COMPARATIVE GENOMICS AND TRANSCRIPTOMICS OF THE 'IMMORTAL JELLYFISH' (*Turritopsis dohrnii*) FROM BOCAS DEL TORO, PANAMA (ATLANTIC)

A Dissertation

by

YUI MATSUMOTO

Submitted to the Graduate and Professional School of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Maria Pia Miglietta
Committee Members,	Jaime Alvarado-Bremer
	Noushin Ghaffari
	Jessica Labonte
	Anja Schulze
Head of Department,	Daniel Roelke

May 2022

Major Subject: Marine Biology

ABSTRACT

This dissertation introduces a new research system, *Turritopsis dohrnii* (Cnidaria, Hydrozoa), that can further our understanding of tissue regeneration, cellular plasticity and aging using a genomic approach. When exposed to physical damage, adverse environmental conditions, or senescence, medusae of *T. dohrnii* will transform into a cyst stage, which then reverts back to the juvenile polyp. The underlying mechanism that enables its ontogenetic reversal is called cell transdifferentiation (i.e., cell reprogramming), a process in which mature, fully differentiated cells can switch into another needed cell type of any lineage. The following interconnected projects are presented: I) Transcriptome assembly/annotation of *T. dohrnii* stages involved in reverse development and bioinformaticss analysis to identify genes and networks that underlie its transdifferentiation and ontogeny reversal; II) Expression profiling of genetic networks involved in mammalian longevity, regeneration, and pluripotency, such as Sirtuins, Heat-shock Proteins, and POU domain factors, and subsequent gene-tree analyses to investigate the evolutionary history of such factors in *T. dohrnii*; III) Construct a draft genome for *T. dohrnii* through generating and assembling genomic reads and subsequent scaffolding of the genome using RNA-sequencing libraries and transcriptome.

Genetic constituents and networks associated to aging/lifespan, transposable element regulation, and response to DNA damage, were identified to be highly active in cyst stage, where cell transdifferentiation occurs. The reversed polyp developed from

transdifferentiation processes show marked differences in its transcriptional profile compared to colonial polyps produced from budding, such as heightened activity of genes associated to chromatin remodeling, matrix metalloproteinases, and embryonic development. Furthermore, profiling analyses revealed that *T. dohrnii* highly regulates genes imperative to the interconnected networks of regeneration, pluripotency, and longevity in mammals during its ontogeny reversal sequence, such as SIRT3, HSP90, and POU factors, and homology-based and phylogenetic analyses support the presence of Yamanaka factor homologs in *T. dohrnii,* warranting further exploration and advanced orthology analyses. Lastly, the generation of the first draft genome for *T. dohrnii* has provided insight into the complexity of the assembly of its genome and represents an essential initial step in developing the species as an unparalleled research system to investigate the genetics of cellular plasticity and reprogramming, *in vivo.*

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my graduate adviser and committee chair, Dr. Maria Pia Miglietta, for providing valuable guidance throughout the duration of my Ph.D program. Her patience, enthusiasm and knowledge has been fundamental to my deep understanding of genomics and the advancement of my academic and research career. Furthermore, Dr. Miglietta's investment in the success of my research and graduate career has facilitated my further development as a genomic researcher.

Additionally, I would like to thank the rest of my research committee, Dr. Jaime Alvarado-Bremer, Dr. Noushin Ghaffari, Dr. Anja Schulze, Dr. Jessica Labonte for the encouragement and support they provided throughout the progression of my research. Without their guidance, the obstacles encounter during my projects would have been much harder to overcome. Additionally, I would like to thank my current and former labmates in the Miglietta lab, Sarah Pruski, Alex Frolova, Kade Muffet, the entire TAMUG MARB graduate department, and collaborator Dr. Stefano Piraino for the encouragement and support they provided throughout the progression of my research. Without their guidance, my project would have failed to incorporate different perspectives which drastically expanded the scope of my project. I would also like to thank my undergraduates Delaney Quinn and Ashley Whitehead for being great assistants for my lab work and data organization.

Furthermore, I would like to thank Dr. Jessie Lie from the UC Davis Bioinformaticss Core for aiding me with data analyses and helping me gain valuable bioinformaticss skills, and Dr. Michael Dickens from the Texas A&M High Performance Researching Cluster (HPRC) for aiding me by downloading necessary software on the HPRC clusters to complete my dissertation projects and helping me identify and fix errors in my scripts. Lastly, sample collection would not have been possible without the research facilities of the Smithsonian Tropical Research Institute and support from the members of the Systematics and Biology of Hydrozoa Summer Course (2015) who invested long hours helping me collect specimens in Bocas del Toro, Panama.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a Ph.D. dissertation research committee consisting of Professor Maria Pia Miglietta [adviser], Jaime Alvarado-Bremer, Dr. Jessica Labonte and Anja Schulze of the Department of Marine Biology in Texas A&M University in Galveston (TAMUG), and Dr. Noushin Ghaffari of the Computer Science Department at Prairie View A&M University (PVAMU).

All work for this dissertation was completed by the student, under the advisement of Dr. Maria Pia Miglietta of the Department of Marine Biology.

Funding Sources

This work was made possible by the NSF ARTS (grant № DEB-1456501) and NSF EAGER (grant № 1936565). Additionally, attending conferences to present my work was made possible in part by TAMUG's Marine Biology Graduate Student Mini Grants, Galveston Graduate Student Association Travel Grant, and Erma Lee and Luke Mooney Graduate Student Travel Grant.

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

CHAPTER I

INTRODUCTION

I.1. Phylum Cnidaria, Class Hydrozoa

Cnidaria (jellyfish, polyps, sea anemones, corals) is a fascinating phylum to further understand the evolution of development in animals, as they are the sister taxon to Bilateria and therefore, a critical phylogenetic link to protostome and deuterostome divergence (Putnam et al. 2007; Watanabe et al. 2009; DuBuc et al. 2014). Many cnidarians have high regenerative capacities (Kortschak et al. 2003; Wenger and Galliot 2013; Tomczyk et al. 2015) and as a consequence, can have very long life-spans (Martínez and Bridge 2012; Vaupel et al. 2004; Prouty et al. 2011; Reiter et al. 2012). Additionally, they possess high conservation in genomic content as compared to vertebrates (Spring et al. 2000; Gauchat et al. 2000; Miller et al. 2000), making them potentially useful to investigate the development and evolution of animals.

Within the phylum Cnidaria, many hydrozoans (class Hydrozoa) undergo drastic changes in body plan physique and behavior throughout their life cycle, referred to as metamorphosis (e.g., polyp-to-medusa transition), thus exhibiting a diverse array of complex life cycles. In a genomic perspective, this implies that a single genome has the capacity and flexibility to encode for a variety of developmental forms during its ontogenetic sequence. In the typical hydromedusan, a hydrozoan with a medusa (jellyfish) stage, its life cycle starts with the planula larvae which settles on a substrate and develops into a juvenile polyp (Figure I.1, indicated in blue arrows). The polyp

asexually propagates and develops into a larger colony which bud free-swimming medusae. Sexually mature medusae will release gametes into the water column, and larvae are produced through external fertilization. Planulae settle and develop into a juvenile polyp, closing the life cycle. The hydrozoan *Turritopsis dohrnii* (Filifera, Family Oceaniidae), is particularly an interesting taxon to investigate metamorphosis and the flexibility of life cycles in metazoans, as it exhibits an additional developmental sequence that can provide a new paradigm for studies in tissue regeneration, cell plasticity and aging (i.e., biological senescence).

I.2. Reverse development and transdifferentiation in *Turritopsis dohrnii*

The medusae of *T. dohrnii* can avoid death caused by physical damage, adverse environmental condition or aging by undergoing reverse development, also known as life-cycle reversal. A senescent or weakened medusa will reverse metamorphose into a juvenile polyp (Figure I.1, indicated in red arrows). During the metamorphosis, the medusa first settles onto a substrate and transforms into a cyst like structure mainly composed of uncharacterized cells and tissue with a chitinous exterior, the perisarc (Piraino et al. 1996; Schmich et al. 2007; Matsumoto et al. 2019). The cyst then rejuvenates back into a juvenile polyp and is re-introduced to its ontogenetic sequence. The unique ability to continuously extend its lifecycle has granted *T. dohrnii* the name, 'Immortal Jellyfish'.

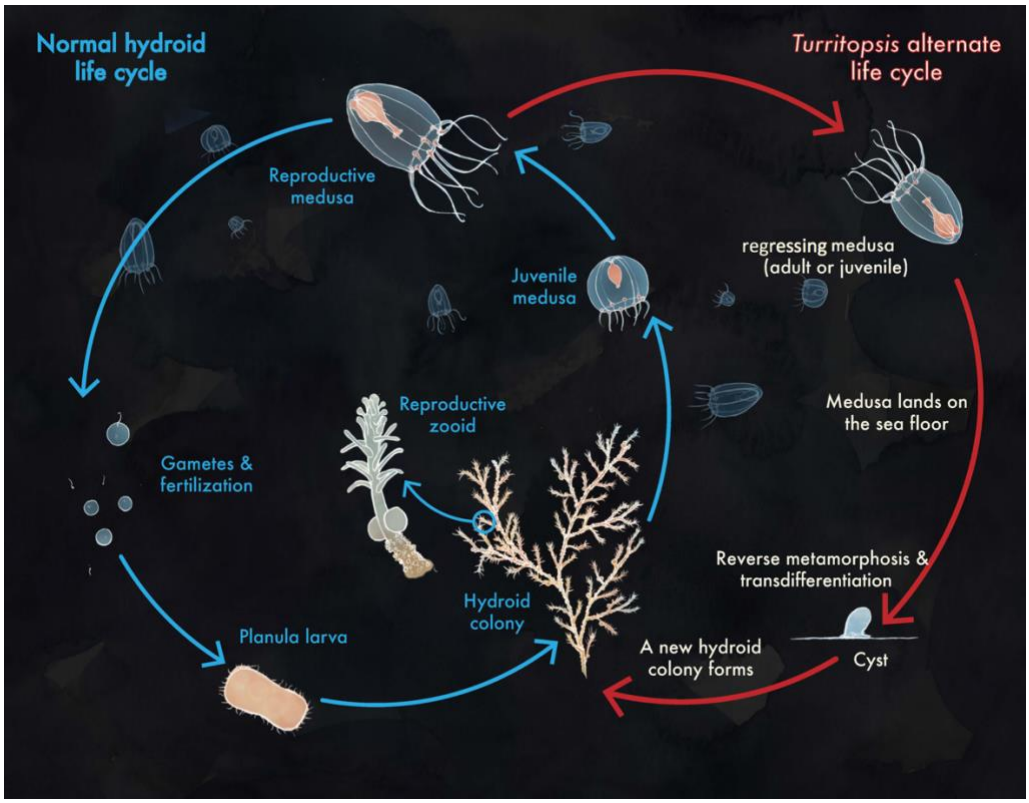


Figure I.1: Life cycle of *Turritopsis dohrnii*. A typical hydromedusan lifecycle is indicated in blue arrows (left), while the alternate life cycle of *T. dohrnii* is indicated in red arrows (right) (Matsumoto et al., 2019).

