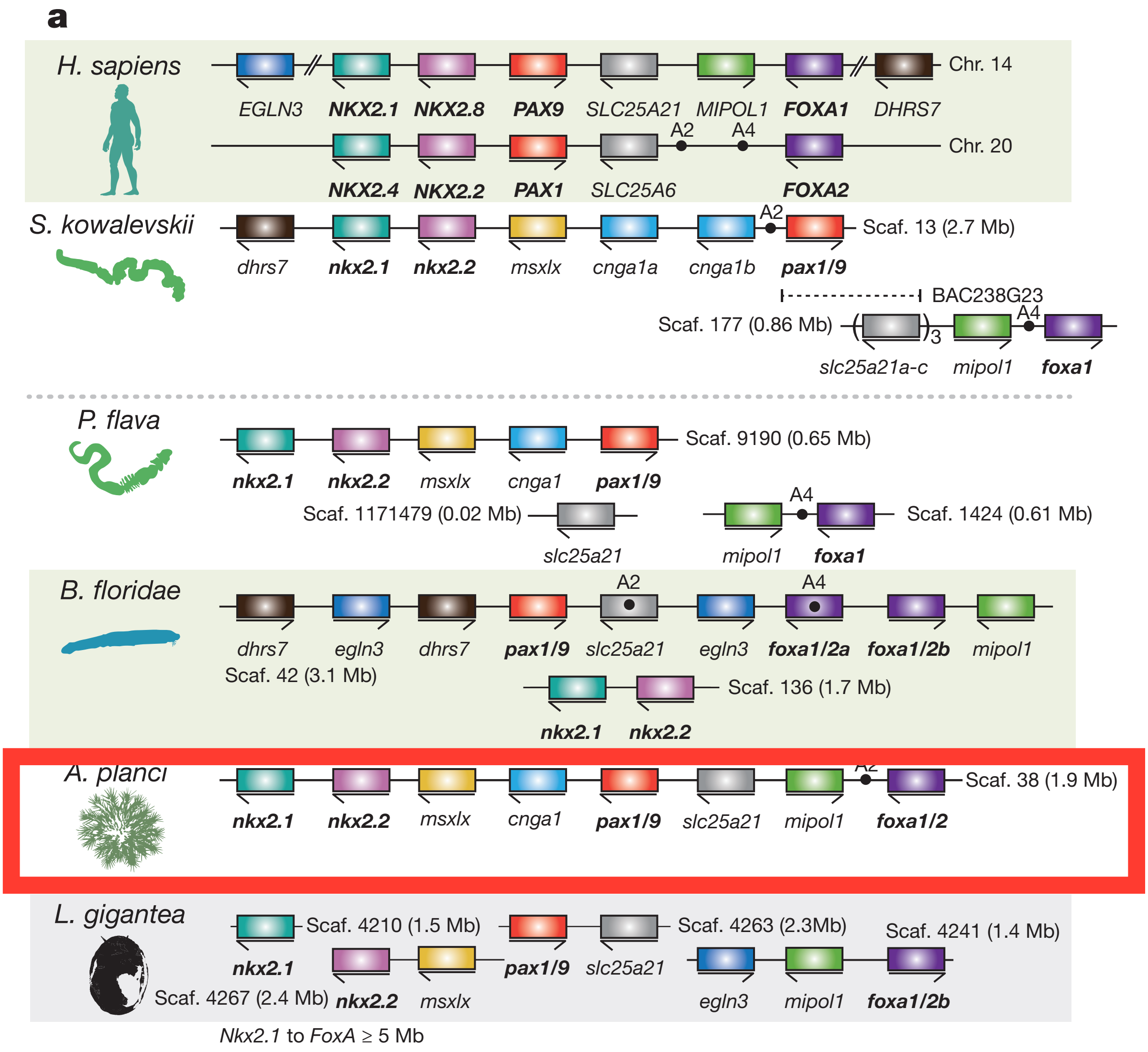


Figure 3.17. Conservation of a pharyngeal gene cluster across deuterostomes.
Linkage and order of six genes including the four genes encoding transcription factors Nkx2.1, Nkx2.2, Pax1/9 and FoxA, and two genes encoding non-transcription factors Slc25A21 (solute transporter) and Mipol1 (mirror-image polydactyly 1 protein), which are putative 'bystander' genes containing regulatory elements of *pax1/9* and *foxA*, respectively. The pairings of *slc25A21* with *pax1/9* and of *mipol1* with *foxA* occur also in protostomes, indicating bilaterian ancestry. The cluster is not present in protostomes such as *Lottia* (Lophotrochozoa), *Drosophila melanogaster*, *Caenorhabditis elegans* (Ecdysozoa), or in the cnidarian, *Nematostella*. *SLC25A6* (the *slc25A21* paralogue on human chromosome 20) is a potential pseudogene. The dots marking A2 and A4 indicate two conserved non-coding sequences first recognized in vertebrates and amphioxus, also present in *S. kowalevskii* and, partially, in *P. flava* and *A. planci* or COTS (highlighted in the red box). Adapted from(Simakov et al. 2015)

(b) Methods

The OKI V1.0 genome assembly used for all analyses, was assembled as described in chapter 2: Methods.

(c) Results and discussion

As identified by the hemichordate genome study (Simakov et al. 2015), the pharyngeal gene cluster contains four transcription factor genes in the order *nkx2.1*, *nkx2.2*, *pax1/9* and *foxA*, along with two non-transcription-factor genes *slc25A21* and *mipol1*, whose introns harbor regulatory elements for *pax1/9* and *foxA*, respectively (Santagati et al. 2003; Lowe et al. 2006; W. Wang et al. 2006). The cluster was first found conserved across vertebrates including humans (chromosome 14; 1.1 Mb length from *nkx2.1* to *foxA1*) (Santagati et al. 2003). In *S. kowalevskii*, it is intact with the same gene order as in vertebrates (Figure 3.17)(0.5 Mb length from *nkx2.1* to *foxA*), implying that it was present in the deuterostome and ambulacrarian ancestors. The fully ordered gene cluster also exists on a single scaffold in the crown-of-thorns sea star. Since these genes are not clustered in available protostome genomes, there is no evidence for deeper bilaterian ancestry. Two non-coding elements that are conserved across vertebrates and amphioxus (S. Wang et al. 2009) are found in the hemichordate and *A. planci* clusters at similar locations (Figure 3.17).

The hemichordate study found that on a more local scale, hundreds of tightly linked conserved gene clusters of three or more genes (‘micro-synteny’) including Hox (Freeman et al. 2012) and ParaHox (Ikuta et al. 2013) clusters in both acorn worms, as also found in echinoderms (Cameron et al. 2006; Baughman et al. 2014). Conservation of micro-syntenic linkages can occur due to low rates of genomic rearrangement or, more interestingly, as a result of selection to retain linkages between genes and their regulatory elements located in neighboring genes (Irimia et al. 2012).

The hemichordate study also found that *pax1/9* gene, at the center of the cluster, is expressed in the pharyngeal endodermal primordium of the gill slit in hemichordates, tunicates, amphioxus, fish, and amphibians (Ogasawara et al. 1999; Gillis et al. 2011), and in the branchial pouch endoderm of amniotes (which do not complete the last steps of gill slit formation), as well as other locations in vertebrates. The *nkx2.1* (thyroid transcription factor 1) gene is also expressed in the hemichordate pharyngeal endoderm in a band passing through the gill slit, but not localized to a thyroid-like organ (Lowe et al. 2003).

The presence of this cluster in COTS, an echinoderm that lacks gill pores, and in amniote vertebrates that lack gill slits, suggests that the cluster's ancestral role was in pharyngeal apparatus patterning as a whole, of which overt slits (perforations of apposed endoderm and ectoderm) were but one part, and the cluster is retained in these cases because of its continuing contribution to pharynx development. Genomic regions of the pharyngeal cluster have been implicated in long-range promoter–enhancer interactions, supporting the regulatory importance of this gene linkage (Kokubu et al. 2009). Alternatively, genome rearrangement in these lineages may be too slow to disrupt the cluster even without functional constraint. The clustering of the four ordered transcription factors, and their bystander genes, on the deuterostome stem may have served a regulatory role in the evolution of the pharyngeal apparatus, the foremost morphological innovation of deuterostomes.

(e) Further Analysis of Nkx Cluster, based on final genome assemblies.

Following the publication of the hemichordate genome study, Nk clusters genes were identified in both GBR and OKI V1.0 assemblies, and gene models for all Nk cluster genes were present (Figure 3.18). Phylogenetic analysis Nk gene models confirms the published Nk gene model identities (Figure 3.19). RNAseq expression data for the four Nk cluster genes confirms expression during early COTS development, consistent with Hemichordate expression data (Figure 3.20).

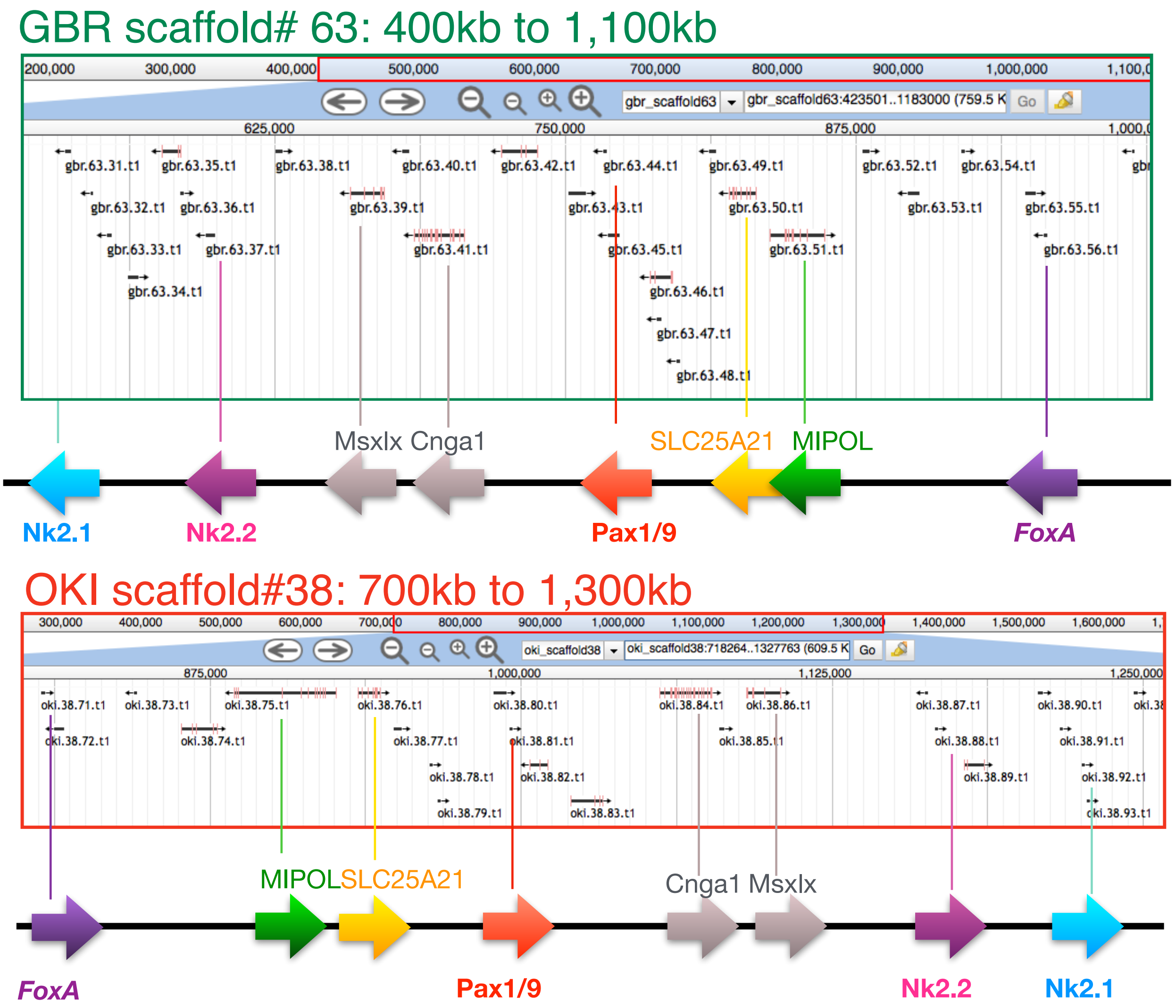

GBR scaffold# 63: 400kb to 1,100kb
200,000 300,000 400,000 500,000 600,000 700,000 800,000 900,000 1,000,000 1,100,0
gbr_scaffold63
gbr_scaffold63:423501..1183000 (759.5 K
Go
625,000 750,000 875,000 1,000,0
gbr.63.31.t1 gbr.63.35.t1 gbr.63.38.t1 gbr.63.40.t1 gbr.63.42.t1 gbr.63.44.t1 gbr.63.49.t1 gbr.63.52.t1 gbr.63.54.t1 gbr
gbr.63.32.t1 gbr.63.36.t1 gbr.63.39.t1 gbr.63.43.t1 gbr.63.50.t1 gbr.63.53.t1 gbr.63.55.t1
gbr.63.33.t1 gbr.63.37.t1 gbr.63.41.t1 gbr.63.45.t1 gbr.63.51.t1 gbr.63.56.t1
gbr.63.34.t1 gbr.63.46.t1
gbr.63.47.t1
gbr.63.48.t1
Msxlx Cnga1
SLC25A21 MIPOL
Nk2.1
Nk2.2
Pax1/9
FoxA
OKI scaffold#38: 700kb to 1,300kb
300,000 400,000 500,000 600,000 700,000 800,000 900,000 1,000,000 1,100,000 1,200,000 1,300,000 1,400,000 1,500,000 1,600,000 1,
oki_scaffold38
oki_scaffold38:718264..1327763 (609.5 K
Go
875,000 1,000,000 1,125,000 1,250,000
oki.38.71.t1 oki.38.73.t1 oki.38.75.t1 oki.38.76.t1 oki.38.80.t1 oki.38.84.t1 oki.38.86.t1 oki.38.87.t1 oki.38.90.t1 oki.38
oki.38.72.t1 oki.38.74.t1 oki.38.77.t1 oki.38.81.t1 oki.38.85.t1 oki.38.88.t1 oki.38.91.t1
oki.38.78.t1 oki.38.82.t1 oki.38.89.t1 oki.38.92.t1
oki.38.79.t1 oki.38.83.t1 oki.38.93.t1
MIPOLSLC25A21
Cnga1 Msxlx
FoxA
Pax1/9
Nk2.2
Nk2.1

Figure 3.18. Nkx Pharyngeal gene clusters in OKI and GBR V1.0 Genomes.
Screen shots from the COTS genome browser (http://marinegenomics.oist.jp/cots/viewer/info?project_id=46), with gene models for all Nk cluster genes highlighted in OKI and GBR V1.0 genome assemblies.