Specifically, during the reverse development, the medusa (Figure I.2A) will settle on a substrate and reduce its tentacles to form a bubble-like structure where the ring canal and tentacle bulbs fuse together, and decreases to about half the size of its original form (Figure I.2B) (Schmich et al. 2007). The bubble-like structure then further reduces, and the manubrium also fuses to the ring canal and tentacle bulbs (Figure I.2C). The medusa structures then are completely reduced and turn into an oval structure (Figure I.2D), which eventually forms into a cyst with a perisarc that attaches to the substrate (Figure I.2E). Stolon tips will start to form (Figure I.2F), develop into hydrorhizal

stolons (Figure I.2G), then start to reduce the cyst to leave behind an empty perisarc while continuing to elongate stolons (Figure I.2H). Tentacle buds start to form (Figure I.2I) and finally, a functional reversed polyp develops (Figure I.2J). The cyst stage has been reported to retain its ability to rejuvenate back into the polyp stage for up to 3 months under stressful but biologically manageable conditions (i.e., cold temperature) (Piraino et al. 1996). Reverse development can occur anytime during the medusa stage, from newborn to post-reproductive (biologically senescent).

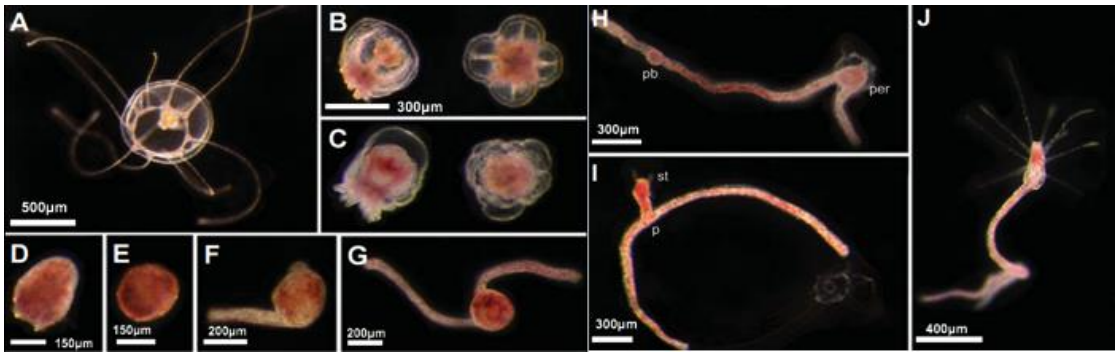


Figure I.2: Reverse development of *T. dohrnii* (Schmich et al. 2007). A) Newborn medusa stage; B) Reduced medusa stage; C) Bubble-stage; D) Oval stage; E) Cyst stage; F) Cyst with short stolon; G) Cyst with elongated stolons; H) Empty perisarc stage with long stolons; I) Bud stage; J) Reversed polyp.

Life-cycle reversal has been reported in other *Turritopsis* species and hydrozoans, such as *Turritopsis* sp.2 (Kubota 2005, 2011), sp.5 (Li et al. 2018), *Laodicea undulata* (De Vito et al. 2006; Schmich et al. 2007), and recently for the first time in a scyphozoan, *Aurelia* sp.1 (He et al. 2015). These species, however, are limiting as model systems for regeneration and reverse development as their reversal success rate and time is highly variable, and evidence of ontogeny reversal have only been reported

in laboratory settings. In comparison, the success rate of life-cycle reversal is much higher in both the adult and newborn medusae of *T. dohrnii* as compared to other *Turritopsis* species (Li et al. 2018; Miglietta et al. 2018). Additionally, *T. dohrnii* has been reported to be carried in ship ballast water and naturally undergo life-cycle reversal, silently invading foreign coastal waters and becoming a prolific invasive species (Miglietta and Lessios 2009; Miglietta et al. 2018; Matsumoto et al. 2019). Thus, *T. dohrnii* appears to be a promising candidate for investigating processes such as cellular transdifferentiation and plasticity associated with its reverse development.

In laboratory settings, reverse development of *T. dohrnii* can be easily induced by starvation, mechanical damage, heat shock (both increase and decrease in temperature), changes in salinity or cesium chloride exposure (Piraino et al. 1996; Piraino et al. 2004; Schmich et al. 2007). Tissue excision experiments on the medusae of *T. dohrnii* have provided insight on the cellular mechanism that underlies its reverse development. The manubrium of a medusa contains a large population of interstitial cells (I-cells), cnidarian stem cells that play a large role in cnidarian regeneration (Tardent 1963). On the other hand, the exumbrellar epidermis does not contain such I-cells. When the two types of tissue were excised and separated, the manubrium with the I-cells were not able to transform into perisarc-secreting tissue, the chitinous exterior that protects the soft tissue in the cyst and the polyp stage, but absent in the medusa (Piraino et al. 1996). The exumbrellar epidermis however, transformed into perisarc-secreting tissue, and eventually reversed back into a polyp. This signifies that *T. dohrnii* cells undergo cellular transdifferentiation (cell reprogramming) during reverse development.

Transdifferentiation is a mechanism in which a fully mature and specialized cell can switch into a new cell type (Okada 1991). The process primarily occurs in two steps: the cells de-activate their current developmental program and de-differentiate. Then a new developmental program is activated, re-programming the cell into a new (needed) type. Therefore, cell transdifferentiation is highly regarded in the biomedical sector as a potential mechanism to transform mature cells into any needed cell type after tissue loss or damage without the need of stem and progenitor cells (Collas and Håkelien 2003). Such reprogramming takes place during the medusa-to-polyp rejuvenation, where weakened medusae revert to the juvenile polyp stage. Overall, the medusa-to-polyp transformation in *T. dohrnii* displays dynamic ontogenetic events, in which a combination of transdifferentiation, cell proliferation, and apoptotic events occur to rejuvenate into an earlier life-cycle stage.

Understanding the genetic basis of transdifferentiation has significant implications in tissue regeneration and stem cell research in metazoans. Much of the effort in regenerative research has been conducted *in vitro* (outside of a living organism) with isolated cells on a petri dish/test tube. Although *in vitro* approaches are useful to identify key genetic components of specific cellular mechanisms, it disregards the true complexity of natural transdifferentiation, as removing a cell from its natural environment has commonly induced changes in phenotype, cellular stability and behavior (Holtfreter 1946; Lorian 1989; Alvarado and Yamanaka 2014). Thus, it is essential to have a model system that can address cell transdifferentiation *in vivo*, where the cells are assessed in their natural environment within a living organism. *T. dohrnii*

embodies a relatively simple system to investigate transdifferentiation *in vivo* and examine the underlying mechanisms by which cells spontaneously leave a specialized state and develop into a new cell type. In addition, *T. dohrnii* possess sister species (e.g. *T. nutricula, T.* sp.1) that do not have the ability to undergo reverse development (Miglietta et al. 2018). This notable characteristic allows for notable comparative analyses that can shed light on how cellular reprogramming and reverse development can occur in one species (i.e., *T. dohrnii*), but not the other (e.g., *T. nutricula, T.* sp.1). Further deciphering the genetic basis that underlies cell transdifferentiation in *T. dohrnii* will have significant impacts on sectors that explore evolutionary development, tissue regeneration, cell plasticity and biological aging/senescence in metazoans.

I.3. Current knowledge: Transcriptomics of *Turritopsis dohrnii*

Transcriptomics, the study of gene expression profiles, is a popular approach to investigate how gene activity is influenced by development, different environmental conditions, or the progression of disease. Gene expression landscapes can be explored using a broad-ranged, high-throughput technique called RNA-sequencing (RNA-seq). RNA-seq enables the quantification of RNA transcripts, including non-coding RNA, to create a snapshot of the genes expressed at a specific stage. Transcriptomes can offer advantages over genome (entire genetic profile) sequencing when the experimental design focuses on the changes of gene expression and activity (i.e., transcription and translation into protein), as opposed to gene structure. However, despite the complexities assembling a genome may bring due to the necessity to decipher sequence repeats, spacer sequences and non-coding regions (Lynch and Conery 2003; Lynch and Walsh

2007; Pop 2009; Gurevich et al. 2013), it is an essential step to cataloging the complete set of genes found within a species to conduct widespread evolutionary and functional comparison among other taxa.

Three separate transcriptomes from the polyp colony and medusa from the Mediterranean Sea in Italy, and the cyst from the Atlantic coast of Bocas del Toro, Panama, have been produced and annotated (Matsumoto et al, 2019). A preliminary comparative functional gene enrichment analyses of the three stages, with the cyst as the central stage of comparison, reported differences in biological processes among the stages. Overall, the results reported a substantially larger portion of under-expressed biological processes in the cyst as compared to the other stages (Matsumoto et al., 2019). The results were attributed to the uncharacteristic state of the cyst in which morphological structures are highly reduced. Gene Ontology (GO) categories involving lifespan and aging, cellular characterization, division, differentiation, and development, and response to stimuli were reported to be under-expressed in the Cyst as compared to the polyp and medusa. On the other hand, DNA synthesis and metabolic processes, telomere maintenance, DNA repair and integration, and transposition categories were reported to be over-expressed in the cyst as compared to the polyp. Further experiments and quantification of the genes and transcripts are needed to further confirm significant GO processes and determine their role, if any, in the cell transdifferentiation and reverse development of *T. dohrnii.*

I.4. Dissertation chapter overview

The ultimate goal of this dissertation is to develop *T. dohrnii* into an alternate research system by producing genomic tools to further understand evolution, reverse development, regeneration, cellular pluripotency and aging in metazoans. The remainder of the dissertation is organized into the following four chapters.

Chapter II (Project 1) generates an annotated transcriptome assembly of *T. dohrnii'*s life cycle stages involved in reverse development. Differential gene expression (DGE) analyses provide insight on the genes that are differentially expressed (DE) during life cycle reversal in *T. dohrnii*, with a focus on the cyst stage where cell transdifferentiation occurs, along with supplementary analyses that distinguish transcriptional differences between the medusa and colonial polyp, and the polyp stages that differ in developmental trajectories (asexual budding vs. cell transdifferentiation).

Chapter III (Project 2) scaffolds previously assembled transcripts (Project 1) into super-transcripts to best represent genes. The activity of genes recognized as fundamental genetic modulators of regenerative, pluripotency and longevity pathways, namely Sirtuin proteins (SIRT), telomerase and telomere maintenance-related genes, heat-shock proteins (HSPs), and the Yamanaka Factor gene family (Oct, Sox, Klf, and Myc) were profiled during *T. dohrnii's* reverse development to provide insight on how the species manipulates genetic networks of high relevance in biomedical studies among mammals. Additionally, the evolutionary history of highly-expressed genes of interest were inferred through gene tree analyses.

Chapter IV (Project 3) produces the first draft genome assembly of *T. dohrnii*, assembling newly generated Pacific Biosciences (PacBio) reads and scaffolding the genome using the previously generated RNA-seq libraries and assembled transcripts (Project 1). The chapter provides new insight into the complexity of genome construction for *T. dohrnii* and produced draft genomes delivers an essential foundation to furthering the species as a non-traditional research system.

Lastly, Chapter V will present the concluding synthesis of the three combined projects, which will include research findings, implications, and future direction.

CHAPTER II

CELLULAR REPROGRAMMING AND IMMORTALITY: EXPRESSION PROFILING REVEALS PUTATIVE GENES INVOLVED IN *Turritopsis dohrnii*'s LIFE CYLE REVERSAL[*]

II.1. Introduction

The ultimate goal of regenerative research is to replace damaged cells in response to injuries and aging (Shenoy and Blelloch 2012). Transdifferentiation (or cell reprogramming), a process through which a mature somatic cell transforms into a new type of mature somatic cell (Jopling et al. 2011), can achieve this goal. *In vitro* reprogramming of somatic cells has proven to be a powerful tool and has allowed the identification of some of the genetic factors required to change identity (Motoyama et al. 2009; Dobrovolskaia et al. 2003; Limbert et al. 2011). However, the mechanisms and molecular drivers by which cells spontaneously (*in vivo*) leave a differentiated state to become a new lineage are poorly understood (Merrell and Stanger 2016), in large part, because of the difficulties of inducing transdifferentiation in live model systems (Alvarado and Yamanaka 2014; Abad et al. 2013). Additionally, because transdifferentiation in available model systems takes place over weeks, modeling its underlying gene regulatory network is problematic (Kaity et al. 2018). Therefore, it is

necessary to identify non-traditional species with relevant life history and physiological traits suited for the investigation of genetic mechanism(s) of cellular stability and plasticity *in vivo*.

The discovery of reverse development in the cnidarian *Turritopsis dohrnii* (Class Hydrozoa) represents a promising new research system (Bavestrello et al. 1992; Piraino et al. 1996; Miglietta and Lessios 2009). Faced with unfavorable circumstances, the medusa naturally undergoes cellular reprogramming to revert to a younger life cycle stage (the polyp), thus avoiding death indefinitely. While the majority of hydromedusae become reproductively mature, release gametes, and die, medusae of *T. dohrnii* that are stressed, damaged, or senescent settle on a surface and transform into a cyst stage that, in 24-72 hours, metamorphoses back into a single juvenile polyp (Piraino et al. 1996; Schmich et al. 2007; Miglietta et al. 2018; Matsumoto et al. 2019). By asexual reproduction, the polyp can develop into a larger colony (i.e., colonial polyp) that can then release new medusae. Because of its unique life cycle, *T. dohrnii* has been popularized as the 'Immortal Jellyfish'.

T. dohrnii is also heavily understudied, with fewer than a dozen papers published since the discovery of its life cycle (Alvarado and Yamanaka 2014). This is mostly due to the difficulties in collecting *T. dohrnii* in the field and in handling the species in the laboratory. However, with its unique potential for rejuvenation and the ability to induce cellular reprogramming under controlled laboratory conditions in approximately 24 hours (Piraino et al. 1996; Matsumoto et al. 2019), *T. dohrnii* represents a promising albeit unexplored system to investigate molecular mechanisms of cell stability and

regeneration that may be relevant to the development of other animals, including humans.

Recently, a comparison of three individually assembled *de novo* transcriptomes of the colonial polyp, cyst, and medusa of *T. dohrnii* has provided preliminary insight on the differences in the number of transcripts (i.e. raw count) annotated with specific Gene Ontology (GO) terms (Matsumoto et al. 2019). Transcripts associated with telomere organization and maintenance, DNA integration, repair and damage response were among those that showed high expression in the cyst relative to medusa and polyp. Processes associated with cell signaling, division, differentiation, and development showed low expression in the cyst relative to the medusa and polyp. However, the three sequenced libraries (polyp, cyst, and medusa) were not assembled into a single annotated transcriptome because the stages originated from different sampling locations, lacked biological replicates, and were sequenced using different platforms. As a result, differential gene expression (DGE) analyses of the life cycle stages were not conducted, and changes in expression levels between the different stages could not be quantified. Matsumoto et al. (2019) also did not sample reversed polyps (i.e., polyps that originate from the medusa through reverse development). Finally, differences in gene activity between the colonial polyp and medusa stage, and polyps developed from different developmental processes, namely asexual budding and reverse development, have yet to be explored.

We investigate the sequential changes in gene expression that occur during the reverse development of *T. dohrnii* (from colonial polyp to medusa, to cyst, to reversed

polyp), with an emphasis on genes that are upregulated in the cyst, where cellular transdifferentiation occurs. We also conduct a pair-wise DGE analysis to compare the transcriptional profiles of the benthic colonial polyp and planktonic medusa, and those of the colonial polyp, developed in the wild through asexual budding within a larger colony, and the reversed polyp, developed from the cyst through cellular reprogramming. These analyses reveal differences in gene activity between ecologically and morphologically different stages (benthic and colonial polyp vs. planktonic and solitary medusa) and between morphologically similar stages (polyps) generated by two very different developmental pathways, asexual budding and reverse development.

In summary, this research aims to clarify the genetic pathways involved in the reverse development and cell transdifferentiation of *T. dohrnii* through sequential and pair-wise transcriptomic comparison of life cycles stages involved in both forward and reverse development. These genomic tools will further the potential of *T. dohrnii* as a system for the study of the mechanisms and molecular drivers by which cells spontaneously leave a differentiated state to become a new lineage.

II.2. Results

Transcriptome Assembly and Characterization

The Metazoa BUSCO (Simão et al. 2015) analysis reported 97.9% completeness (95.6% complete, 2.2% partial), indicating that our initial transcriptome assembly is highly complete in terms of gene content (Appendix II.A). The final transcriptome assembly (~265.685 Mbp) resulted in 204,031 transcripts and 127,645 unigenes with a

GC content of 38.29% (Appendix II.B). The N50 of the transcriptome is 1,734 bp with a median contig length of 832 bp and an average length of 1,258.07 bp. Approximately 80% of the transcriptome (162,010 out of 204,031 contigs) was annotated with the following annotation pipelines: B2G (NCBI Non-Redundant), InterProScan, COG/EggNOG, KEGG, Rfam, and the Hydrozoa EST; and GO terms were found among ~42% (85,960 out of 204,031) of contigs of the transcriptome (Appendix II.C). Most of the top BLAST hits belonged to six cnidarian species, *Hydra vulgaris*, *Exaiptasia pallida*, *Stylophora pistillata*, *Acropora digitifera*, *Orbicella faveolata* and *Nematostella vectensis*, ranked 1st to 6th, accounting for ~52% (59,859 out of 116,110) of total hits (Figure II.1).

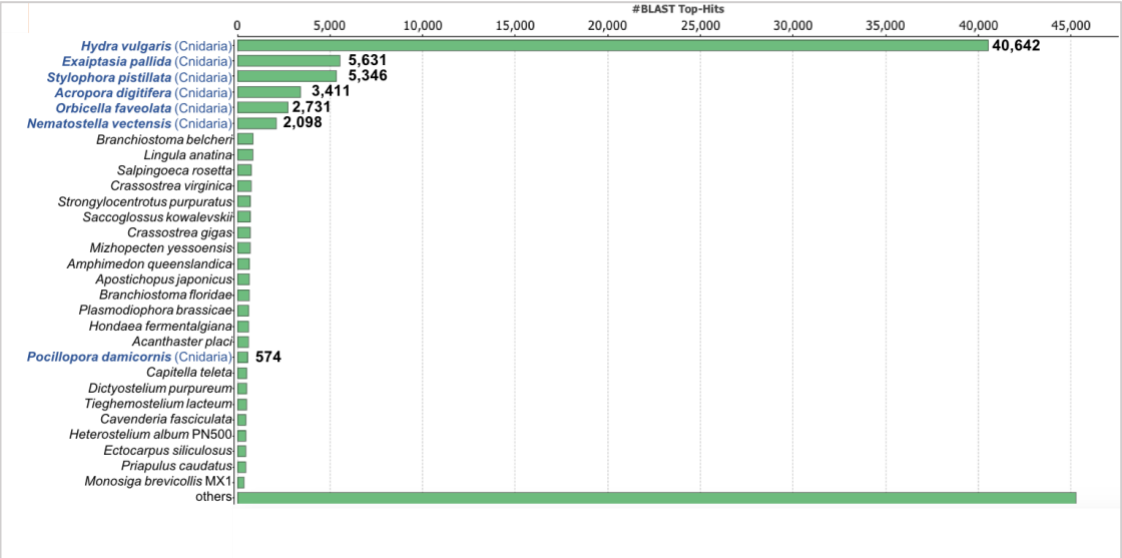


Figure II.1: Species distribution of top BLASTx hits in final assembly. [Total contigs with BLASTx hit(s): 116,110 transcripts; Blue font: Cnidarian taxa]

DGE analysis of Life Cycle Stages Involved in Reverse Development

A DGE analysis of the life cycle stages involved in reverse development (Nueda et al. 2014) was performed in the following sequence: 1) Colonial Polyp, 2) Medusa, 3) Cyst, and 4) Reversed Polyp. This sequence reflects the natural ontogenetic trajectory observed in *T. dohrnii* during its life cycle reversal. Overall dataset quality was verified with a multidimensional scaling plot (Appendix II.D). A total of 7,003 differentially expressed genes (DEGs) were identified, where 1,585 DEGs had R^2 values larger than 0.7 and considered significant for analyses with less than five biological replicates per stage (as recommended by Nueda et al. (2014)) (Figure II.2A). Nine different gene expression profiles (i.e., clusters) were formed through hierarchical clustering, including 50-239 genes per group (Figure II.2B; expression profiles in Appendix II.E). Among the nine clusters, Cluster 5 with 224 genes, represented genes that were exclusive and/or associated to the Cyst stage, reporting a statistically significant increase and subsequent decrease in gene activity during the Medusa to Reversed Polyp transition (Figure II.2B (Green) and 2C). Cluster 5 showed a higher number of novel genes (i.e., genes with no annotation) (26.8%) compared to the overall transcriptome (20.7%), particularly in the top 50 (44%) and 100 (38%) most significant DEGs (Table 1A). The 44% among the top 50 DEGs being novel is a striking comparison to the 20.7% of unannotated sequences in the overall transcriptome. The most significant DEG among the 224 genes was a gene with no annotation (Td_DN103197_c0_g1). It had no evidence of any expression in all *T. dohrnii* samples except for the Cyst replicates (Figure II.3A). Additionally, there were

other unannotated genes that showed a significant peak at the Cyst, such as gene Td_DN92408_c0_g2 (Figure II.3B).