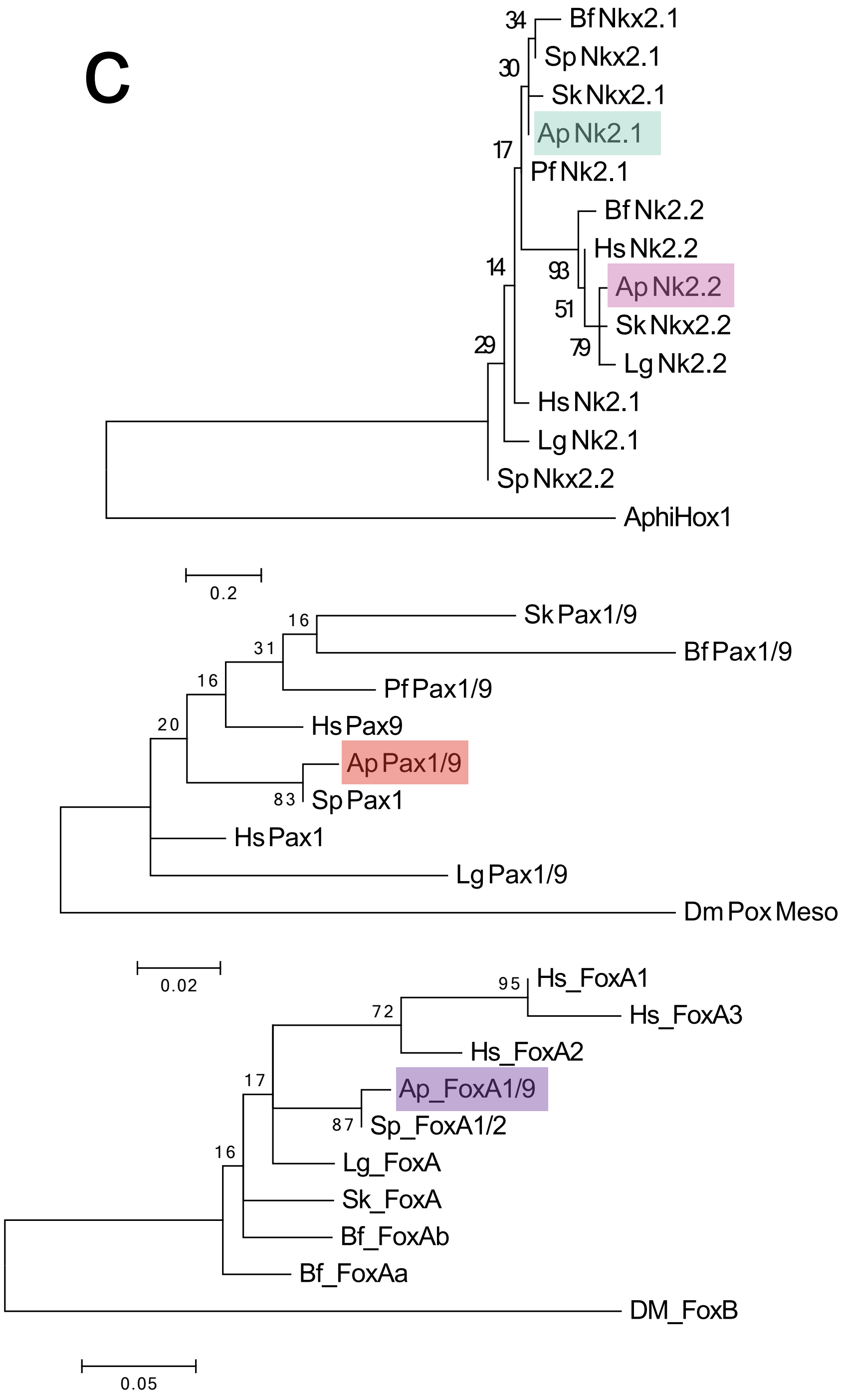
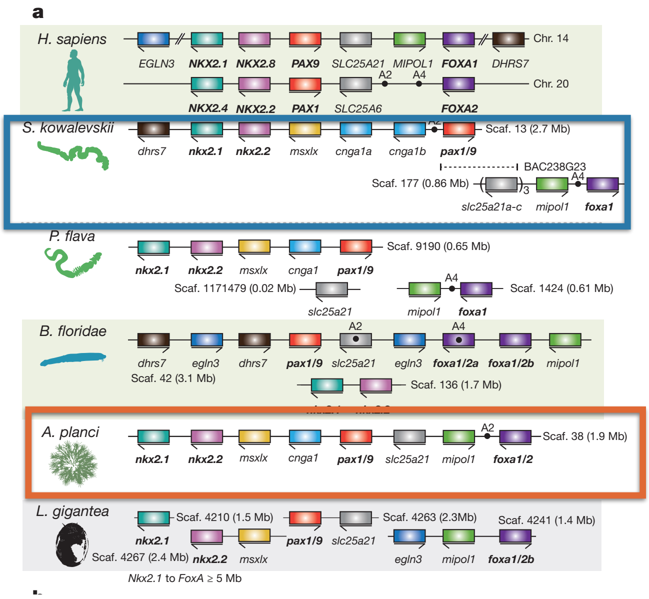
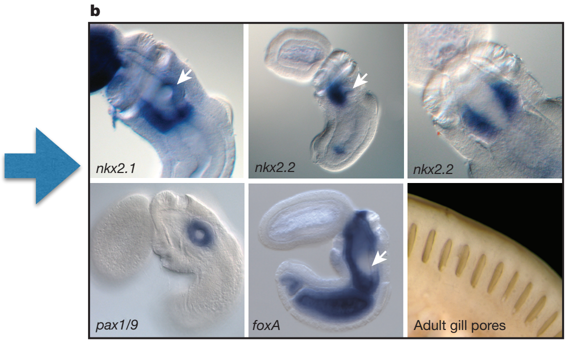


Figure 3.19. Phylogenetic Analysis of COTS Nk cluster genes.
Molecular phylogenetic analysis of Nkx gene domains from *Acanthaster planci (Ap), Saccoglossus kowalevskii (Sk), Ptychodera flava (Pf), Drosophila melanogaster (Dm), Homo Sapiens (Hs), Branchiostoma floridae (Bf), Lotia gianta (Lg)* and *Stronglocentrotus purpuratus (Sp).*



Simakov, O. et al. Hemichordate genomes and deuterostome origins. Nature (2015).



Simakov, O. et al. Hemichordate genomes and deuterostome origins. Nature (2015).

GENE NAME	oki gene model	gbr gene model	Oocyte (RNAseq FPKM)	Early gastrula	Mid gastrula
>Bra	oki.scaffold15.190	gbr.scaffold25.115	0.0	23.84	20.76
>Nk2.1	oki.38.92	gbr.63.31	0.67	2.86	10.95
>Nk2.2	oki.38.88	gbr.63.37	0.0	0.0	0.98
>Pax1/9	oki.38.81	gbr.63.44	1.45	0	0
>FoxA	oki.38.71	gbr.63.56	1.9	93.9	74.22

Figure 3.20. COTS Nk cluster RNAseq expression.
Hemichordate Nk cluster gene expression (outlined in blue) by *in situ* hybridization. The table on bottom shows RNAseq expression in COTS in FPKM (fragments per kilobase of exon per million fragments mapped), for Oocyte, Early gastrula, and Mid gastrula. Brachyury (Bra) for reference.

3.4 Systems biology analysis of COTS 1-MA-dependent oocyte maturation

The follow text was adapted and updated from a final presentation submitted to the OIST A402 Computational and Mathematical Biology Course.

(a) Introduction:

Almost 50 years ago, 1-methlyadenine (1-MA) was identified as the hormone responsible for inducing starfish oocytes to prepare for fertilization, via resuming meiosis(Kanatani 1964). This discovery resulted in a number of exciting findings related to the basic cell biology of meiosis(Ikegami et al. 1967; Shirai et al. 1972; Kishimoto & Kanatani 1976), and led to the concept of molecular control of the cell cycle(Draetta et al. 1989). Importantly, the advent of an ‘timed’ induction of maturation for a vast quantity of eggs was particularly useful for biochemical methods. 1-methyladenine (1-MA) is a hormone released by radial nerves, which induces female starfish to eject oocytes, which in turn causes the oocytes to initiate meiosis in preparation for fertilization (Kanatani 1964). Oocytes undergo a variety of cell signaling events upon 1-MA stimulation, many of which have been described in detail (Figure 3.21). I used a Systems Biology Graphical Notation (SBGN) approach to summarize these events, and map them to COTS gene models and RNA transcripts. By connecting 1-MA signaling on the oocyte plasma membrane to the cyclin-dependent cell cycle resumption mechanisms in the nucleus in a quantitative manner, these results connect COTS genomic data to the current sea star literature.

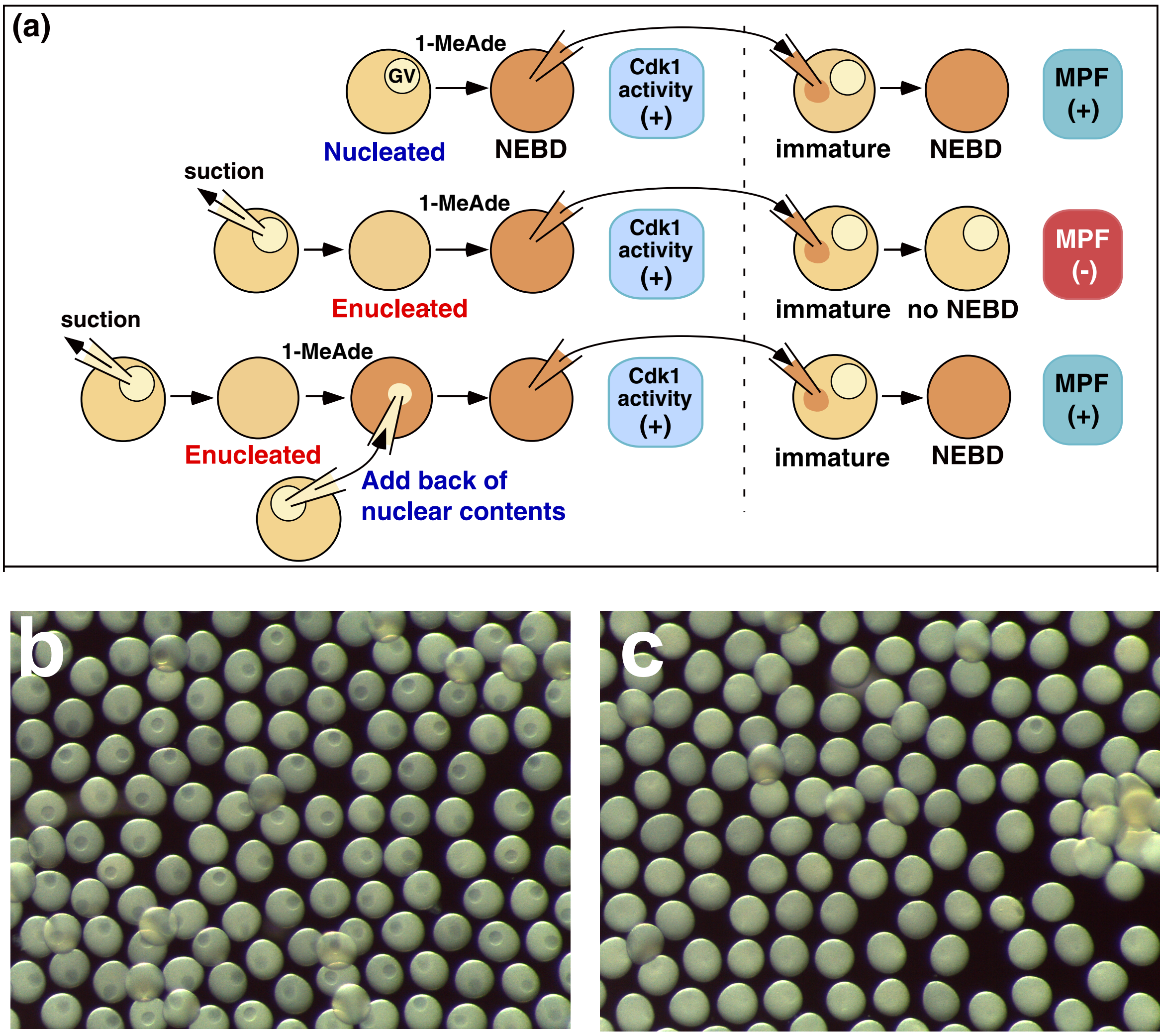


Figure 3.21. 1-MA-mediated oocyte maturation.

a, Summary of 1-MA (1-MeAde) mediated nuclear envelope breakdown (NEBD), and the myriad experiments confirming that cytoplasm along with nuclei, are required for resumption of the cell cycle, and oocyte maturation. Adapted from(Kishimoto 2015) **b,** COTS oocytes. **C,** COTS oocytes after 15 minutes of 1 uM 1-MA treatment.

(b) Methods

Primary literature on starfish oocyte biology was downloaded using http://scholar.google.com. The 1-MA oocyte meiotic resumption pathway was mapped in CellDesigner4.3 (http://www.celldesigner.org), and converted to Systems Biology Graphical Notation (SBGN) via the conversion option in the CellDesigner4.3 software. Two recent reviews were used as reference to summarize both components of, and evidence for, various steps in the starfish1-MA oocyte induction pathway (Kalachev 2013),(Kishimoto 2011)

(Figure 3.21a). Figure 3.22b is taken from figure#9 in (Mita et al. 1999), which describes the synthesis of 1-MA from ATP, in detail.

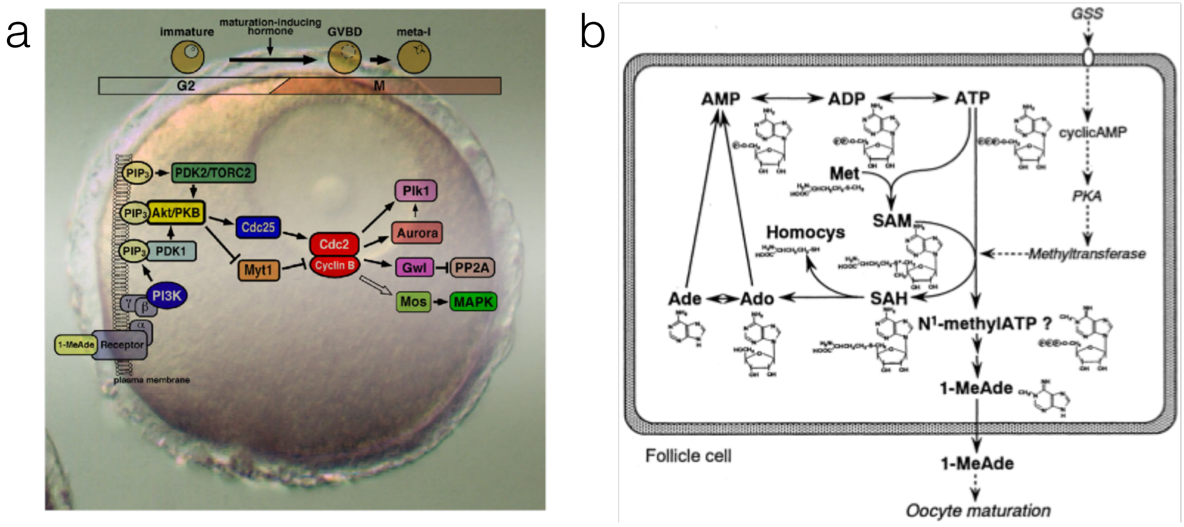


Figure 3.22. 1-MA signaling in starfish oocytes.
a, A model for cell cycle signaling components in starfish oocyte germinal vesicle (GV) breakdown and meiotic resumption. Adapted from (Kishimoto 2011). **b,** 1-MA chemical synthesis. Adapted from (Mita et al. 1999).