A.

	# genes
Total DE genes	7,003
Total signficant DE genes (R-squared > 0.7)	1,585
Start	1,287
Linear	1,334
Curved	1,445

B.

Cluster ID	Genes	Median profile	Point of interest
1	239	1. High expression in Pc 2. Low expression in M 3. Lowest expression in C 4. Low expression in RP	Genes associated with Pc
2	237	Similar pattern to Cluster 1, less expression of Pc	Genes associated with Pc and RP
3	174	1. Low expression in Pc 2/3. Lowest expression in M and C 4. High expression in RP	Genes associated with RP
4	194	1. High expression in Pc 2. Low expression in M 3. Lowest expression in C 4. Medium expression in RP	Genes associated with Pc and RP
5	224	1. Almost no expression in Pc 2. Low expression in M 3. High expression in C 4. Low expression in RP	Genes associated with C
6	133	1/2/3. Low expression in Pc, M and C 4. High expression in RP	Genes associated with RP
7	50	Similar pattern to Cluster 6, with almost no expression in Pc, M and C	Genes associated with RP
8	104	1. Low/medium expression in Pc 2/3. High expression in M and C 4. Low/medium expression in RP	Genes associated with M and C; supressed in Pc and RP
9	230	1. Low expression in Pc 2. High expression in M 3/4. Low expression in C and RP	Genes associated with M

C.

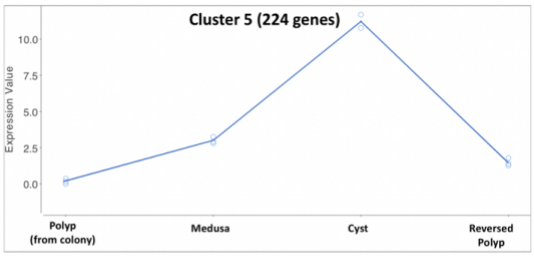


Figure II.2: Results from the DGE analysis of *T. dohrnii*'s life cycle reversal. A) Number of significant DEGs. B) Number of significant DEGs in each profile cluster based on hierarchical clustering. C) Averaged gene expression profile of cluster 5 with 224 DEGs, where expression values that are closer to 0 indicate suppressed genes, larger values indicate enriched genes. [Pc- Polyp (from wild colony), M- Medusa, C- Cyst, RP- Reversed Polyp; Green: Genes associated with Cyst]

Table II.1: Annotations of differentially expressed genes in Cluster 5. A) Categories of GO terms and the number of GO terms within category among the top 50, top 100, and all 224 genes. B) Gene codes, BLAST protein assignment, and GO function (GO term ID) of representative DE genes for each named category. The expression profiles for the listed DE genes are found in Figure II.3.

A.

Categories	# of Sequences		
	Top 50	Top 100	All 224 genes
Unannotated (novel genes)	22 (44%)	38 (38%)	60 (26.8%)
Aging and Lifespan	2	4	10
Transposable elements	4	7	10
DNA repair and damage response	0	4	13
Cancer and tumor	3	3	4
Cell and tissue differentation	2	4	16
Development	5	15	60

B.

Gene Codes	BLAST Protein Assignment	GO function (GO Term ID)	Representative Category
Td_DN103197_c0_g1	N/A	N/A	Unannotated
Td_DN92408_c0_g2	N/A	N/A	Unannotated
Td_DN99579_c2_g3	Serine racemase-like gene	Aging (GO:0007568)	Aging and Lifespan
Td_DN111257_c5_g5	Peptide methionine sulfoxide reductase-like (MsrA)	Determination of adult lifespan (GO:0008340)	Aging and Lifespan
Td_DN111173_c2_g5	Transposable element Tc3 transposase	Transposition, DNA-mediated (GO:0006313) DNA integration (GO:0015074)	Transposable elements and DNA integration
Td_DN108173_c0_g1	Uncharacterized protein LOC110239820	DNA integration (GO:0015074)	Transposable elements and DNA integration
Td_DN110869_c3_g4	E3 Ubiquitin-protein ligase Torpors	Mitotic G2 DNA-damage checkpoint (GO:0007095) Intrinsic apoptotic signaling pathway in response to DNA damage (GO:0097193)	DNA repair and damage response
Td_DN103087_c0_g3	Tumor necrosis factor receptor superfamily member 16-like (TPRG1L)	Response to wounding (GO:0009611) Induction of programmed cell death (GO:0012501) Positive regulation of MAPK cascade (GO:0043410) Nerve development (GO:0021675) Negative regulation of cell cycle (GO:0045786)	Cancer and tumor
Td_DN97842_c1_g1	Protein FEV	Negative regulation of cell differentiation (GO:0045596) Cell fate determination (GO:0001709) Stress-activated MAPK cascade (GO:0051403)	Cell and tissue differentiation
Td_DN105664_c0_g1	Methionine aminopeptidase 2	Wnt signaling pathway (GO:0016055) Stem cell differentiation (GO:0048863)	Cell and tissue differentiation

Representative DEGs for Each Annotation Category

Besides several unannotated genes (see Figure II.3AB), genes associated with aging and lifespan, transposable elements, DNA repair and damage response, cancer/tumors, and cell differentiation and development were found in Cluster 5 (Table II.1; Figure II.3C to H). Representative DEGs for each category, with their gene code, BLAST protein assignment, GO function, and term IDs, are reported in detail in Table II.1B. Below we summarize our findings and highlight annotation categories and their representative DEGs found in Cluster 5 that show potential to be involved in the cellular reprogramming and reverse development process in *T. dohrnii*.

We report genes involved in lifespan and aging (i.e., Serine-racemase-like' and 'Peptide methionine sulfoxide reductase-like (MsrA))' and in transposition and DNA integration (Table II.1, Figure II.2C to F). Among them is 'Transposable element Tc3 transposase', largely explored in *C. elegans* as one of the most active transposons (Bessereau 2006). Genes associated with DNA repair and response to DNA damage, such as Ubiquitin-related genes, proteins that targets cellular destruction by proteasomes in response to DNA damage, were also found in Cluster 5 (Table II.1, Figure II.2G). Several cancer and tumor-related genes were also reported. Among them were 'Breast cancer type 1 susceptibility protein isoform X1 (BRCA1)', 'Tumor protein p63-regulated gene 1-like protein (TPRG1L)', and 'Tumor necrosis factor receptor superfamily member 16-like (Tnfrsf16)' (Appendix II.F, Figure II.2H).

Finally, we report numerous genes associated with cell/tissue differentiation and development, such as 'Protein PEV' and 'Methionine aminopeptidase 2' with functions

related to cell fate determination and embryonic development, respectively (Table II.1, Figure II.3IJ). A functional gene enrichment analysis of Cluster 5 reported that terms associated with larval development, the response to DNA damage, and protein monoubiquitination were among the most enriched (Appendix II.E). On the other hand, processes associated with cytoskeleton and chromosome organization, both specific child-GO terms of broader mitotic cell division processes, were among the most suppressed in Cluster 5 (Appendix II.E).

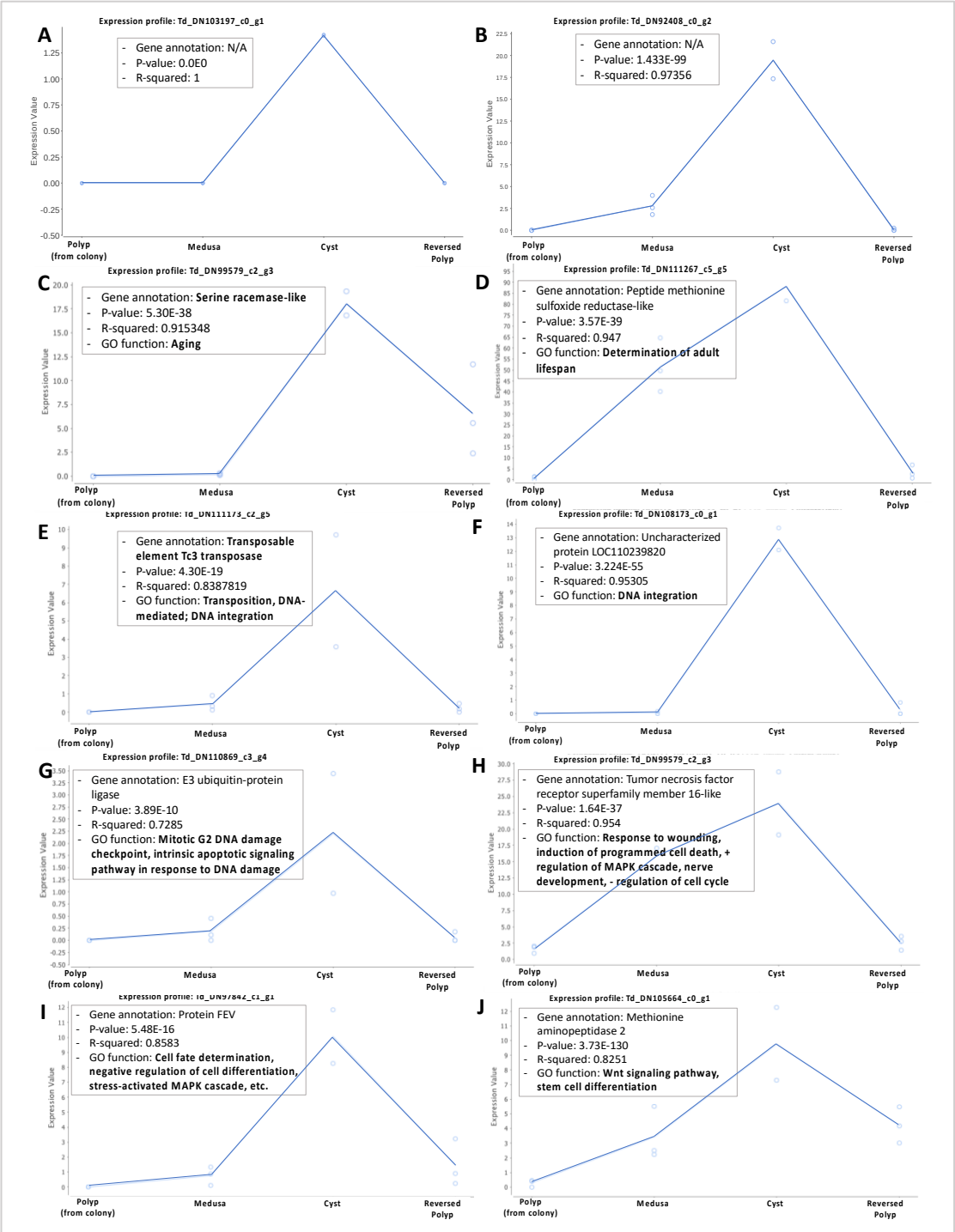


Figure II.3: Expression profile of DEGs. A. Expression profile of an unannotated gene Td_DN103197_c0_g1 from cluster 5; B. Expression profile of an unannotated gene Td_DN92408_c0_g2 from cluster 5; C. Serine racemase-like (Td_DN99579_c2_g3); D. Peptide methionine sulfoxide reductase-like (Td_DN111257_c5_g5); E. Transposable element Tc3 Transposase (Td_DN111173_c2_g5); F. Uncharacterized protein

LOC110239820 (Td_DN108173_c0_g1); G. E3 Ubiquitin-protein ligase Torpors (Td_DN110869_c3_g4) H. Tumor necrosis factor receptor superfamily member 16-like (Td_DN103087_c0_g3); I. Protein FEV (Td_DN97842_c1_g1); J. Methionine aminopeptidase 2 (Td_DN105664_c0_g1).

Pair-wise DGE analyses

Pair-wise DGE analyses were performed to identify differences in transcriptional profile between 1) Colonial Polyp and Reversed Polyp, and 2) Polyp and Medusa. Due to the large number of total DEGs in both comparisons, only the top 50 were further analyzed. DEGs and functional categories that may reflect important physiological differences between the life cycle stages are reported below and summarized in Table II.2.

Table II.2: **Summary of the pair-wise DGE analyses.** Colonial Polyp vs. Reversed Polyp (top), and Polyp vs. Medusa (bottom).

Analysis	Representative Genes	Funtional Category
Colonial Polyp vs. Reversed Polyp	Bromodomain adjacent to zinc finger domain protein 1A (BAZ1A) (Oppikofer et al., 2017)	Chromatin Remodeling
	Sushi, von Willebrand factor type A, EGF and pentraxin domain-containing protein 1 (SVEP1) (Shur et al., 2006)	Chromatin Remodeling
	Matrix Metalloproteinases (MMPs)-II; MMP-14 isoform X1	MMPs
	Thyrotroph embryonic factor-like (TEF) (Gavriouchkina et al., 2010)	Embryonic development
Polyp vs. Medusa	ATP-binding cassette sub-family A member 3-like (ABCA3)	Cellular and transmembrane transport
	Collagen alpha-1, Collagen alpha-6 chain	Collagen
	Clathrin heavy chain (Vassilopoulos et al. 2014) Actin (Vassilopoulos et al. 2014) Myosin (Rayment et al. 1993) Calmodulin (Chin 2005) Dystroglycan (Adams and Brancaccio 2015) Elongation Factor 1A (Murray et al.,1996; Owens et al., 1992)	Muscle development/assembly and contraction
	Dystroglycan (Yatsenko et al., 2014) Clathrin heavy chain (Sato et al., 2009)	Nervous system development

Colonial Polyp vs Reversed Polyp

While sharing 16,614 transcripts, the Reversed Polyp had a larger number of unique transcripts when compared to both the colonial polyp and to the rest of the stages (respectively 11,603, and 9,729). The majority of the 50 most significant DEGs were enriched in the Reversed Polyp and suppressed in the Colonial Polyp (Figure II.4A/B, Appendix II.H). The most enriched gene in the Reversed Polyp was 'Bromodomain adjacent to zinc finger domain protein 1A (BAZ1A)', followed by 'Sushi, von Willebrand factor type A, EGF, and pentraxin domain-containing protein 1 (SVEP1)' (Appendix II.H), both involved in chromatin remodeling (Zaghlool et al. 2016; Wang et al. 2004). Matrix metalloproteinases (MMPs), such as MMP-II and MMP-14 isoform X1', were also upregulated in the Reversed Polyp. Furthermore, 'Thyrotroph embryonic factor-like (TEF)', a gene active during mammalian embryogenesis (Drolet et al. 1991), with GO terms associated to embryonic development, was over-expressed in the Reversed Polyp (Appendix H).

An enrichment analysis using all significant DEGs showed the Reversed Polyp enriched in a variety of GO categories, such as reproduction, development/growth, symbiotic processes, actin filaments/microtubule, and response to stimuli (Figure II.5A). On the other hand, the Colonial Polyp was enriched in genes associated with the regulation of RAC protein signaling transduction and G-protein receptor signaling pathway (Figure II.5A), both overlapping processes that use Rho GTP-binding proteins to initiate signaling cascades (Ridley 2006).

A.

	# Genes	
	Reversed Polyp vs. Polyp	Medusa vs. Polyp
Total signficant DE genes (FDR < 0.05)	2,905	4,485
Up-regulated	1,622	2,432
Down-regulated	1,283	2,053

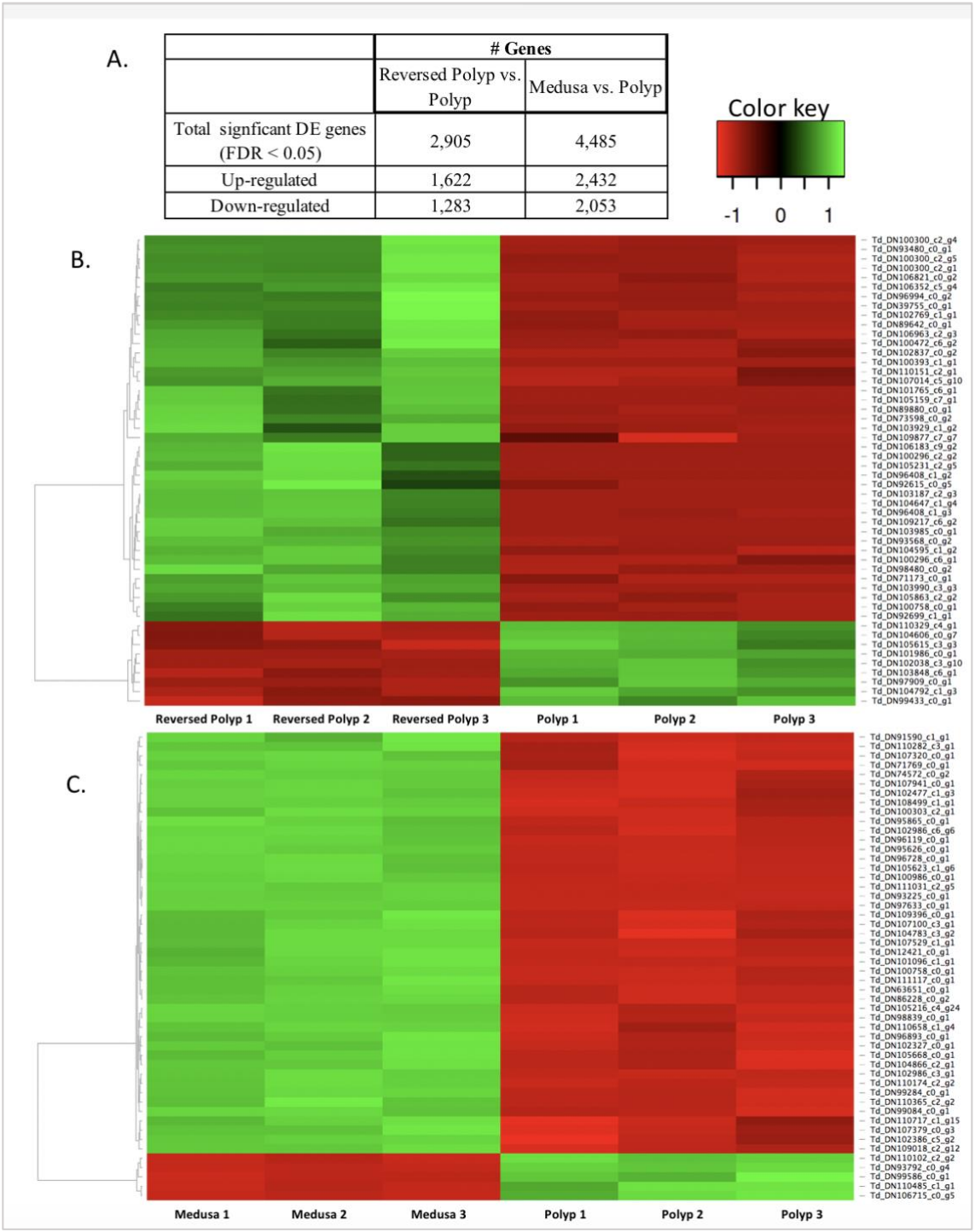


Figure II.4: A. Number of over-expressed and under-expressed genes in the Reversed Polyp vs. Polyp and Medusa vs. Polyp pair-wise DGE analyses. B./C. Visualization of the top 50 most DEGs (red: under-expressed, green: over-expressed) when Polyp was compared to the Reversed Polyp, and Medusa was compared to the Polyp. Heatmap created by the Heatmapper software (Babicki et al. 2016).