(c) Results and Discussion.

Figure 3.23 is a proposed Systems Biology Graphical Notation (SBGN) model that summarizes the key events of the 1-MA signaling pathway. The diagram includes three cellular compartments. The first compartment is the “Radial Nerve Cell” in which gonad-stimulating substance (GSS) stimulates 1-MA production. Radial nerves are found relatively close to arrested oocytes, and are known to contain substances capable of inducing cell cycle resumption(Kanatani 1964). Because the focus of the SBGN model is the events connecting 1-MA receptor binding through cyclin activation in the oocyte, both 1-MA production and events downstream of cyclin activation were either abbreviated, or omitted. In cases where intermediate steps have been omitted, all displayed stoichiometry is correct. The second compartment is the plasma membrane of the oocyte, on which the 1-MA receptor, known to be a G-protein-coupled receptor (GPCR) yet still unidentified, is activated upon 1-MA binding. The third compartment is the interior of the oocyte, in which the activated GPCR pathway initiates germinal vesicle breakdown, and resumption of the cell cycle.

The reactions in Figure 3.23 can be summarized as:

- biosynthesis of 1-MA (Table 3.8: re7),
- canonical G-protein coupled receptor activation via PIP2 and PIP3 (re12, re13, re21, re22),
- downstream signaling via PDK and Akt/PKB activation (re23, re27),
- and finally, cyclin activation (re28, re29, re30).

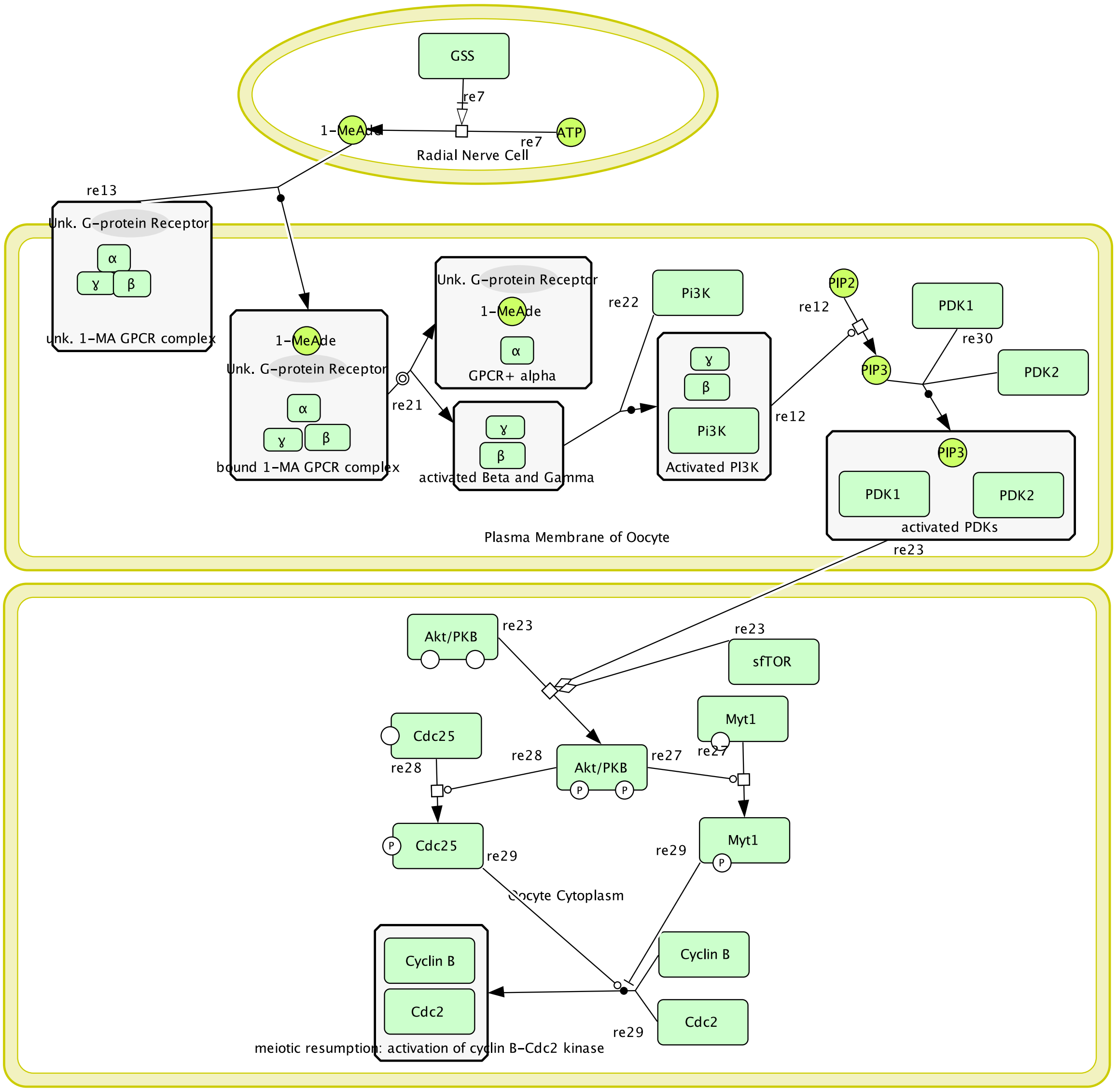


Figure 3.23. 1-MA oocyte resumption in SBGN notation.
3 cellular components are defined; Radial Nerve Cell, Plasma Membrane of the Oocyte, and Oocyte interior (including cytoplasm and nuclei). A COTS candidate gene model or transcript has been identified for each protein component of the pathway.

Table 3.7 includes Template IDs for all components in “Figure 3.23: 1-MA oocyte resumption in SBGN notation,” and lists the associated COTS RNA transcript and genome scaffold. Table 3.8 includes Template PMIDs for all reactions in “Figure 3.23: 1-MA oocyte resumption in SBGN notation.”

Based on the 1-MA oocyte resumption in SBGN notation, there are three conclusions. First, the central role of Akt/PKB (Okumura et al. 2002) become much clearer. In contrast to other styles of signaling diagrams where this central role can only be alluded to, SBGN notation requires both validation via published results as well as accurate accounting of reactants and reactions, resulting in an easily viewable diagram. This diagram is also biochemically robust(Le Novère et al. 2009). Second, an extensive review of the literature with regard to sfTOR and PDK2, coincidentally the most important contribution of the review by Kishimoto (Kishimoto 2011), made clear that while this association has been alluded to be several authors, the basic biochemistry of the reaction remains unknown. Thirdly, the 1-MA receptor remains unknown, though strong evidence indicates that it must a a G-protein coupled receptor mediated(Jaffe 1993).

All three observations could be made without the use of SBGN notation, but quantitative nature of SBGN allows for a higher level of confidence, while the simplified graphical nature of SBGN diagrams makes them easy to rapidly assess with regard to complex signaling events. Finally, because COTS transcripts and gene models exist for each of the protein components identified by the SBGN diagram, it is possible to consider testing these specific hypotheses in COTS oocytes.

Table 3-7. SBGN Template IDs and associated COTS RNA transcripts.

EPN Name	EPN Type	EPN ID	EPN ID	Synonyms	Organism	paper reference	COTS genome scaffold #	COTS transcript ID
1-MeAde	Metabolite	CHEBI:18083		Methyladenine	Starfish	1. Shirai, H., Kanatani, H. & Taguchi, S. 1-methyladenine biosynthesis in starfish ovary: action of gonad-stimulating hormone in methylation. Science 175, 1366–1368 (1972).		
Pi3K	Protein	GenBank:	Z29090.1	phosphatidylinositol 3-kinase	Human	Molecular Cloning, cDNA Sequence, and Chromosomal Localization of the Human Phosphatidylinositol 3-Kinase p110α (PIK3CA) Gene Original Research Article Genomics, Volume 24, Issue 3, December 1994, Pages 472-477 Stefano Volinia, Ian Hiles, Elizabeth Ormondroyd, Dean Nizetic, Rachele Antonacci, Mariano Rocchi, Michael O. Waterfield	scaffold13	comp179390_c1_seq1
Akt/PKB	Protein	UniProt:	Q95YJ0	kinase Akt/PKB	Starfish	1. Okumura, E. et al. Akt inhibits Myt1 in the signalling pathway that leads to meiotic G2/M-phase transition. Nat. Cell Biol. 4, 111–116 (2002).	scaffold24	comp180357_c1_seq1
Myt1	Protein	UniProt:	Q95YJ1	inactivator of cyclin B-Cdc2	Starfish	1. Okumura, E. et al. Akt inhibits Myt1 in the signalling pathway that leads to meiotic G2/M-phase transition. Nat. Cell Biol. 4, 111–116 (2002).	scaffold469	comp174497_c0_seq2
Cdc2	Protein	UniProt:	Q17066		Starfish	1. Okumura, E., Sekiai, T., Hisanaga, S.-I., Tachibana, K. & Kishimoto, T. Initial triggering of M-phase in starfish oocytes: a possible novel component of maturation-promoting factor besides cdc2 kinase. J. Cell Biol. 132, 125–135 (1996).	scaffold453	comp169732_c1_seq1
CyclinB	Protein	UniProt:	P90681		Starfish	1. Okumura, E., Sekiai, T., Hisanaga, S.-I., Tachibana, K. & Kishimoto, T. Initial triggering of M-phase in starfish oocytes: a possible novel component of maturation-promoting factor besides cdc2 kinase. J. Cell Biol. 132, 125–135 (1996).	scaffold361	comp170404_c0_seq2
GSS	Protein	UniProt:	C9K4X8		Starfish	1. Kanatani, H. Spawning of starfish: action of gamete-shedding substance obtained from radial nerves. Science 146, 1177–1179 (1964).	scaffold295	comp157746_c0_seq1
ATP	Metabolite	(CHEBI:15422)			Starfish	1. Mita, M., Yoshikuni, M. & Nagahama, Y. 1-Methyladenine production from ATP by starfish ovarian follicle cells. Biochim. Biophys. Acta 1428, 13–20 (1999).		
PIP2	Metabolite	CHEBI:18348		Phosphatidylinositol-4,5-bisphosphate	Starfish	1. Sadler, K. C. & Ruderman, J. V. Components of the signaling pathway linking the 1-methyladenine receptor to MPF activation and maturation in starfish oocytes. Developmental Biology 197, 25–38 (1998).		
PIP3	Metabolite	CHEBI:16618		phosphatidylinositol (3,4,5)-trisphos- phate	Starfish	1. Sadler, K. C. & Ruderman, J. V. Components of the signaling pathway linking the 1-methyladenine receptor to MPF activation and maturation in starfish oocytes. Developmental Biology 197, 25–38 (1998).		
PDK1	Protein	UniProt:	Q76BX2	Phosphoinositide dependent kinase-1	Starfish	1. Hiraoka, D., Hori-Oshima, S. & Fukuhara, T. PDK1 is required for the hormonal signaling pathway leading to meiotic resumption in starfish oocytes. Developmental … (2004).	scaffold292	comp174076_c1_seq1
PDK2	Protein	None (hypothetical protein)		3- phosphoinositide-dependent kinase-2	Starfish	Dong, L. Q. PDK2: the missing piece in the receptor tyrosine kinase signaling pathway puzzle. *AJP: Endocrinology and Metabolism* **289,** E187–E196 (2005).		
sfTOR	Protein	UniProt:	E2RWP8	Target of rapamycin	Starfish	1. Hiraoka, D., Okumura, E. & Kishimoto, T. Turn motif phosphorylation negatively regulates activation loop phosphorylation in Akt. Oncogene 30, 4487–4497 (2011).	scaffold51	comp171410_c0_seq9
Cdc25	Protein	GenBank:	AB076395.1		Starfish	1. Draetta, G. et al. Cdc2 protein kinase is complexed with both cyclin A and B: evidence for proteolytic inactivation of MPF. Cell 56, 829–838 (1989).	scaffold1	comp171899_c0_seq2
1-MA GPCR	Protein	(unknown)			Starfish	1. Sadler, K. C. & Ruderman, J. V. Components of the signaling pathway linking the 1-methyladenine receptor to MPF activation and maturation in starfish oocytes. Developmental Biology 197, 25–38 (1998).		
Gα	Protein	UniProt:	P30676	Guanine nucleotide-binding protein G(i) subunit alpha	Starfish	1. Jaffe, L. A. Oocyte maturation in starfish is mediated by the beta gamma-subunit complex of a G-protein. The Journal of Cell Biology 121, 775–783 (1993).	scaffold321	comp163322_c0_seq1
Gβ	Protein	GenBank	AB894321.1	Guanine nucleotide-binding protein G(i) subunit beta	Starfish	Mita,M., Haraguchi,S., Watanabe,M., Takeshige,Y., Yamamoto,K. and Tsutsui,K.TITLE Involvement of Gs-alpha proteins in the action of relaxin-likegonad-stimulating substance on starfish ovarian follicle cellsJOURNAL Unpublished	scaffold157	comp157533_c0_seq1
Gγ	Protein	UniProt:	Q568H1	Guanine nucleotide-binding protein G(i) subunit gamma	Zebrafish	1. Jaffe, L. A. Oocyte maturation in starfish is mediated by the beta gamma-subunit complex of a G-protein. The Journal of Cell Biology 121, 775–783 (1993).	scaffold56	comp174571_c1_seq1

Table 3-8. 1-MA Reactants.