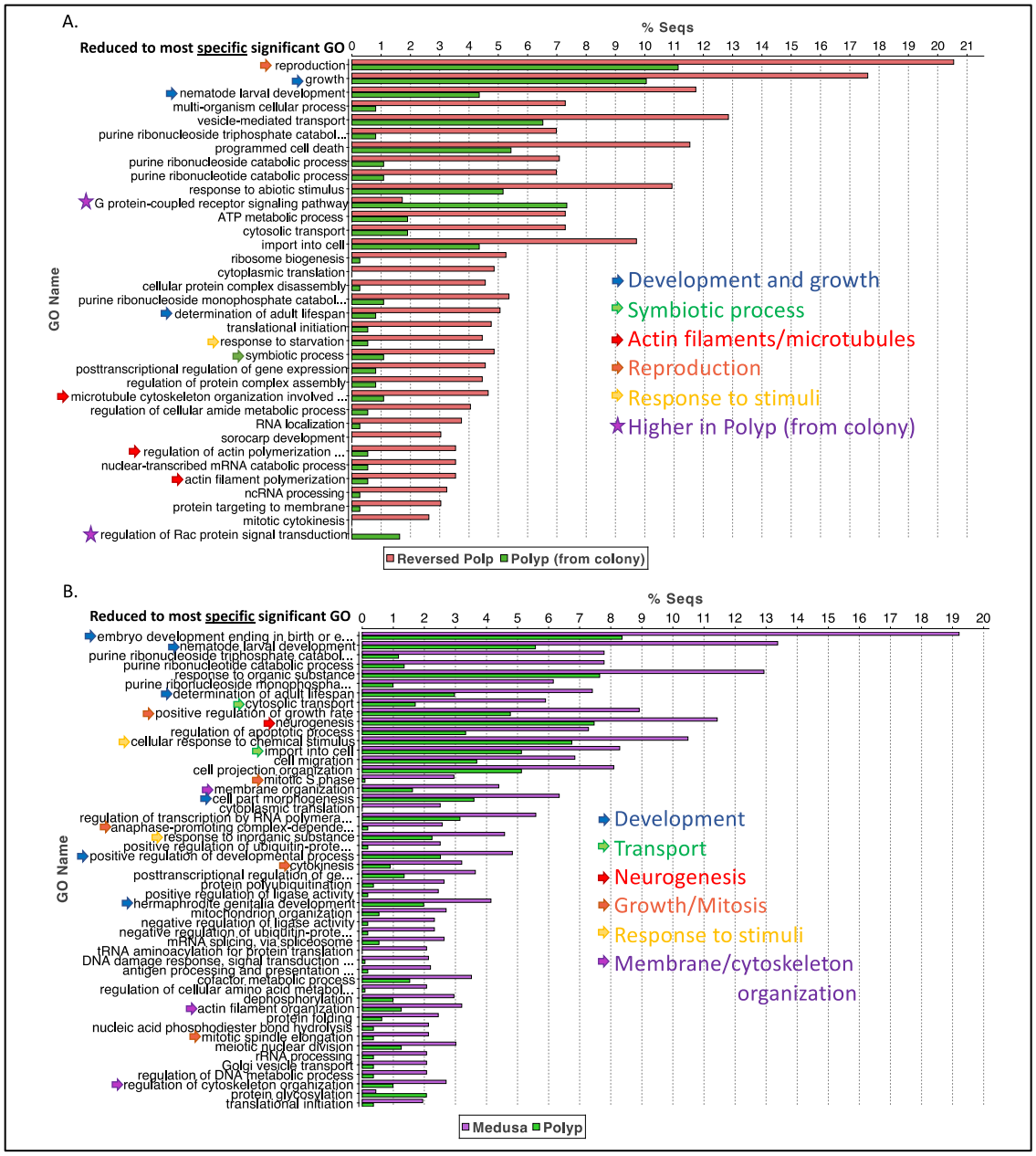


Figure II.5: Functional gene enrichment analyses. A. Reversed Polyp vs. Polyp and B. Medusa vs Polyp analysis, where significantly enriched and suppressed categories were reduced to the most specific terms (sorted by most different between Reversed Polyp/Medusa and Polyp).

Medusa vs. Polyp

In the Medusa vs Polyp comparison, the majority of the most significant 50 DEGs were over-expressed in the Medusa, with only three annotated genes over-expressed in the Polyp (i.e., Ectin-like, Galaxin-like, Prefoldin subunit 5 (Appendix H)). Notably, in the 50 most significant DEGs, 17 (34%) transcripts were novel (Appendix H). This is substantially higher than the Polyp vs. Reversed Polyp comparison that had 10 (20%) novel transcripts (Appendix H). The most enriched gene in the Medusa was 'ATP-binding cassette sub-family A member 3-like (ABCA3)' (Appendix II.H). In addition to ABCA3, five other genes with GO terms involving cellular and transmembrane transport were over-expressed (Appendix II.H). Other over-expressed functional categories in the medusa were related to collagen, muscle development/assembly, and contraction, response to external stimuli (light, salt, osmotic stress), glycolysis and glucose catabolism, and nervous system development (Table II.2, Appendix II.H). Likewise, an enrichment analysis of all significant DEGs also reported an up-regulation of development, transport, nervous system, response to stimuli, and membrane/cytoskeleton organization related categories in the Medusa (Figure II.5B). On the other hand, there were only 9 GO categories (reduced to most specific) that were found to be enriched in the Polyp (Appendix II.I). Among these categories were 'Chitin metabolic process', 'Digestion', and the "Formation of the primary germ layer'.

Unique and shared transcripts among life cycle stages

Comparison of the BLASTx annotations revealed that the Reversed Polyp and the Colonial Polyp share 16,614 transcripts, and have 11,603 and 3,806 unique

transcripts respectively (Figure II.6A). The Medusa and Polyp share 15,946 transcripts and have 2,537 and 4,474 unique transcripts, respectively (Figure II.6B). The four stages involved in life cycle reversal (Polyp, Medusa, Cyst, and Reversed Polyp) share 12,388 transcripts, with 2,046 polyp-specific, 1,016 medusa-specific, 1,455 cyst-specific, and 9,729 reversed polyp-specific transcripts (Figure II.6C).

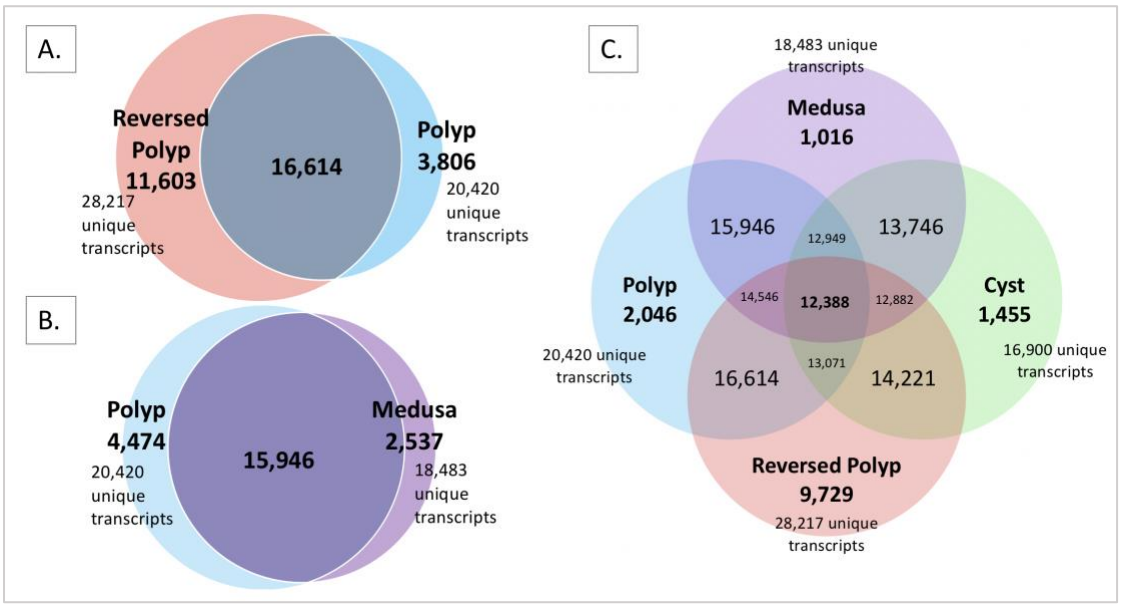


Figure II.6: Venn diagrams of BLASTx annotations for shared and unique transcripts. A) Reversed Polyp compared to the Polyp (from colony), B) Medusa compared to the Polyp (from colony), C) All four lifecycle stages involved in reverse development.

II.3. Discussion

Genes Enriched in the Cyst During Reverse Development

The cyst is the focal stage of interest in *T. dohrnii*, because the analyses of genes overexpressed in the cyst during life cycle reversal (represented in Cluster 5) help us identify processes directly involved in *T. dohrnii*'s reverse development and cellular reprogramming (i.e., transdifferentiation). We find aging and lifespan-related genes, serine racemase and MsrA, are enriched in the cyst. Suppression of serine racemase activity can cause aging-related cognitive dysfunctions in mammals (Turpin et al. 2011), while MsrA has a role in protecting cells from oxidative damage by destroying reactive intermediates or repairing damaged DNA, subsequently having a profound effect on regulating longevity in mammals (Moskovitz et al. 2001; Weissbach et al. 2002; Koç and Gladyshev 2007). Matsumoto et al. (2019) reported fewer transcripts associated with aging and lifespan GOs in the cyst in comparison to the polyp (colony) and medusa stages. Although the two analyses were different in nature (GO count vs. DEG analyses) and limited comparisons can be made, both indicate that the significant regulation of named genes that influence biological aging and lifespan may be necessary to the reverse development in *T. dohrnii.*

In the cyst, we also found an increased expression of genes controlling transposable element activity, DNA repair, and response to DNA damage, such as Ubiquitin-related factors (Figure II.3, Table II.1). This result indicates that the maintenance and regulation of genome integrity have an important role during the ontogenetic reversal and transdifferentiation processes of *T. dohrnii*. A similar result was

found in Matsumoto et al. (2019), where an analysis of the raw GO counts in the DNA repair, integration, and transposition categories showed significantly more transcripts in the cyst than the polyp and the medusa. In addition, GO analyses of the cyst stage (Matsumoto et al. 2019) reported that the cellular response to DNA damage stimulus was the only over-represented term associated with the response to stimuli compared to the polyp and medusa. Our results are consistent, where DNA repair, integration, transposition, and response to damage being overexpressed categories in the cyst. They represent major focus points for future research on the genetics of regenerative and reprogramming mechanisms in *T. dohrnii*.

We also show that several cancer and tumor-related genes known to control the cell cycle and active during embryonic development (Calvo and Drabkin 2000; Slamon and Cline 1984; Monk and Holding 2001) and tissue regeneration (Beausejour and Campisi 2006; Oviedo and Beane 2009), are enriched in the cyst (Figure II.3, Table II.1). Such genes include tumor suppressor BRCA1, which repairs DNA and inhibit cells from dividing erratically (Chen and Parmigiani 2007), and TPRG1K, a gene regulated by the TERC component of the telomerase enzyme that can activate the NF-κB (Nuclear Factor kappa B) signaling pathway and controls immune response (Liu et al. 2019). The NF-κB pathway has also been recognized as a vital regulator of the initiation and advancement of cancer (Hoesel and Schmid 2013), and its manipulation has been shown to extend lifespan in mice and halt epigenetic aging (Adler et al. 2007).