Model ID#	Process Type	Process	Modifier	EC	Modification Type	Organism	Cell Type	Subcellular Location	PMID/lit ref.
id=re7	Metabolic reaction	ATP->1-MeAde	GSS		Catalysis	Starfish	Radial Nerve Cell		1. Mita, M., Yoshikuni, M. & Nagahama, Y. 1-Methyladenine production from ATP by starfish ovarian follicle cells. Biochim. Biophys. Acta 1428, 13–20 (1999).
id=re12	Metabolic reaction	PIP2->PIP3	Pi3K, Gβ, Gγ		Catalysis	Starfish	Oocyte	Plasma Membrane	1. Sadler, K. C. & Ruderman, J. V. Components of the signaling pathway linking the 1-methyladenine receptor to MPF activation and maturation in starfish oocytes. Developmental Biology 197, 25–38 (1998).
id=re13	Receptor/Ligand binding	1-MeAde +GPCR			Binding	Starfish	Oocyte	Plasma Membrane	1. Shirai, H., Kanatani, H. & Taguchi, S. 1-methyladenine biosynthesis in starfish ovary: action of gonad-stimulating hormone in methylation. Science 175, 1366–1368 (1972).
id=re21	Complex dissociation	GPCR complex-> Gβ, Gγ			GPCR subunit activation	Starfish	Oocyte	Plasma Membrane	1. Shirai, H., Kanatani, H. & Taguchi, S. 1-methyladenine biosynthesis in starfish ovary: action of gonad-stimulating hormone in methylation. Science 175, 1366–1368 (1972).
id=re22	Complex formation	Gβ, Gγ + Pi3K			Activation	Starfish	Oocyte	Plasma Membrane	1. Sadler, K. C. & Ruderman, J. V. Components of the signaling pathway linking the 1-methyladenine receptor to MPF activation and maturation in starfish oocytes. Developmental Biology 197, 25–38 (1998).
id=re23	Signalling	Akt -> Akt (PP)	PDK1, PDK2, TOR		Phosphorylation	Starfish	Oocyte	Cytoplasm	1. Okumura, E. et al. Akt inhibits Myt1 in the signalling pathway that leads to meiotic G2/M-phase transition. Nat. Cell Biol. 4, 111–116 (2002).
id=re27	Signalling	Myt -> Myt (P)	Akt (PP)		Phosphorylation	Starfish	Oocyte	Cytoplasm	1. Okumura, E. et al. Akt inhibits Myt1 in the signalling pathway that leads to meiotic G2/M-phase transition. Nat. Cell Biol. 4, 111–116 (2002).
id=re28	Signalling	Cdc25 -> Cdc25 (P)	Akt (PP)		Phosphorylation	Starfish	Oocyte	Cytoplasm	1. Okumura, E., Sekiai, T., Hisanaga, S.-I., Tachibana, K. & Kishimoto, T. Initial triggering of M-phase in starfish oocytes: a possible novel component of maturation-promoting factor besides cdc2 kinase. J. Cell Biol. 132, 125–135 (1996).
id=re29	Complex formation	Cdc2+CyclinB	Myt(-), Cdc25(+)		Activation	Starfish	Oocyte	Plasma Membrane	1. Draetta, G. et al. Cdc2 protein kinase is complexed with both cyclin A and B: evidence for proteolytic inactivation of MPF. Cell 56, 829–838 (1989).
id=re30	Complex formation	PIP2+PDK1+ PDK2			Activation	Starfish	Oocyte	Plasma Membrane	1. Hiraoka, D., Hori-Oshima, S. & Fukuhara, T. PDK1 is required for the hormonal signaling pathway leading to meiotic resumption in starfish oocytes. Developmental … (2004).

Chapter 4 : Discussion

4.1 The COTS genome as a guide for biocontrol measures: Next steps.

4.2. Are COTS marine pests?

4.3 Are there differences between aggregating and endemic COTS populations?

4.4 What causes COTS aggregations?

4.5 COTS as model system for the study of genomic structure.

4.1 The COTS genome as a guide for biocontrol measures: Next steps.

The crown-of-thorns starfish genome, as presented in this thesis, is notable for three reasons. First, the overall quality of the assembly was remarkably good, an order of magnitude better than other marine invertebrates, echinoderms in particular. Second, the likely reason for this excellent assembly is the lack of overall heterozygosity, both within each genome, and between OKI and GBR assemblies. Last, the long scaffolds of the assembly have biological significance; the discovery of two evolutionarily-relevant gene clusters confirms that the sequenced genome assembly likely represents the true order of the COTS genome. Taken together, this suggests that the COTS genome is sequenced at a higher resolution than previous echinoderm genome assemblies.

In a recent publication, the COTS genome was used as a reference for identifying peptides secreted by COTS under different behavioral conditions, and subsequently for a bioinformatics approach that found COTS-specific peptides to be used as targets for COTS biocontrol measures (Figure 4.1) (Hall et al. 2017). Perhaps deemphasized in that report was the fact that the surprisingly high quality of the COTS assemblies was a data point itself; the aspects of the COTS genomic structure that made it amenable to short-read sequencing technology, also directly address an open question in COTS biology; have COTS population dynamics recently been perturbed by anthropomorphic causes? In other words, has human activity over the past 50 to 100 years led to a dramatic (e.g. 4 to 5 orders of magnitude) increase in the total COTS populations size? The structure and analysis of the COTS genome are consistent with a recent and rapid COTS population expansion, and thus highlight the next steps that should be taken to definitively answer this critical question, in a quantitative manner.

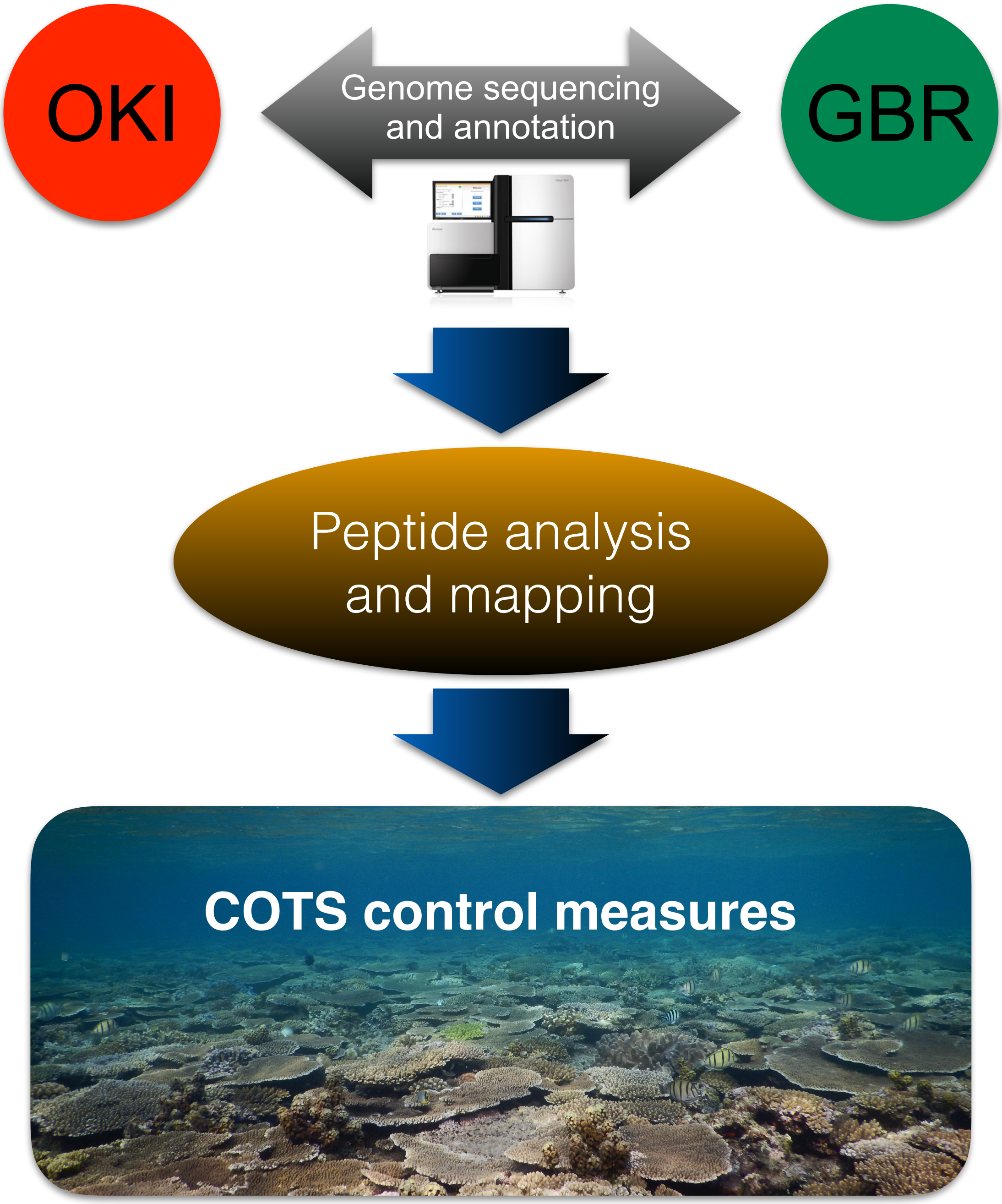

OKI
Genome sequencing and annotation
GBR
Peptide analysis and mapping
COTS control measures

Figure 4.1. . Summary of bioinformatics work flow related to (Hall et al. 2017)

4.2. Are COTS marine pests?

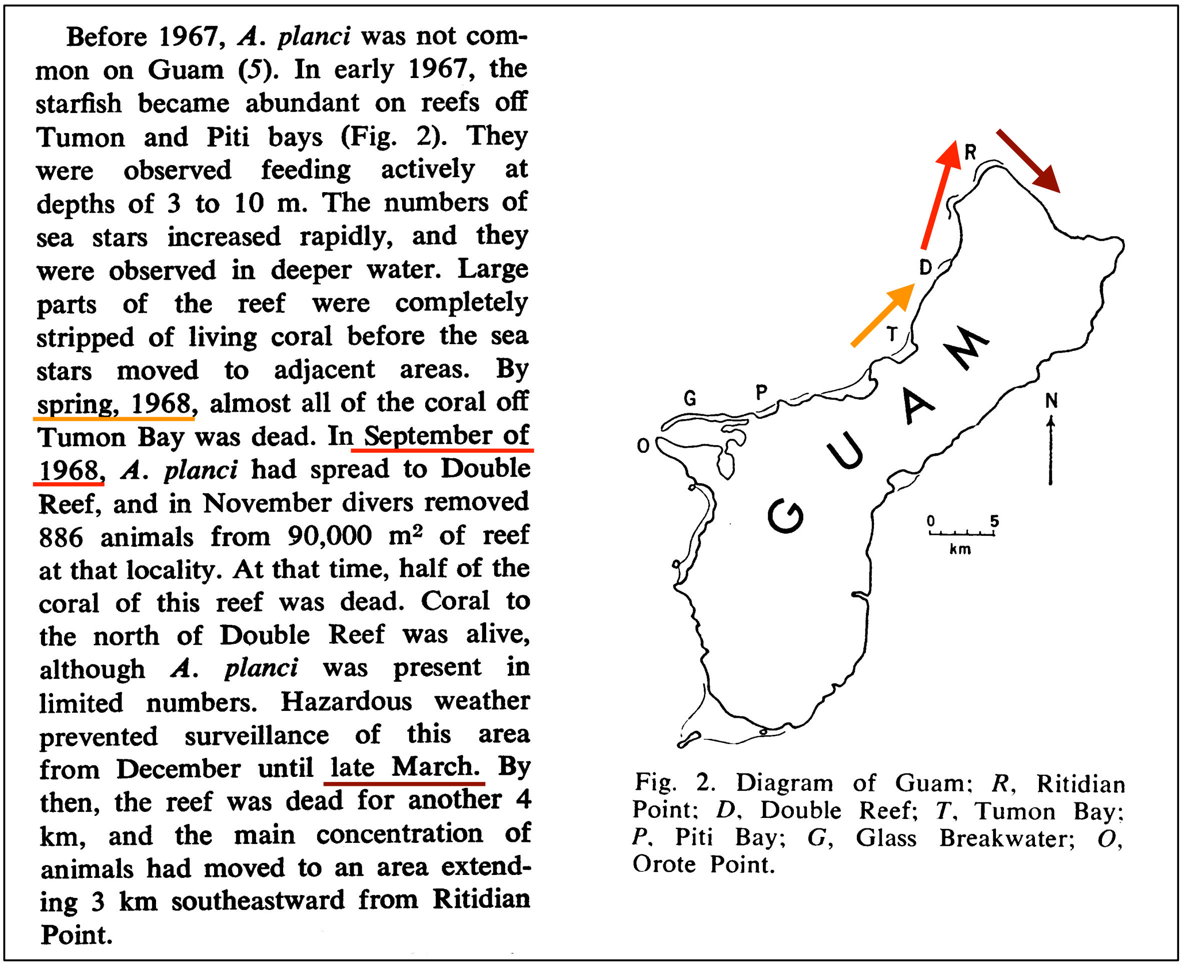
The COTS genome addresses three aspects of the 'COTS as pests' discussion. First, the hallmarks of a high-quality genome assembly, namely long scaffolds, intact gene synteny, and low heterozygosity within the assembly, result from lower than expected heterozygosity, consistent a recent COTS population bottleneck. Second, the comparison between OKI and

GBR directly addresses the question of whether differences between aggregating and endemic or non-aggregating populations exist, finding no evidence for a genome-based difference. Last, based on the assumption that COTS population densities have increased recently, the COTS genome does not provide a mechanism for these increases, but similarity between two genomes from specimens collected over 6000 km apart is consistent with the currently hypothesis, namely that COTS larva have increased survivorship in nutrient-rich seawater.

To address how the COTS genome data are relevant to COTS population size history, it is important to review how opinions on this topic have evolved over the past 50 years. As summarized in the introduction, COTS ecological research can roughly be broken into three phases, which related to how researchers have answered the question of 'Are COTS pests? (Sapp 1999). In the first phase, from 1960s to the late 1970s, the answer was 'undoubtedly yes,' with the primary focus on collecting reef survey data to quantify the damage. In the second phase, as the initial COTS aggregations subsided, some coral recovery was observed and the answer shifted to 'perhaps no, COTS populations naturally fluctuate' on the basis that geological data seemed to indicate that COTS naturally followed boom/bust cycles. In the last phase beginning in 1980s, extensive ecological observation of additional major COTS infestations on the Great Barrier Reef and the subsequent lack of effective coral recovery support our current answer; 'Yes COTS are pests, though both previous observations are true.' In other words, our current understanding is that COTS population size naturally fluctuates and population outbreaks are natural in a sense, but the frequency of those population expansion events has recently increased, and now exceeds the natural recover rate of corals (De'ath et al. 2012).

Prior to the 1960s, COTS were described as being an exceedingly rare organism, generally observed only once or twice on region-wide, multi-year sampling excursions (Sapp

1999). The first evidence for COTS causing extensive damage to coral reefs began with observations of large numbers of COTS on Miyako island of the Ryukyu Islands in 1957 (Yamaguchi 1986). By the early 1970s, numerous examples of COTS aggregations had been described across the Pacific (see Table 1.1 for a summary of high profiles publications). Generally, devastation begins as local COTS populations expand dramatically, and thousands to millions of starfish aggregate in one area before systematically eating and migrating together en mass, decimating all coral in their path (Chesher 1969; Sapp 1999). These aggregations move along the reef (Figure 4.2), persist until all coral are eaten, and can spawn secondary aggregations that occur in subsequent years and adjacent regions (Sapp 1999; Birkeland & Lucas 1990). Thus, the first phase of COTS research established methods for quantifying COTS population densities, and confirming that 'COTS population density increases' over the past half century correlate with measurable damage to coral reefs across the indo-pacific region (Chesher 1969; Pearson 1972; De'ath et al. 2012)



Before 1967, *A. planci* was not common on Guam (5). In early 1967, the starfish became abundant on reefs off Tumon and Piti bays (Fig. 2). They were observed feeding actively at depths of 3 to 10 m. The numbers of sea stars increased rapidly, and they were observed in deeper water. Large parts of the reef were completely stripped of living coral before the sea stars moved to adjacent areas. By spring, 1968, almost all of the coral off Tumon Bay was dead. In September of 1968, *A. planci* had spread to Double Reef, and in November divers removed 886 animals from 90,000 m² of reef at that locality. At that time, half of the coral of this reef was dead. Coral to the north of Double Reef was alive, although *A. planci* was present in limited numbers. Hazardous weather prevented surveillance of this area from December until late March. By then, the reef was dead for another 4 km, and the main concentration of animals had moved to an area extending 3 km southeastward from Ritidian Point.

Fig. 2. Diagram of Guam: *R*, Ritidian Point; *D*, Double Reef; *T*, Tumon Bay; *P*, Piti Bay; *G*, Glass Breakwater; *O*, Orote Point.

Figure 4.2. COTS Outbreak on Guam (1969).
A high profile 1969 description of a early COTS outbreak of 1968 on the island of Guam. The timeline for the progression of the COTS aggregation around Guam is highlighted in yellow, orange and red, both in the text and on the map. (Chesher 1969)

The frequency of and damage caused by this first wave of aggregations subsided in the mid-1970s. After the Great Barrier Reef recovered, some authors began to suggest that COTS population density fluctuated naturally, raising questions about the extent to which COTS 'aggregation' behavior was abnormal (Sapp 1999). A small number of high profile, contrarian papers promoted the notion that the outbreaks were naturally occurring. The most compelling evidence was a geochemical analysis of reef-front sediment which suggested that large spikes in COTS-related chemical signatures had periodically occurred over the past 5000-7000 years (Walbran, R. A. Henderson, Faithful, et al. 1989; Walbran, R. A. Henderson, Jull, et al. 1989). These studies suggested that periodicity to COTS population size was a naturally occurring phenomenon with large increases occurring on the order of once per 100 years (Figure 4.3), importantly, prior to the intervention of modern anthropogenic impacts.

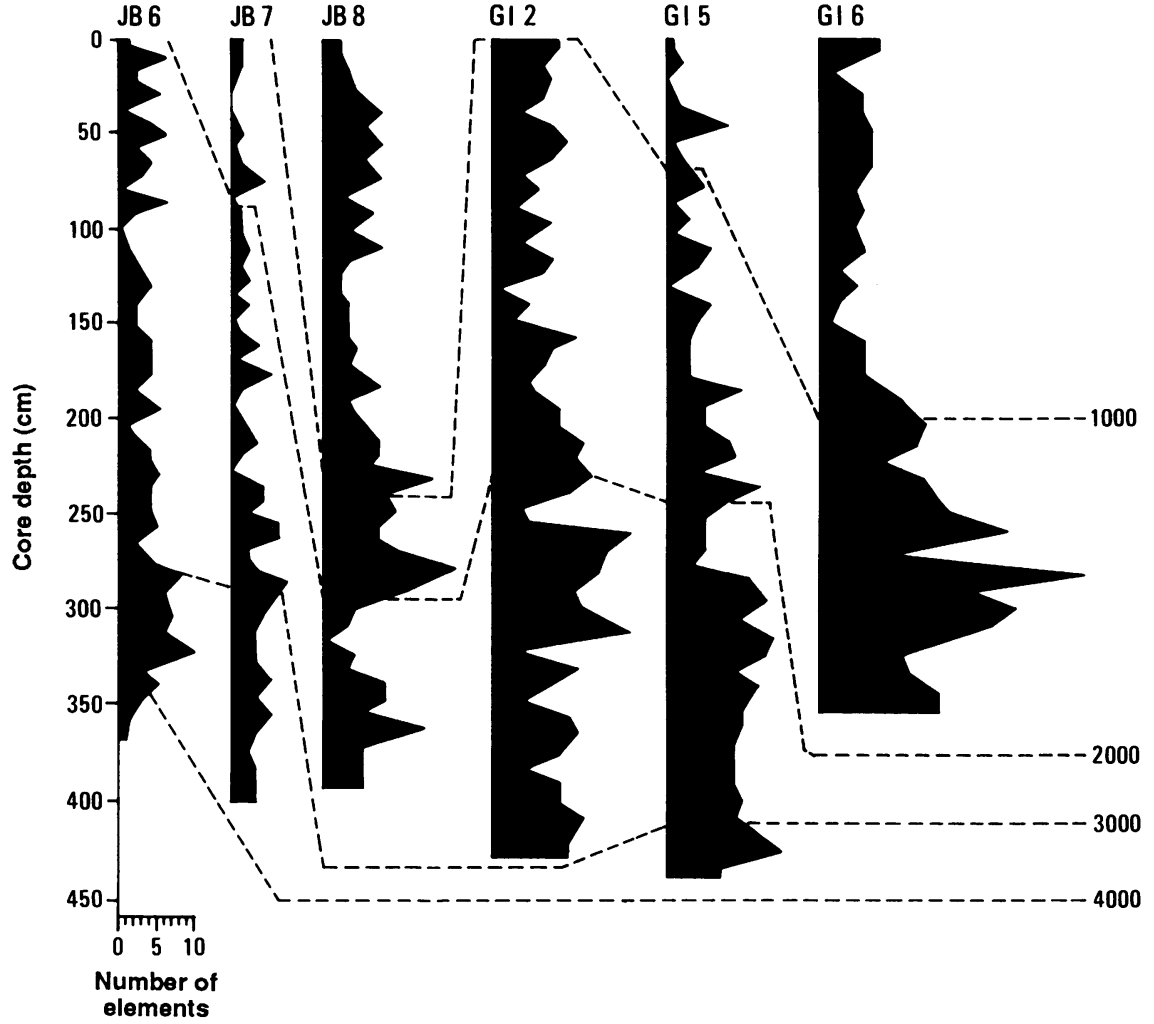


Figure 4.3. Geochemical analysis of COTS-related chemicals in reef-front sediment.
A graph showing the number of COTS skeletal elements found in core samples taken from sediment core samples at John Brewer (JB) and Green Islands (GI)Reef, Great Barrier Reef, Australia. The depth of the core samples is used to calculate the age of the sample. These data suggest that over 4-5000 years, COTS population densities have varied dramatically, prior to human intervention. From (Walbran, R. A. Henderson, Jull, et al. 1989).

Other authors suggested that the outbreaks were natural from an ecological view point, with COTS proposed to be an r-strategist, and outbreaks, not unlike forest fires, serving to increase the biodiversity of coral species on reefs (Moore 1978). Moreover, the failure to describe a direct causal mechanism between human interventions and COTS outbreaks was also cited as evidence for COTS aggregations occurring naturally (Moore & Huxley 1976). These authors also proposed that ‘ COTS aggregations’ were simply the result of oversampling, given that scuba diving and snorkeling activity dramatically increased in usage in the 1960s and 1970s (Sapp 1999). It should be noted that this period of COTS research was highly politicized; as the first efforts to protect coral reefs entered the public discussion, the proposed legislation was met with severe resistance from commercial concerns. The conflicts that characterize this second phase of COTS research were only heightened by a lack of common methods and standardized data for quantifying COTS and coral reef coverage over longer terms, and wider geographic regions (Sapp 1999).

Thus, the third and most recent era of COTS research began with the advent of standardized methods for measuring coral reef coverage and COTS population size in the field, and longitudinal studies done over larger geographic regions. By quantifying evidence for periodicity of COTS population aggregations, the question of whether COTS were in fact damaging coral reef cover was reopened; a large number of publications in lower impact journals established that the rate and magnitude of COTS-related destruction greatly outstripped the reefs ability to rebound. These studies further quantified the immergence of new outbreaks in the second half of the 1970s, and described the correlation of COTS-related aggregations and coral reef destruction with regions with high human exposure (Birkeland & Lucas 1990; Birkeland 1982; Kettle & Lucas 1987). These studies included a major discovery; the observation that outbreaks were preceded by above-average rainfall two to three years prior to the outbreak, generally through increased typhoon-related rainfall following a drought period (Birkeland 1982; Birkeland & Lucas 1990).

Alarmingly, the current literature has been updated with more recent data that confirms an increase in the frequency of COTS outbreaks over the past 50 years, with at least three major outbreaks on the Great Barrier Reef observed since 1966 and a fourth outbreak potentially ongoing (figure 4.4). Importantly, the frequency of these outbreaks is well above the ‘once per century’ historical estimate proposed by COTS geochemical data and reef recovery rates (Fabricius et al. 2010). Moreover, COTS population increases have been observed to be more localized and endemic in some areas. For example, on the Okinawan islands, COTS have been actively removed from reefs by divers since the 1960s, yet show population densities above historical measures (Nakamura et al. 2014).

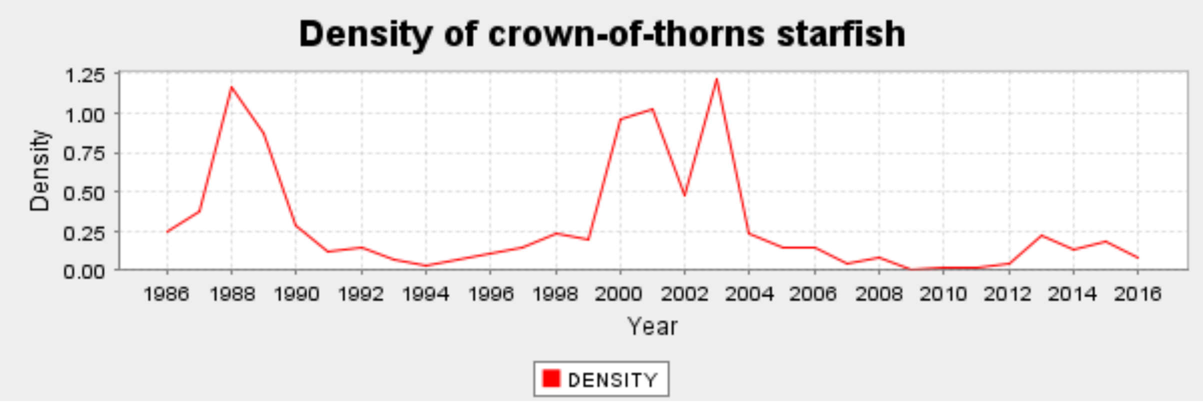


Figure 4.4. COTS Outbreaks on the GBR (1986-2016)
Evidence for a 2nd and 3rd major COTS population density outbreak on the Great Barrier Reef. The Average COTS density per two-minute tow, across the entire Great Barrier Reef increases begin in 1986 and 2000. A 4th event may be occurring, given the elevated levels beginning in 2012. Taken From: http://data.aims.gov.au/waCOTSPage/cotspage.jsp

Although the early COTS studies effectively describe localized devastation of coral reefs (Sapp 1999; Birkeland & Lucas 1990; Moran & De'ath 1992), the fundamental question of whether COTS aggregations were having a measurable impact on coral reefs on a global scale required longitudinal studies done over years or decades, across geographic regions, importantly based on quantification of coral cover. The most definitive study of reef monitoring data to date, taken over 27 years along the entirety of the Great Barrier Reef, revealed that COTS starfish account for 42% of coral loss and are the second most important factor impacting reef coverage following cyclones (De'ath et al. 2012). This monumental report (Table 4.1) established that regardless of the frequencies of COTS aggregations, coral mortality directly correlates with the presence of COTS outbreaks on the same reefs. Although the report definitively confirms and quantifies the role for COTS aggregations in increased coral mortality, given that over the past 50 years COTS have been the focus of more reef management efforts than any other species (Birkeland & Lucas 1990), the most alarming aspect of the report is that it suggests these efforts have not been enough to mitigate the continuing loss of coral cover, at least on the Great Barrier Reef.

Table 4-1. COTS and coral reef mortality.
Results from observing 27 years of Coral Decline on the Great Barrier Reef. The table summarizes the modeled impacts of 3 main causes of coral mortality, based on 27 years of reef sampling data. COTS mortality (highlighted in red) is the only cause for which immediate intervention is possible. Adapted from (De'ath et al. 2012).

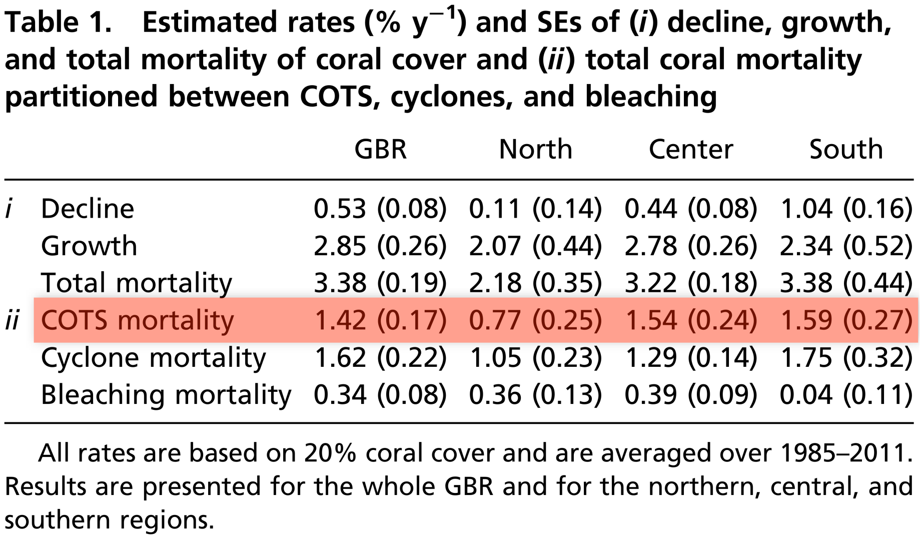


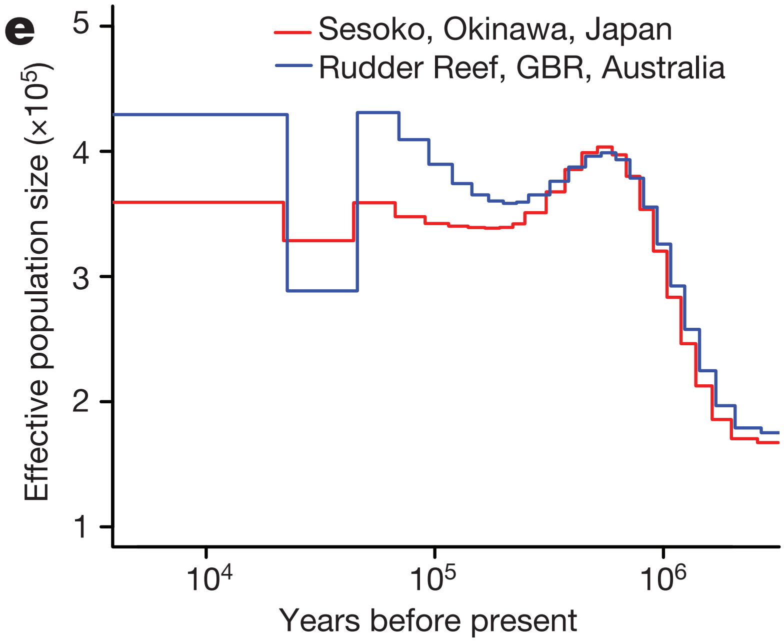
Table 1. Estimated rates (% y^{-1}) and SEs of (*i*) decline, growth, and total mortality of coral cover and (*ii*) total coral mortality partitioned between COTS, cyclones, and bleaching

		GBR	North	Center	South
i	Decline	0.53 (0.08)	0.11 (0.14)	0.44 (0.08)	1.04 (0.16)
	Growth	2.85 (0.26)	2.07 (0.44)	2.78 (0.26)	2.34 (0.52)
	Total mortality	3.38 (0.19)	2.18 (0.35)	3.22 (0.18)	3.38 (0.44)
ii	COTS mortality	1.42 (0.17)	0.77 (0.25)	1.54 (0.24)	1.59 (0.27)
	Cyclone mortality	1.62 (0.22)	1.05 (0.23)	1.29 (0.14)	1.75 (0.32)
	Bleaching mortality	0.34 (0.08)	0.36 (0.13)	0.39 (0.09)	0.04 (0.11)

All rates are based on 20% coral cover and are averaged over 1985–2011. Results are presented for the whole GBR and for the northern, central, and southern regions.