In summary, our findings illustrate a scenario where maintenance and regulation of genome integrity, regulation of cell division and proliferation, as well as genes

canonically involved in the regulation of aging/lifespan, are active during the ontogenetic reversal and transdifferentiation processes of *T. dohrnii*. Furthermore, we find that 44% of the top 50 genes in Cluster 5 (thus upregulated in the cyst and with little or no expression in the other stages) are novel/non-annotated genes (Table II.1A). This is a significantly higher ratio of novel/not annotated genes than found in the entire transcriptome. Moreover, the most significant DEG among the 224 genes in Cluster 5 also has no annotation. This indicates a further avenue of research where an effort should be spent toward the annotation of such genes that seem to play a crucial role in the events of rejuvenation and cellular transdifferentiation that occur in *T. dohrnii*.

Colonial Polyp vs. Reversed Polyp

The comparison between the colonial polyp and reversed polyp aims to identify differences in gene activity, if any, between the same life cycle stage (the polyp) produced by two different developmental pathways (budding and reverse development through the cyst). Our pair-wise DGE and enrichment analyses show that polyps produced through reverse development and polyps originating from a wild colony via asexual budding show remarkably different transcriptome profiles (see Figure II.4AB, 5A, and 6A), overall having the most distinct patterns of gene expression among the pairwise comparisons performed in this paper. Below, we highlight the roles of genes associated with chromatin remodeling, MMPs, and embryonic development, which were found among the most significantly enriched in the reversed polyp (Table II.2).

Chromatin remodeling gene BAZ1A repairs damaged chromatin and promotes survival in human cells (Oppikofer et al., 2017), while chromatin binding SVEP1 is

involved in cell adhesion. Polymorphism of SVEP1 gene has also been correlated with human longevity (Shur et al., 2006; Yashin et al., 2010). In metazoans, MMPs function to maintain the extracellular matrix and regulate the interaction between cells and matrix during regeneration (Yang and Byant 1994; Bai et al. 2005; Vinarsky et al. 2005; Tseng and Levin 2008). These proteinases are also involved in tissue remodeling and engineering, and play important roles in the regeneration of zebrafish (Bai et al. 2005), newts (Vinarsky et al. 2005) and axolotls (Yang and Byant 1994). Among the genes enriched in the reversed polyp, some are associated with embryonic development, such as TEF, which plays a prominent role in activating DNA repair genes in response to stress and damage in zebrafish embryonic cells (Gavriouchkina et al., 2010).

We also show that the colonial polyp is enriched in genes related to the Rho GTP-binding signaling cascades (Figure II.5A), which is involved in a diversity of processes, including development, metabolic regulation, cellular growth and survival, and changes in the actin cytoskeleton (Yan and Jin 2012; Neves et al. 2002) (Figure II.5A).

In conclusion, our colonial vs. reversed polyp comparison shows that reparative and regenerative processes, which include maintaining chromatin, MMP induced tissue remodeling, and genetic pathways vital to embryonic development, are enriched in the reversed polyp but not in the colonial polyp. Such differences may reflect the fact that reverse development as a whole is a regenerative process that includes the re-establishment of cell types and physical features of the polyp from a different life cycle stage, the medusa.

Medusa vs. Colonial Polyp

The comparison between medusa and colonial polyp aims to identify differences in genetic networks between the planktonic and solitary stage (medusa) and the colonial stage (polyp). Medusae are considered more complex than polyps, being able to swim, sexually reproduce, and having nerves and sensory organs. Morphologically, medusae have a thick mesoglea, while polyps have chitinous protective exteriors. We discuss genes and networks attributable to the physiological differences of the medusa and polyp stages.

The most enriched gene in the medusa (Appendix II.H), ABCA3, plays a role in the transmembrane transportation of surfactants. These substances reduce the surface tension of liquids and crucial for normal respiratory development and processes in vertebrates (Anandarajan et al. 2009). This may be because hydromedusae produce mucins, a natural polymeric surfactant (Petrou and Crouzier 2018; Uzawa et al. 2009) in their mesoglea, an extracellular matrix absent in polyps. Collagens, components of the mesoglea (Schmid et al. 1991), were also upregulated in the medusa (Table II.2). Similarly, collagen is enriched in the medusa stage in comparison to polyps in the hydrozoan *Podocoryna carnea* (Sanders and Cartwright 2015b).

Consistent with the notion that medusae (but not polyps) contain smooth muscles and have a more complex sensory and nervous system, we found numerous genes that contribute to muscle development and contractions (Table II.2), and genes involved in neuron development and transmission (Table II.2, Appendix II.H) upregulated in the medusa. Among those is dystroglycan, the major mediator of the integrity of muscles

with some consistent functions from invertebrates to mammals (Adams and Brancaccio 2015; Shcherbata et al. 2007). Dystroglycan also functions as a critical regulator of proper development and of the nervous system in Metazoa (Yatsenko et al., 2014; Lindenmaier et al., 2019). On the other hand, we show an enrichment of genes involved in chitin metabolic processes and the formation of a primary germ layer in the polyps (Appendix II.I), likely reflecting the polyps' chitinous protective exterior and continuous allocation of energy towards asexual budding to expand the colony.

Overall, consistent with morphological and physiological differences between the medusa and polyp stages, our analyses identify DEGs associated to transmembrane transport, components of the mesoglea (e.g., mucin, collagen, etc.), muscular development and usage, neuronal development, and increased ability to respond to external stimuli in the medusa, and DEGs associated with chitin metabolism and formation of primary germ layer in the polyp. Furthermore, the top 50 DEGs in the medusa showed an unusually high number of novel transcripts. This may reflect the low number of published genomes of Hydrozoa with a medusa stage (the top BLASTx hit distribution shown in Figure II.1 includes Cnidaria without medusa) and the fact that, unlike other cnidarians, the medusae of *T. dohrnii* undergo reverse development. The study of Cnidaria with a variety of life cycles is needed to progress toward a better understanding of the genetics and evolution of their life stages.

II.4. Conclusion

The transcriptome assembly and profiling revealed that the cyst stage of *T. dohrnii* is enriched with genes that are associated with aging/lifespan, regulation of transposable elements, DNA repair and damage response, and Ubiquitin-related processes. We show that a large portion (44%) of the top 50 DEGs in the cyst, a unique life cycle stage within the Hydrozoa, are novel and not annotated. We also show that polyps from the two different developmental trajectories (medusa reversal and budding within a wild colony) exhibit significant differences in gene activity, with processes of chromatin remodeling, matrix metalloproteinases, and embryonic development being highly active in the reversed polyp. The polyp and medusa also show major differences, with transmembrane transport, nervous system, components of the mesoglea, and muscle contraction-related categories being over-expressed in the medusa. In contrast, categories related to chitin metabolism and the formation of primary germ layers are over-expressed in the polyp. Ultimately, we produce genomic tools in the form of a high-quality transcriptome and provide insight into genes and genetic networks associated with each of its life cycle stages and with the reverse development of *T. dohrnii*.

II.5. Material and Methods

Specimen collection, rearing and identification

A *T. dohrnii* colony bearing medusa buds was collected from Bocas del Toro, Panama in July 2015. Polyp hydranths were cut off the colony and preserved in

RNAlater (ThermoFisher Scientific- Catalog #: AM7020). Released medusae were isolated into petri dishes, and a few of the medusa and the remaining colony were preserved in RNAlater. Remaining medusae were starved until they formed into cysts, and subsequently reversed into the polyp stage. As there is variability in the time it takes for individual medusa to form into a cyst, this stage was preserved when the following physical attributes were visible (Schmich et al. 2007): 1) Cyst has no visible physical attributes remaining from the medusa (i.e. tentacles, mesoglea, etc.); 2) Settled and attached onto a substrate; 3) Cyst shows a smooth, yellow-orange perisarc (protective chitinous exterior); 4) Cyst has not formed any stolon. Prior to RNA extraction, total DNA was extracted from spare tissue from the colony all the specimens originated from (Zietara et al. 2000). A fragment of the mitochondrial 16S ribosomal RNA gene was amplified (Miglietta et al., 2006) to confirm the species of the samples (Accession.#: MH029858, (Miglietta et al. 2018)).

RNA extraction/library construction/sequencing

Total RNA was extracted from biological triplicates of polyp hydranths, medusae, cysts and reversed polyps using the Epicentre kit (MasterPureTM RNA Purification kit #MCR85102). The Poly A-based SMARTer Ultra Low Input RNA kit for Sequencing v4 kit (Clontech Laboratories) was utilized on all RNA samples (total of twelve). The cDNA generated in the amplification was constructed into libraries by the AgriLife Genomic and Bioinformatics Services at TAMU and sequenced in a single lane of the Illumina HiSeq 4000 platform (Illumina Sequencing Technologies). 150bp paired-end reads were generated for each library. Only two of the three generated cyst RNA-seq

replicate datasets were viable for further use, which is elaborated in detail below (described in detail in Matsumoto et al. 2019 and in Appendix II.A). The remaining eleven RNA-seq datasets were utilized in our paper. The RNA-seq datasets from the two cyst replicates were utilized in preliminary analyses (Matsumoto et al. 2019) and can be found under GenBank accession #: SRR10053756 and SRR10053757 (Under BioProject #PRJNA563171, BioSample ID #SAMN1266994).

De Novo transcriptome assembly

Specific parameters, databases and command lines for bioinformatics software and algorithms can be found in detail in Appendix II.K. All datasets of RNA-seq reads were mapped to the existing polyp and medusa *T. dohrnii* transcriptomes from the Mediterranean Sea (Matsumoto et al. 2019), and their % mapped, broken reads, and estimated paired distances were used to ensure proper and adequate sequencing quality. One of the cyst replicates had poor alignment rates against the polyp transcriptome, high percentage of broken reads, few mapped PE reads, and an abnormal range of estimated paired distance in comparison to the other cyst replicates and other sequenced stages (Matsumoto et al., 2019; Appendix II.A). Thus, that replicate was removed from subsequent analyses and transcriptome assembly. Remaining datasets were trimmed for quality, normalized and assembled using Trinity (Haas et al. 2013).

Completeness of transcriptome

Results outputted from Trinity's *in silico* normalization and all individual libraries were separately mapped back to the assembled transcriptome using two different stringencies in the CLC Genomic Workbench v8 alignment software

(Appendix II.A). BUSCO v2.0 (Simão et al. 2015) with the Metazoa database was used to evaluate the completeness of the transcriptome in terms of gene content and broadly assess transcript fragmentation.

Functional annotation of transcriptome and biological contaminant removal

The Kraken2 sequence classifier tool (Wood and Salzberg 2014) was used to filter contaminant sequences from Bacteria, Archaea and virus sources, and contigs less than 400bp were removed from the transcriptome prior to annotation as they tend to provide less biological meaning than longer transcripts, commonly having poor coverage and quality, and often lacking proper protein assignments and function (Kitchen et al. 2015; Appendix II.B). Moreover, trimming reads less than 400bp has been applied to other transcriptomes among Cnidaria (Kitchen et al. 2015; Sanders and Cartwright 2015a, 2015b).