The strongest evidence for recent COTS population expansion that the genomic data can provide is reduced heterozygosity rates, both between the OKI and GBR genome assemblies by BLAST alignment (Figure 3.3) and within each genome assembly by SNP analysis (Figure 3.4). In other words, the statistical characteristics that lead to a high-quality genome assembly, namely low heterozygosity, are consistent with either a recent population bottleneck, or conversely, a population expansion. Notably, COTS are sedentary broadcast spawners that eject small propagules into the water column during mating, which is associated with increased heterozygosity, suggesting even more significance to the low heterozygosity rates in the COTS genome (Romiguier et al. 2014).

Additionally, a statistical analysis of SNP distribution within the COTS genomes which I did not do but was included as part of the recent report (Hall et al. 2017), attempted to determine historical COTS population size (Figure 4.5). The figure shows a large drop off and recovery (in both genomes) between 10^4 and 10^5 years ago following a longer drop off at 10^6 years ago, all during the 'late Pleistocene epoch.' In short, the method works by inverting the classical population genetics approach; instead of measuring heterozygosity in many individuals at a single genetic locus, many individual unlinked loci within a single genome are compared. There are two iterations of the method. 'Pairwise sequentially Markovian coalescent' (PSMC) (Li & Durbin 2011) was the first version and was used for single genomes. The second, multiple sequential Markovian Coalescent Analysis (MSMC) (Schiffels & Durbin 2014) was developed to address multiple genomes from the same species.



(Schiffels & Durbin 2014)

Figure 4.5. COTS MSMC Analysis.

Reproduction of Figure 1e, (Hall et al. 2017). Historical effective population sizes inferred from OKI and GBR genomes using multiple sequential Markovian coalescent analysis (Schiffels & Durbin 2014), assuming a generation time of 3 years and a substitution mutation rate of 1.0 × 10−8 per generation.

There are several major caveats with using this method (which was repeatedly request by one of the manuscript reviewers) to address the question of recent COTS population dynamics. The first problem is the time scale for which both methods were developed, namely human migrations during ice ages over the past hundred thousand to 2 million years, is four orders of magnitude longer than 50 or 100-year time scale relevant to COTS. The minimum limit of resolution for both methods is at least 5000 generations, which for COTS (that take two years to sexually mature) is 10,000 years. Therefore, it is suspicious that COTS population sizes for both OKI and GBR show a major decrease exactly at the minimal time frame for the method (e.g. 10,000 years, or 5000 generations).

The second issue is that COTS populations are known to fluctuate once per century, both from ecological field observations (De'ath et al. 2012)and geochemical analysis of reef front sedimentation (Walbran, R. A. Henderson, Faithful, et al. 1989). Given that COTS are known to have primary and secondary outbreaks, it is unclear how cyclical population dynamics impact the PSMC and MSMC methods (which assume stable population growth). Chapter three of this thesis highlights that the COTS genome in known to be somewhat abnormal with regard to heterozygosity, genome length, and other structural features, at least with regard to other marine invertebrates and echinoderms. In other words, COTS population history and genomic structure present unique challenges to any quantitative analysis or calculation, simply because COT genome structure is already unique.

Finally, it is notable that although extensive COTS population genetics analyses have been undertaken over the past 50 years, none of these reports have directly addressed the recent population size question. In other words, classical population genetics methods have been unable to address perhaps the single most important question of COTS biology; “What is the recent COTS population size?” Thus, attempting to draw population genetics conclusions from a data set from only two COTS genomes and a novel methodology, is unlikely to provide a definitive answer. At the very least, caution should be taken in drawing any conclusions from the PSMC/MSMC results.

There are three main suggestions for a potential COTS population genomics attempt to address the question of recent COTS population size. First, low coverage sequencing should be done from many COTS genomes, from many regions. Given that PCR-based analysis of mitochondrial DNA (mtDNA) have suggested 4 clades within the COTS species world-wide, with evidence for a single pacific clade (Gérard et al. 2008; Yasuda et al. 2014; Vogler et al. 2012; Timmers et al. 2012), it would be critical to confirm that genome-based methods are able to recover these known population subgroups. Additionally, individual samples should be taken from COTS that predate the modern era. For example, genomic DNA may be isolated from museum specimens.

Second, any attempt to determine COTS population size should develop novel analytical tools from first principles. Due to the unique challenges of a shortened time line (e.g. less than 50 generations), the divergence of the COTS genomic structure, and known COTS historical population cyclicality, simply reusing existing bioinformatics techniques are unlikely to provide meaningful results.

Lastly, given the long scaffold lengths of the published assembly, chromosome level resolution for the genome is not an unreasonable goal. A recent attempt to improve the scaffolding using both Dovetail (https://dovetailgenomics.com) and BioNano (http://bionanogenomics.com) methods for genome polishing have resulted in a COTS scaffold N90 of 79 scaffolds. In other words, 90% of the genome length is accounted for by just 79 scaffolds, is in line with the observation that several starfish species are known to have 43 chromosomes. A COTS genomic map (e.g. chromosome-level assembly) would also allow for an EvoDevo-oriented analysis of whether genomic structural variation is related to gene regulator network function.

The high quality of the COTS genomic assembly, the extensive COTS sampling done over the past 50 years, and the eminent threats COTS pose to coral reefs suggest that such an approach to COTS population genomics is likely to provide a robust and quantitative result to the question of recent COT population dynamics.

4.3 Are there differences between aggregating and endemic COTS populations?

A separate question from whether COTS populations have expanded recently, is whether the COTS found in aggregations are distinct from those that are not. Changes in COTS behavior during outbreaks have been observed; outbreak starfish feed during daylight hours, can be non-selective in their coral consumption, and preferentially aggregate with other COTS (Moran 1990). In contrast, non-outbreak starfish feed nocturnally, are highly specific in the coral species they consume, and show no preference or tendency to seek out other starfish (Sapp 1999; Birkeland & Lucas 1990). Moreover, captive COTS show dramatic behavioral changes in response to an as yet undefined genetically, spawning factors (Beach et al. 1975). It is likely that spawning and aggregating behaviors are interrelated, but to date what these factors are, how they are sensed, and how their molecular mechanisms function remain unknown. Genes related to conspecific communication may be related to the behavioral differences between aggregating and non-aggregating COTS.

In the late 1980s, an allozyme approach was used to analyze population genetics of COTS and found allelic variation between aggregating and non-outbreak populations, as well as differences between animals from different regions (J. Benzie & Stoddart 1992; Katoh & Hashimoto 2003; J. A. Benzie 1999; Nishida & Lucas 1988; Nash et al. 1988). In this technique, protein extracts from individual animals are run on native starch gels, in order to separate allozymes by charge. Adding detection reagents directly to the gels and recording enzymatic activity can then be used to distinguish sample specific differences in charge-separated enzymes. For example, in an early analysis, 13 of 30 enzymes analyzed were genetically distinguishable, and 10 were polymorphic (Nash et al. 1988). These allozyme studies established genetic differences between aggregating and non-outbreak populations, as well as differences between animals from different regions. Interestingly, all aggregating starfish are genetically similar, both between regions, and within a region but from outbreaks separated by 15 years (Katoh & Hashimoto 2003).

As a proxy for aggregating versus non-aggregating COTS, I compared OKI and GBR genomes, due to the differences in COTS management ethos. Over 40 years of active starfish-management efforts in different regions has led to different local COTS population dynamics. On the Great Barrier Reef, 10-year cycles of massive "outbreak" events are seen, where little to no active intervention was made prior. Conversely, on Okinawa more endemic and localized COTS population outbreaks are observed, where active removal of COTS has been undertaken for decades (Yamaguchi 1986; Nakamura et al. 2014). This difference in management has changed the COTS size distribution. On the western coast of Okinawa, continuous collection of adult COTS has resulted in approximately 85 ± 8% of starfish belonging to the 10 to 25 cm diameter size class (Nakamura et al. 2014). COTS found on the Great Barrier Reef of Australia are generally larger, with an mean size of 35.4cm (Moran 1990; Sapp 1999) Size distribution has been reported to be related to population density dynamics, with several distinct patterns emerging (Moran 1988). I found no significant differences between OKI and GBR genomes, with regard to genes that were highlighted to be related to conspecific communication (Hall et al. 2017).

Sequencing of the COTS mitochondrial genome in 2006 provided a molecular tool kit for phylogenetics, larval detection, and PCR-based methods for population genetics (Yasuda et al. 2006; Gérard et al. 2008; Vogler et al. 2008; Yasuda et al. 2009; Timmers et al. 2012; Vogler et al. 2012) (Yasuda et al. 2006). One study found evidence for four different clades across the Indo-pacific region, though Okinawan and Australian populations were grouped as a single species (Vogler et al. 2008). Given the large range over which COTS is found, from the Pacific coast of South America to the western coasts of Africa, as well as the contrast in behavior between outbreak and non-break populations, sub-species or cryptic speciation may be prevalent (Appeltans et al. 2012). My analysis of the OKI and GBR mitochondrial genomes confirms that both individuals belong to the pacific clade, and finds no evidence for ‘cryptic’ speciation, at least within the pacific clade.

4.4 What causes COTS aggregations? The larval survivorship hypothesis.

Sufficient data exist today to confirm that human development, particularly with regard to water quality, plays a role in COTS outbreaks (Fabricius et al. 2010). Termed the *larval survivorship hypothesis*, the current consensus is that anthropogenic increases in seawater nitrogen levels result in increased algal load, which are a COTS larval food source, and in turn increased COTS larval survivorship (Brodie et al. 2005; Fabricius et al. 2010; Uthicke et al. 2015; Wolfe et al. 2015; Miller et al. 2015). The genesis of the theory was that localized COTS outbreaks tended to occur two to three years after years with increased rainfall, likely causing nitrate-rich run-off and thus COTS larval food increases (Brodie et al. 2005; Fabricius et al. 2010). Recent studies have confirmed that increased food levels consistent with nitrate-rich run off leads to increased COTS larval survivorship (Wolfe et al. 2015). Other authors rephrase the hypothesis as the *enhanced nutrient hypothesis*, to distinguish it from the *larval resilience hypothesis,* which states that COTS larva are oligotrophic, and thrive in low nutrient and oxygen depleted water (Caballes et al. 2016). My analysis of the COTS genome does not directly impact the COTS larval survivorship hypothesis, though the long gestational period of COTS larva may be the key to understanding the cyclical nature of COTS outbreaks, and is consistent with the low overall heterozygosity observed.

4.5 A new hypothesis for COTS aggregations: sick corals.

How COTS overcome coral defenses and consume live corals remains unclear, though the question raises an interesting possibility about the outbreaks, and a potential means for reconciling seemingly contradictory observations. Given that corals are known to be particularly sensitive to water quality (Shinzato et al. 2011), it is possible that COTS outbreaks are the symptom of unhealthy coral populations, rather than a root cause? Although a careful review of the current literature does not provide any meaningful evidence for or against this notion, three potentially interesting comments can be made.

First, the COTS coral feeding literature from the 1970s used broken coral, coral extracts, or corals in tanks i.e. corals that were stressed, as the fragility of corals was unknown at that time. Although no study of COTS preference for weak or sick coral was done, revisiting these early behavioral experiments and field observations raises some interesting possibilities. For example, the basic aggregation behavior, in which groups of starfish remained together on a single coral head until it was completely digested, before moving to adjacent healthy corals, now appears quite logical, in contrast to the quandary this observation raised when it was first reported (Chesher 1969; Branham et al. 1971). Moreover, perhaps the damaged coral is the source of the aggregation factor, versus starfish themselves (Ormond et al. 1973). A behavioral study reported that amino acids in coral extracts, but not nematocysts, induced COTS behavioral changes (Moore & Huxley 1976).

Second, unhealthy coral populations may explain why COTS among all other echinoderms multiply during outbreak conditions, which the larval hypothesis does not address. To date, explanations for the outbreaks are based on two observations; that COTS aggregating behavior is an abnormal historical anomaly (De'ath et al. 2012), and that outbreaks occur in regions with increased human pressure, due to elevated seawater nitrogen and increases COTS larval survivorship (Fabricius et al. 2010; De'ath et al. 2012). Yet, the *larval survivorship hypothesis* fails to explain why COTS populations increase relative to all other echinoderms. Indeed, the similarity of COTS larvae to other echinoderm larvae has been a frustrating hurdle for COTS population genetics until the advent of mtDNA screening tools (Yasuda et al. 2006). One possibility is that other echinoderm species are also undergoing similar dramatic population fluctuations, but these masses of starfish and urchins are not as visible or easy to observe as COTS. I propose that at the larval level, there may not be a COTS-specific 'larval survivorship' effect, but that these other echinoderms lack adequate food sources and settlement locations for adults to mature. If corals weakened by local water conditions (temperature, salinity, or nitrate levels) are simply unable to fend off predation, this dramatic increase in forage for COTS may lead to the positive feedback cycles observed during outbreaks. With this view in mind, the endemic nature of the COTS on the islands of Okinawa, in the face 50 years of heroic starfish collection efforts, may seem more reasonable (Yamaguchi 1986). For example, in 1997 the town of Onna-son collected over 59 tons or 169,631 individual starfish alone during an outbreak, but in intermittent non-outbreak years, yields rarely fell below 3 to 5 tons or on around 10,000 starfish (Katoh & Hashimoto 2003; Nakamura et al. 2014). No matter how many starfish are collected, if coral are unable to protect themselves, starfish populations will continue to remain at population densities that are elevated relative to historical levels. Intriguingly, a recent report confirms that both fish and coral larva dramatically prefer water samples from healthy protected reefs versus fished, damaged reefs; sick corals give off a unique chemosensory signal (Dixson et al. 2014).

Lastly, the notion that sick coral may be driving the outbreaks, and not vice versa, can explain some of the historically contentious data sets and observations. The initial reports on COTS from the late 1960s and early 1970s overstated the long term impacts of the starfish by usage of poor extrapolation techniques (Sapp 1999). Yet these studies developed population measurement methods that were very accurate, and have resulted in the datasets COTS biologists have access to today (De'ath et al. 2012). Conversely, the second wave of reports, which highlighted the predictive failure of the first reports, also clearly established that at least aspects of COTS behavior are rooted in pre-human ecological cycles. Hence, data have now accrued for both sides debate enters its fifth decade… are the outbreaks natural? Or are they caused by humans? If human behavior and ecological impact is damaging corals and making them more susceptible to COTS outbreaks, both positions (and data sets) can be true. If outbreaks are a symptom of unhealthy coral, and not the converse, outbreaks can follow from naturally occurring ecology, and yet be magnified by human ecological impacts. Definitively showing that COTS outbreaks are dependent on unhealthy coral is beyond the scope of a study focused on genomic analysis, but an annotated genome for population genomics-based methods is an excellent place to begin exploring this hypothesis. Based on this new theory, I would predict large COTS outbreaks in two to three years (2019, 2020) due to the extensive coral bleaching of the past two years (2016, 2017).

Chapter 5 : Conclusions

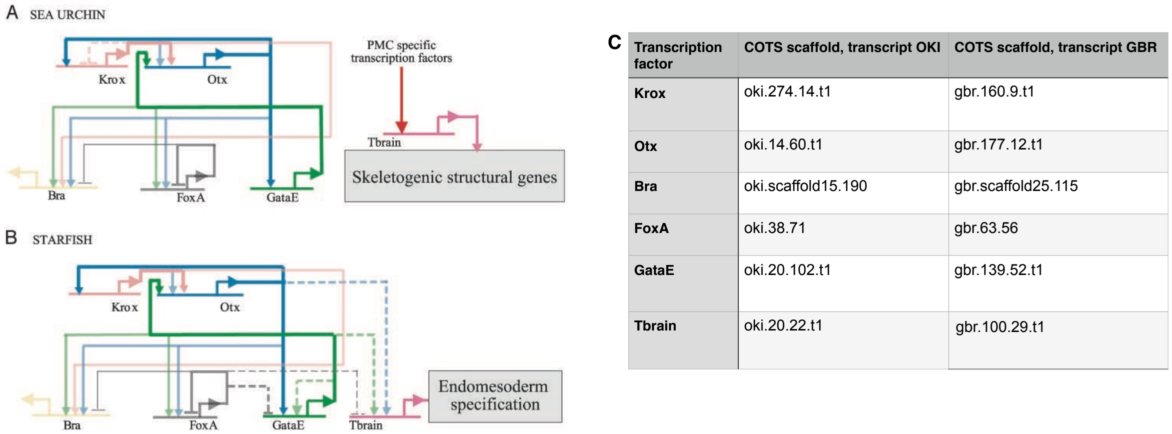
The necessary and sufficient information required to organize and grow complex body plans exists within the genome. How and when this information is used remains the critical question of developmental biology in the current era. More recently, comparisons between how diverged species use the same genetic toolkits to assemble myriad body plans and novel structures has shed light on the fundamental mechanics of evolution. Therefore, sequencing and analyzing the genomes, particularly for species that display divergent body plans, such as echinoderms, is a potentially useful way to understanding how body plan organization has evolved. Moreover, the dramatic reduction in costs and technical challenges associated with short-read or next-generation sequencing now allow for genome projects of non-model organisms. To this end, I chose to study the Crown-of-Thorns Starfish (*Acatasther planci* or 'COTS').

5.1 COTS as model system for the study of genomic structure

In reflecting on the meaning and value of having sequenced the COTS genome, I was deeply moved by the historical significance of this technical achievement, in the context of biology and evolution, in particular. Aside from the potential utility that a COTS genome provides for the eminent threats COTS pose to coral reefs, the primary discoveries I made were directly related to how the information coded in genomes is related to development, and how changes and similarities in the gene clusters that control development, between species, can begin to shed light on the origins of life itself. In other words, if the genome contains the directions for how to build an organism, it follows that changes in the genome will lead to changes in the resulting organism, ergo speciation. That I was able to find both Hox and Nkx clusters intact, at high genomic resolution, was likely directly related to the life history of COTS. Therefore, in thinking about next steps, I would propose that the unique genome characteristics of COTS make it an excellent organism to study how evolution of the gene regulatory networks (GRN) may be related to structural genomics.

In comparing the extensive sea urchin GRN literature with collinearity mechanisms of the Hox cluster, I am struck by the awkwardness of the lack of genomic coordinates in GRNs versus the failure to find or identify additional Hox-like clusters; to date, the Hox cluster remains the only developmentally relevant gene cluster in which genomic organization correlates with gene function (collinearity). It seems obvious to me that within the milieu of enhancers, transcription factor cascades, and GRN-defined tissue specification, there must be a roll for genomic order, synteny, collinearity, perhaps even an 'enhancer code.' That development is so precise, so timely, so reproducible, given the malleability of the genomes that drive it, indicates to me that a missing layer of structure, of linguistics, must exist.

The most important lesson I have learned from sequencing the COTS genome is that this intermediate layer of genomic information likely resides in the heuristics of transcriptional control mediated by transcription factors, akin to the collinearity of the Hox cluster, but on a grander scale. Put another way, when I started this thesis work, I was convinced that the pluripotency of pluripotent stem cells (iPS) via four transcription factors had effectively undone a century of careful embryology and developmental biology. How meaningful could the developmental process be? Four transcription factors could function as a cellular reset button! Four years later, I now see that the challenge is the lack of a mechanism for precisely extracting relevant information from genomes, at the right time and place, in the right cell types; iPS cells simply prove how robust the power of transcriptional control actually is. The last figure of this thesis (Figure 5.1) is a mapping of GRN components found to be both conserved and diverged between starfish and urchin, to the COTS genome. Creating a COTS genomic map (e.g. chromosome-level assembly) would allow for interrogation of genome structure, and perhaps the mapping of the genomic context for gene regulatory networks.


A SEA URCHIN
Krox
Otx
PMC specific transcription factors
Tbrain
Skeletogenic structural genes
Bra
FoxA
GataE
B STARFISH
Krox
Otx
Endomesoderm specification
Bra
FoxA
GataE
Tbrain
C

Transcription factor	COTS scaffold, transcript OKI	COTS scaffold, transcript GBR
Krox	oki.274.14.t1	gbr.160.9.t1
Otx	oki.14.60.t1	gbr.177.12.t1
Bra	oki.scaffold15.190	gbr.scaffold25.115
FoxA	oki.38.71	gbr.63.56
GataE	oki.20.102.t1	gbr.139.52.t1
Tbrain	oki.20.22.t1	gbr.100.29.t1

Figure 5.1. Sea urchin and batstar GRNs mapped to the COTS genomes.
Mapping of **a.** sea urchin (*S. purpuratus*) and **b.** starfish (*P. miniata*) gene reglatory networks that have conservation of wiring and some divergence in function **c.** GRN genes mapped to COTS genome. Note that GataE and Tbrain are on the same scaffold in OKI (scaffold#20) but different scaffolds in GBR.From (Hinman et al. 2003).

References

Appeltans, W. et al., 2012. The Magnitude of Global Marine Species Diversity. *Current biology : CB*, 22(23), pp.2189–2202.

Barnes, D.J., 1970. Field and Laboratory Observations of the Crown-of-Thorns Starfish, Acanthaster planci: Locomotory Response of Acanthaster planci to Various Species of Coral. 228, pp.342–344.

Baughman, K.W. et al., 2014. Genomic organization of Hox and ParaHox clusters in the echinoderm, Acanthaster planci. *genesis*, 52(12), pp.952–958.

Beach, D.H., Hanscomb, N.J. & Ormond, R.F., 1975. Spawning pheromone in crown-of-thorns starfish., 254(5496), pp.135–136.

Benzie, J. & Stoddart, J.A., 1992. Genetic structure of crown-of-thorns starfish (Acanthaster planci) in Australia. *Marine Biology*, 112(4), pp.631–639.

Benzie, J.A., 1999. Major genetic differences between crown-of-thorns starfish (Acanthaster planci) populations in the Indian and Pacific Oceans. *Evolution*, pp.1782–1795.

Birkeland, C., 1982. Terrestrial runoff as a cause of outbreaks of Acanthaster planci (Echinodermata: Asteroidea). *Marine Biology*, 69(2), pp.175–185.

Birkeland, C. & Lucas, J., 1990. *Acanthaster Planci*, CRC Press.

Boetzer, M. & Pirovano, W., 2012. Toward almost closed genomes with GapFiller. *Genome Biology*, 13(6), p.R56.

Boetzer, M. et al., 2011. Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics (Oxford, England)*, 27(4), pp.578–579.

Bolger, A.M., Lohse, M. & Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics (Oxford, England)*.

Branham, J.M. et al., 1971. Coral-Eating Sea Stars Acanthaster planci in Hawaii. *Science*, 172(3988), pp.1155–1157.

Brauer, R.W., Jordan, M.R. & Barnes, D.J., 1970. Triggering of the stomach eversion reflex of Acanthaster planci by coral extracts., 228(5269), pp.344–346.

Brenner, S., 2010. Sequences and consequences. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1537), pp.207–212.

Brodie, J. et al., 2005. Are increased nutrient inputs responsible for more outbreaks of crown-of-thorns starfish? An appraisal of the evidence. *Marine Pollution Bulletin*, 51(1-4), pp.266–278.

Caballes, C.F. et al., 2016. The Role of Maternal Nutrition on Oocyte Size and Quality, with Respect to Early Larval Development in The Coral-Eating Starfish, Acanthaster planci J. G. Knott, ed. *PloS one*, 11(6), pp.e0158007–21.

Camacho, C. et al., 2009. BLAST+: architecture and applications. *BMC bioinformatics*, 10(1), p.421.

Cameron, R.A. et al., 2015. Do echinoderm genomes measure up? *Marine Genomics*, 22, pp.1–9.

Cameron, R.A. et al., 2006. Unusual gene order and organization of the sea urchin hox cluster. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution*, 306B(1), pp.45–58.

Chesher, R.H., 1969. Destruction of Pacific corals by the sea star Acanthaster planci. *Science*, 165(3890), pp.280–283.

Chikhi, R. & Medvedev, P., 2014. Informed and automated k-mer size selection for genome assembly. *Bioinformatics (Oxford, England)*, 30(1), pp.31–37.

CRICK, F.H. et al., 1961. General nature of the genetic code for proteins. *Nature*, 192, pp.1227–1232.

Danecek, P. et al., 2011. The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, 27(15), pp.2156–2158.

Davidson, E.H., 2010. Emerging properties of animal gene regulatory networks. 468(7326), pp.911–920.

Davidson, E.H., 1997. Evolutionary biology. Insights from the echinoderms., 389(6652), pp.679–680.

De Robertis, E.M., 2008. Evo-Devo: Variations on Ancestral Themes. 132(2), pp.185–195.

De'ath, G. & Moran, P.J., 1998. Factors affecting the behaviour of crown-of-thorns starfish (Acanthaster planci L.) on the Great Barrier Reef:: 1: Patterns of activity. *Journal of experimental marine biology and ecology*.

De'ath, G. et al., 2012. From the Cover: The 27-year decline of coral cover on the Great Barrier Reef and its causes. *Proceedings of the National Academy of Sciences*, 109(44), pp.17995–17999.

Dixson, D.L., Abrego, D. & Hay, M.E., 2014. Reef ecology. Chemically mediated behavior of recruiting corals and fishes: a tipping point that may limit reef recovery. *Science*, 345(6199), pp.892–897.

Draetta, G. et al., 1989. Cdc2 protein kinase is complexed with both cyclin A and B: evidence for proteolytic inactivation of MPF., 56(5), pp.829–838.

Fabricius, K.E., Okaji, K. & De'ath, G., 2010. Three lines of evidence to link outbreaks of the crown-of-thorns seastar Acanthaster planci to the release of larval food limitation. *Coral Reefs*, 29(3), pp.593–605.

Fernandez-Valverde, S.L., Calcino, A.D. & Degnan, B.M., 2015. Deep developmental transcriptome sequencing uncovers numerous new genes and enhances gene annotation in the sponge Amphimedon queenslandica. *BMC Genomics*, 16(1), p.720.

Freeman, R. et al., 2012. Identical Genomic Organization of Two Hemichordate Hox Clusters. *Current Biology*, 22(21), pp.2053–2058. Available at: http://www.sciencedirect.com/science/article/pii/S0960982212010561.

Garcia-Alcalde, F. et al., 2012. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics (Oxford, England)*, 28(20), pp.2678–2679.

Gérard, K. et al., 2008. Assessment of three mitochondrial loci variability for the crown-of-thorns starfish: A first insight into Acanthaster phylogeography. *Comptes Rendus Biologies*, 331(2), pp.137–143.

Gillis, J.A., Fritzenwanker, J.H. & Lowe, C.J., 2011. A stem-deuterostome origin of the vertebrate pharyngeal transcriptional network. *Proceedings of the Royal Society B: Biological Sciences*, 279(1727), pp.237–246.

Gordon, A. & Hannon, G.J., 2010. *Fastx-toolkit*, FASTQ/A short-reads preprocessing tools (unpublished ….

Green, M.R. & Sambrook, J., 2012. Molecular cloning: a laboratory manual.

Haas, B. & Papanicolaou, A., 2012. Transdecoder.

Haas, B.J. et al., 2011. Approaches to Fungal Genome Annotation. *Mycology*, 2(3), pp.118–141.

Haas, B.J. et al., 2008. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, 9(1), p.R7.

Haas, B.J. et al., 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, 8(8), pp.1494–1512.

Hall, M.R. et al., 2017. The crown-of-thorns starfish genome as a guide for biocontrol of this coral reef pest. *Nature*, 544, pp231-234.

Haszpruner, G. & Spies, M., 2014. An integrative approach to the taxonomy of the crown-of-thorns starfish species group (Asteroidea: Acanthaster): A review of names and comparison to recent molecular data. *Zootaxa*, 3841(1), pp.271–284.

Henderson, J.A. & Lucas, J.S., 1971. Larval development and metamorphosis of Acanthaster planci (Asteroidea)., 232(5313), pp.655–657.

Hinman, V.F. et al., 2003. Developmental gene regulatory network architecture across 500 million years of echinoderm evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 100(23), pp.13356–13361.

Holland, L.Z., 2015. Evolution of basal deuterostome nervous systems. *Journal of Experimental Biology*, 218(4), pp.637–645.

Ihama, Y. et al., 2014. Anaphylactic shock caused by sting of crown-of-thorns starfish (Acanthaster planci). *Forensic Science International*, 236, pp.e5–e8.

Ikegami, S., Tamura, S. & Kanatani, H., 1967. Starfish gonad: action and chemical identification of spawning inhibitor. *Science*, 158(3804), pp.1052–1053.

Ikuta, T. et al., 2013. Identification of an intact ParaHox cluster with temporal colinearity but altered spatial colinearity in the hemichordate Ptychodera flava. *BMC Evolutionary Biology*, 13(1), p.129.

Irimia, M. et al., 2012. Extensive conservation of ancient microsynteny across metazoans due to cis-regulatory constraints. *Genome Research*, 22(12), pp.2356–2367.

Jaffe, L.A., 1993. Oocyte maturation in starfish is mediated by the beta gamma-subunit complex of a G-protein. *The Journal of Cell Biology*, 121(4), pp.775–783.

Kalachev, A.V., 2013. A brief summary of neuroendocrine regulation of reproduction in sea stars. *General and Comparative Endocrinology*, 183, pp.79–82.

Kanatani, H., 1964. Spawning of starfish: action of gamete-shedding substance obtained from radial nerves. *Science*, 146(3648), pp.1177–1179.

Katoh, M. & Hashimoto, K., 2003. Genetic similarity of outbreak populations of crown-of-thorns starfish (Acanthaster planci) that were 15�years apart in Okinawa, Japan. *Coral Reefs*, 22(2), pp.178–180.

Kent, W.J. et al., 2002. The Human Genome Browser at UCSC. *Genome Research*, 12(6), pp.996–1006.

Kettle, B.T. & Lucas, J.S., 1987. Biometric relationships between organ indices, fecundity, oxygen consumption and body size in Acanthaster planci (L.)(Echinodermata; Asteroidea). *Bulletin of marine science*, 41(2), pp.541–551.

Kirschner, M.W., Gerhart, J.C. & Norton, J., 2006. *The plausibility of life: Resolving Darwin's dilemma*, Yale University Press.

Kishimoto, T., 2011. A primer on meiotic resumption in starfish oocytes: The proposed signaling pathway triggered by maturation-inducing hormone. *Molecular Reproduction and Development*, 78(10-11), pp.704–707.

Kishimoto, T., 2015. Entry into mitosis: a solution to the decades-long enigma of MPF. *Chromosoma*, pp.1–12.

Kishimoto, T. & Kanatani, H., 1976. Cytoplasmic factor responsible for germinal vesicle breakdown and meiotic maturation in starfish oocyte., 260(5549), pp.321–322.

Kokubu, C. et al., 2009. A transposon-based chromosomal engineering method to survey a large cis-regulatory landscape in mice. *Nature Genetics*, 41(8), pp.946–952.

Komori, T., 1997. Toxins from the starfish Acanthaster planci and Asterina pectinifera. *Toxicon : official journal of the International Society on Toxinology*, 35(10), pp.1537–1548.

Le Novère, N. et al., 2009. The Systems Biology Graphical Notation. *Nature biotechnology*, 27(8), pp.735–741.

Lee, C.-C. et al., 2013. Hemolytic activity of venom from crown-of-thorns starfish Acanthaster planci spines. *The journal of venomous animals and toxins including tropical diseases*, 19(1), p.22.

Lee, C.-C. et al., 2014. Spine venom of crown-of-thorns starfish (Acanthaster planci) induces antiproliferation and apoptosis of human melanoma cells (A375.S2). *Toxicon : official journal of the International Society on Toxinology*.

Li, H. & Durbin, R., 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, 26(5), pp.589–595.

Li, H. & Durbin, R., 2011. Inference of human population history from individual whole-genome sequences. 475(7357), pp.493–496.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16), pp.2078–2079.

Lowe, C.J. et al., 2003. Anteroposterior patterning in hemichordates and the origins of the chordate nervous system., 113(7), pp.853–865.

Lowe, C.J. et al., 2006. Dorsoventral Patterning in Hemichordates: Insights into Early Chordate Evolution. *PLoS Biology*, 4(9), p.e291.

Lowe, C.J. et al., 2015. The deuterostome context of chordate origins. 520(7548), pp.456–465.

Lucas, J.S. & Jones, M.M., 1976. Hybrid crown-of-thorns starfish (Acanthaster planci X A. brevispinus) reared to maturity in the laboratory.

Luo, R. et al., 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, 1(1), p.18.

Luo, Y.-J., Satoh, N. & Endo, K., 2015. Mitochondrial gene order variation in the brachiopod Lingula anatina and its implications for mitochondrial evolution in lophotrochozoans. *Marine Genomics*, pp.1–11.

Mah, C.L. & Blake, D.B., 2012. Global diversity and phylogeny of the Asteroidea (Echinodermata). *PloS one*, 7(4), p.e35644.

Maoka, T. et al., 2010. Structure of minor carotenoids from the crown-of-thorns starfish, Acanthaster planci. *Journal of natural products*, 73(4), pp.675–678. Available at: http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=20180541&retmode=ref&cmd=prlinks.

Marcais, G. & Kingsford, C., 2011. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics (Oxford, England)*, 27(6), pp.764–770.

Martindale, M.Q. & Hejnol, A., 2009. A Developmental Perspective: Changes in the Position of the Blastopore during Bilaterian Evolution. *Developmental Cell*, 17(2), pp.162–174.

McClay, D.R., 2011. Evolutionary crossroads in developmental biology: sea urchins. *Development*, 138(13), pp.2639–2648.

Meselson, M. & Stahl, F.W., 1958. The replication of DNA in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 44(7), pp.671–682.

Miller, I. et al., 2015. Origins and Implications of a Primary Crown-of-Thorns Starfish Outbreak in the Southern Great Barrier Reef. *Journal of Marine Biology*, 2015(1), pp.1–10.

Mita, M., Yoshikuni, M. & Nagahama, Y., 1999. 1-Methyladenine production from ATP by starfish ovarian follicle cells. *Biochimica et biophysica acta*, 1428(1), pp.13–20.

Moore, R.J., 1978. Is Acanthaster planci an r-strategist?

Moore, R.J. & Huxley, C.J., 1976. Aversive behaviour of crown-of-thorns starfish to coral evoked by food-related chemicals., 263(5576), pp.407–409.

Moran, P.J., 1990. Acanthaster planci (L.): biographical data. *Coral Reefs*.

Moran, P.J., 1988. The Acanthaster phenomenon. 7.

Moran, P.J. & De'ath, G., 1992. Estimates of the abundance of the crown-of-throns starfish Acanthaster planci in outbreaking and non-outbreaking populations on reefs within the Great Barrier Reef. *Marine Biology*, 113(3), pp.509–515.

Nagarajan, N. & Pop, M., 2013. Sequence assembly demystified. *Nature Reviews Genetics*, 14(3), pp.157–167.

Nakamura, M. et al., 2014. Spatial and temporal population dynamics of the crown-of-thorns starfish, Acanthaster planci, over a 24-year period along the central west coast of Okinawa Island, Japan. *Marine Biology*, 161(11), pp.2521–2530.

Nash, W.J., Goddard, M. & Lucas, J.S., 1988. Population genetic studies of the crown-of-thorns starfish, Acanthaster planci (L.), in the Great Barrier Reef region. *Coral Reefs*, 7(1), pp.11–18.

Nishida, M. & Lucas, J.S., 1988. Genetic differences between geographic populations of the crown-of-thorns starfish throughout the Pacific region. *Marine Biology*, 98(3), pp.359–368.

Ogasawara, M. et al., 1999. Developmental expression of Pax1/9 genes in urochordate and hemichordate gills: insight into function and evolution of the pharyngeal epithelium. *Development*, 126(11), pp.2539–2550.

Okumura, E. et al., 2002. Akt inhibits Myt1 in the signalling pathway that leads to meiotic G2/M-phase transition. *Nature Cell Biology*, 4(2), pp.111–116.

Ormond, R. et al., 1973. Formation and breakdown of aggregations of the crown-of-thorns starfish, Acanthaster planci (L.).

Parra, G., Bradnam, K. & Korf, I., 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics (Oxford, England)*, 23(9), pp.1061–1067.

Pearson, R.G., 1972. Changes in distribution of Acanthaster planci populations on the Great Barrier Reef. 237, pp.175–176.

Peter, I. & Davidson, E.H., 2015. *Genomic control process: development and evolution*, Academic Press.

Pisani, D. et al., 2012. Resolving phylogenetic signal from noise when divergence is rapid: A new look at the old problem of echinoderm class relationships. *Molecular Phylogenetics and Evolution*, 62(1), pp.27–34.

Romiguier, J. et al., 2014. Comparative population genomics in animals uncovers the determinants of genetic diversity. 515(7526), pp.261–263.

Santagati, F. et al., 2003. Identification of Cis-regulatory elements in the mouse Pax9/Nkx2-9 genomic region: implication for evolutionary conserved synteny. *Genetics*, 165(1), pp.235–242.

Sapp, J., 1999. *What Is Natural?* Oxford University Press.

Satoh, N., 2016. *Chordate Origins and Evolution: The Molecular Evolutionary Road to Vertebrates*, Academic Press.

Satoh, N., Rokhsar, D. & Nishikawa, T., 2014. Chordate evolution and the three-phylum system. *Proceedings of the Royal Society B: Biological Sciences*, 281(1794).

Schiffels, S. & Durbin, R., 2014. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8), pp.919–925.

Schmieder, R. & Edwards, R., 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics (Oxford, England)*, 27(6), pp.863–864.

Sea Urchin Genome Sequencing Consortium et al., 2006. The genome of the sea urchin Strongylocentrotus purpuratus. *Science*, 314(5801), pp.941–952.

Shinzato, C. et al., 2011. Using the Acropora digitifera genome to understand coral responses to environmental change. *Nature*, 476(7360), pp.320–323.

Shirai, H., Kanatani, H. & Taguchi, S., 1972. 1-methyladenine biosynthesis in starfish ovary: action of gonad-stimulating hormone in methylation. *Science*, 175(4028), pp.1366–1368.

Shoguchi, E. et al., 2013. Draft assembly of the Symbiodinium minutum nuclear genome reveals dinoflagellate gene structure. *Current biology : CB*, 23(15), pp.1399–1408.

Sievers, F. et al., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, 7(1), pp.539–539.

Simakov, O. et al., 2015. Hemichordate genomes and deuterostome origins. *Nature*, 527(7579), pp.459–465.

Simakov, O. et al., 2012. Insights into bilaterian evolution from three spiralian genomes. 493(7433), pp.526–531.

Simão, F.A. et al., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics (Oxford, England)*, 31(19), pp.3210–3212.

Smit, A., Hubley, R. & Green, P., 2016. RepeatMasker Open-3.0. 1996. *Source: http://www. repeatmasker. org/faq. html# faq3*.

Stanke, M. et al., 2006. AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic acids research*, 34(Web Server), pp.W435–W439.

Tamura, K. et al., 2013. MEGA6: Molecular Evolutionary Genetics Analysis Version 6.0. *Molecular Biology and Evolution*, 30(12), pp.2725–2729.

Timmers, M.A. et al., 2012. There's no place like home: crown-of-thorns outbreaks in the central pacific are regionally derived and independent events. *PloS one*, 7(2), p.e31159.

Trapnell, C. et al., 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3), pp.562–578.

Uthicke, S. et al., 2015. Climate change as an unexpected co-factor promoting coral eating seastar (Acanthaster planci) outbreaks. *Scientific Reports*, 5, p.8402.

Vogler, C. et al., 2008. A threat to coral reefs multiplied? Four species of crown-of-thorns starfish. *Biology Letters*, 4(6), pp.696–699.

Vogler, C. et al., 2012. Phylogeography of the crown-of-thorns starfish in the Indian Ocean. *PloS one*, 7(8), p.e43499.

Wada, H. & Satoh, N., 1994. Phylogenetic relationships among extant classes of echinoderms, as inferred from sequences of 18S rDNA, coincide with relationships deduced from the fossil record. *Journal of molecular evolution*, 38(1), pp.41–49.

Walbran, P.D., Henderson, R.A., Faithful, J.W., et al., 1989. Crown-of-thorns starfish outbreaks on the Great Barrier Reef: a geological perspective based upon the sediment record. *Coral Reefs*, 8(2), pp.67–78.

Walbran, P.D., Henderson, R.A., Jull, A.J., et al., 1989. Evidence from Sediments of Long-Term Acanthaster planci Predation on Corals of the Great Barrier Reef. *Science*, 245(4920), pp.847–850.

Wang, S. et al., 2009. Up-regulation of C/EBP by thyroid hormones: A case demonstrating the vertebrate-like thyroid hormone signaling pathway in amphioxus. *Molecular and Cellular Endocrinology*, 313(1-2), pp.57–63.

Wang, W. et al., 2006. Comparison of Pax1/9 Locus Reveals 500-Myr-Old Syntenic Block and Evolutionary Conserved Noncoding Regions. *Molecular Biology and Evolution*, 24(3), pp.784–791.

Watson, J.D. & Crick, F., 1953. *Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid*, Nature.

Wolfe, K. et al., 2015. Larval Starvation to Satiation: Influence of Nutrient Regime on the Success of Acanthaster planci S. C. A. Ferse, ed. *PloS one*, 10(3), p.e0122010.

Yamaguchi, M., 1986. Acanthaster planci infestations of reefs and coral assemblages in Japan: a retrospective analysis of control efforts. *Coral Reefs*, 5(1), pp.23–30.

Yasuda, N. et al., 2006. Complete mitochondrial genome sequences for Crown-of-thorns starfish Acanthaster planci and Acanthaster brevispinus. *BMC Genomics*, 7, p.17.

Yasuda, N. et al., 2009. Gene flow of Acanthaster planci (L.) in relation to ocean currents revealed by microsatellite analysis. *Molecular ecology*, 18(8), pp.1574–1590.

Yasuda, N. et al., 2014. Genetic connectivity of the coral-eating sea star Acanthaster planciduring the severe outbreak of 2006-2009 in the Society Islands, French Polynesia. *Marine Ecology*, pp.n/a–n/a.

Zerbino, D.R. & Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5), pp.821–829.

Zhang, M. et al., 2012. Preparation of megabase-sized DNA from a variety of organisms using the nuclei method for advanced genomics research. *Nature Protocols*, 7(3), pp.467–478.