The OmicsBox software's (BioBam) annotation pipeline (Götz et al. 2008; Conesa et al. 2005) was used to assign functional terms to assembled contigs. Blastx via CloudBlast (Matsunaga et al. 2008) was performed using a e-value of e^{-3}, and all contigs with hits were assigned names based on the best Blast description annotator tool within OmicsBox. InterProScan (IPS) (Zdobnov and Apweiler 2001) was used to build upon and confirm existing GO annotations. The IPS annotations were merged and ANNEX Augmentation and the manual removal of 1st level annotations were conducted as recommended (Götz et al. 2008). The EggNOG was also utilized to map, confirm and merge GO terms to the B2G and IPS annotations. KEGG enzyme GO-enzyme code mapping was additionally performed and visualized in Blast2GO. Contigs that were not

annotated by BLASTx, IPS, EggNOG or KEGG were further annotated using the RFAM database for non-coding RNA, and Blastn against the hydrozoan EST database and subsequent Blastx search.

Biological contaminants were inferred through the generated top Blastx hits chart. Contigs that had top-hits to the following genera, *Thecamonas*, *Acanthoamoeba*, *Planoprotostelium*, *Abelmoschus*, and *Acytostelium*, that showed higher than 95% sequence similarity were eliminated from the transcriptome. The remaining contigs that had the highest hit to one of the named genera were re-BLASTed against the Metazoa (taxa ID: 33208) NR database. Contigs that had no metazoan hits were predicted as biological contaminants and removed from the transcriptome before further downstream analyses due to the concern of biological contamination from culturing/rearing and epibiontic relationships. The remaining contigs were used for further downstream analyses. The final transcriptome (fasta file) and annotation (GFF file) can be found at http://www.therealimmortaljellyfish.com/data.

Differential gene expression analyses

Sequencing reads from each of the eleven libraries were individually mapped back to the transcriptome to generate count data for all samples. Gene-level estimation-based count data was generated using RSEM (Li and Dewey 2011). EdgeR (Robinson et al. 2010) was utilized to filter and normalize the count data using the upper quartile approach, and the outputted count data was utilized to conduct gene-level pairwise DGE analyses on the following: 1) Colonial Polyp vs. Reversed Polyp; 2) Medusa vs. Polyp. Sequential DGE analyses were conducted using the maSigPro Bioconductor package

(Nueda et al. 2014) on the stages in the following order: Colonial Polyp, Medusa, Cyst, Reversed Polyp. Significant DE genes were clustered into nine different gene expression profiles (i.e. gene activity pattern- linear, curved, etc.) using hierarchical clustering. Visualized expression profiles of representative genes in Figure 3 were chosen based on the highest DEGs from each category with the most consistent expression among the stage replicates from each category listed in Table II.1A (Appendix II.F).

Functional gene enrichment analyses

As all transcript isoforms belonging to a single unigene are combined for gene-level analyses, the transcript/isoform over 800 bp with the most GO annotated terms was utilized as the representative for the gene-level functional enrichment analyses. If no transcripts were over 800bp, the transcript with the most GO terms was used as the gene representative. Functional gene enrichment analyses were conducted on the DE genes using the FatiGO (Al-Shahrour et al. 2004). Two-tailed Fischer's Exact Tests were performed on pair-wise and sequential DGE analyses. The biological processes GO domain was the focal domain of interest for subsequent analyses.

CHAPTER III

GENETIC NETWORKS OF REGENERATION, CELL PLASTICITY AND LONGEVITY IN THE IMMORTAL JELLYFISH (*Turritopsis dohrnii*)

III.1. Introduction

Comparative studies of developmental processes among diverse metazoan groups have shown that genetic programs for development are conserved. This is a consequence of natural selection acting upon and altering conserved gene networks that may induce new proteins and cellular function (Wagner, 1994; Levine and Tijian, 2003; Butland et al., 2005; Wildwater et al., 2005; Costa and Shaw, 2006). The existence of parallel gene networks among animals have restructured and expanded the biomedical sector, in which valuable experimental research on both metazoans and non-metazoans have contributed to further our understanding of human health and disease (Brockes and Kumar, 2002; Lee et al., 2003; Longo and Kennedy, 2006; Takahashi and Yamanaka, 2006; Fontana et al., 2010; Elliot and Sanchez-Alvarado, 2013). The ability for medusae of *Turritopsis dohrnii* to undergo reverse development via cell transdifferentiation makes it an exceptional *in vivo* system to further expand our understanding of tissue regeneration, cell plasticity and aging/longevity in animals.

Most animals are able to regenerate tissue in the form of wound healing. After a substantial loss of tissue however, such as entire limbs and/or organs, few species show the ability to fully redevelop lost body structures. Various genes involved in embryonic development have been shown to be key components of tissue regeneration (Birnbaum

and Sanchez-Alvarado and Yamanaka, 2014). Despite the drastic differences in regenerative mechanisms among animals, some underlying genetic network and constituents have been reported to be highly conserved across a variety of metazoan groups (Brockes and Kumar, 2008; Bely and Nyberg, 2010).

Cellular pathways involved in regeneration, cell plasticity and longevity are interconnected and regulated in complex genetic networks. The decrease in an individual's regenerative potential causes the accumulation of cellular damage, consequently resulting in aging and senescence among most animals (Conboy et al., 2005; Moskalev et al., 2012; Sousounis et al., 2014). Additionally, self-renewing stem cells and damage repair pathways start to malfunction and slow down when aged, adding to the inability to remedy cell damage (Terman and Brunk, 2004; Lopez-Otin et al., 2013). RNA interference and knock-out experiments with transgenic model organisms have uncovered genes that directly influence regenerative capabilities, cell pluripotency, and the progression of aging. Contributing genetic networks that have been experimentally manipulated and documented to influence regeneration, cellular plasticity and aging, includes Sirtuin (SIRT) proteins (Borradaile et al. 2011; Grabowska et al. 2017; Amano and Sahin 2019), telomere maintenance via telomerase (Amano and Sahin 2019; Whitaker et al. 1995; Nowak et al. 2006; Flores and Blasco 2010), heat shock proteins (HSPs) (Patruno et al. 2001; Tower 2011; Shi et al. 2007; Hsu et al. 2003), and the Yamanaka transcription factor gene family (Oct, Sox, Klf and Myc) (Takahashi et al. 2007; Takahashi and Yamanaka 2006).

Gene enrichment analyses of *T. dohrnii* have reported telomere maintenance/organization and DNA repair processes to be elevated in the cyst stage in comparison to the polyp and medusa stages (Matsumoto et al., 2019). Additionally, DGE analyses reported genes that are associated to aging and lifespan and DNA repair were found to be enriched at the cyst stage during the reverse development (Matsumoto and Miglietta, 2021). The presented research aims to identify and profile the expression of additional genetic constituents and networks that modulate regeneration, pluripotency and aging in mammalian systems during *T. dohrnii*'s reverse development sequence, specifically SIRT proteins, telomerase-related genes, HSPs, and the Yamanaka Factors gene family. In addition, the research aims to infer the evolutionary history of highly-regulated genes within the named networks in *T. dohrnii*. The presented research results have the potential to contribute to the discovery of molecular drivers that underlie cell reprogramming and longevity in metazoans.

III.2. Material and Methods

Construction of Super-Transcripts and BLAST protein assignment

Super-Transcripts (Davidson et al., 2017), an replacement for a reference genome that contain all unique exons from transcript isoforms, were constructed via the Trinity software (Grabherr et al., 2011) from published RNA-sequencing (RNA-seq) libraries of life cycle stages of *T. dohrnii* (colonial polyp, medusa, cyst, reversed polyp) from Bocas del Toro, Panama (NCBI Accession # SAMN12669945, SAMN13924705-SAMN13924707; Matsumoto and Miglietta, 2021). BLASTx against the Metazoa

(TaxID 33208) database (e-value cutoff= e^{-5}) was performed on the Super-Transcripts to add gene descriptions to each of the sequences. The following key terms were used to screen for genes of interest among the BLAST descriptions (i.e., protein assignments):

i. SIRT: Sirt, Sirtuin
ii. Telomere and telomerase-related genes: telomerase, telomere
iii. HSPs: Heat shock, Heat-shock, HSP
iv. Yamanaka factor gene family: POU, Octamer, Oct, Sox, Sex-determining region Y, Klf, Krueppel-like, Myc

Expression normalization and profiling

The RNA-seq libraries of *T. dohrnii* life cycle stages were trimmed using Phred score cutoff of 10 as recommended for gene profiling analyses (MacManes, 2014) and aligned to the super-transcriptome using RSEM (Li and Dewey, 2011) to generate gene-level count data. The maSigPro Bioconductor package (Conesa et al., 2006) was used to perform upper quartile normalization of the expression data using the upper quartile method (75% quantile). The data was analyzed in the reverse development sequence: colonial polyp to medusa, to cyst, to reversed polyp. Normalized expression values for each super-transcript were averaged per life cycle stage. Only genes with expression data consistent across biological stage replicates (i.e., expression present in all replicates or absent in all replicates) were used in further analyses and expression profiling visualization.

Transcripts of interest search for Yamanaka Factors (Oct4, Sox2, Klf4, c-Myc)

Transcripts of interest were identified in the published transcriptome and RNA-seq libraries from Matsumoto and Miglietta (2021) using a multi-approach method: 1) tBlastx search; 2) RNA-seq analysis; 3) Annotation description screening analysis. For the tBlastx search, all variants of the human Yamanaka Transcription Factors were utilized (GenBank Accession found in Appendix D) and a e-value cutoff of e^{-10} was utilized. For the RNA-seq analysis, different stringencies of parameters (0.3, 0.5, 0.7 and 0.9 length and similarity fraction) were used with a mismatch cost value of 2 (linear gap cost), insertion value of 3 and deletion value of 3. Variants with the lowest e-value and highest bit-score larger than 400 bp were prioritized (i.e., Oct4 variant 1).

Construction of gene trees

The PSI-BLAST algorithm was used to find SIRT3, HSP90, and POU-factor homologs among various metazoan taxa with the assembled *T. dohrnii* super-transcripts as the query sequence. Clustal (Higgins et al. 1992) and MUSCLE (Edgar 2004) multiple sequence alignment algorithms used to find the best alignment. Gene trees were constructed using both Maximum Likelihood (ML) (PhyML (Guindon and Gascuel 2003)) and Bayesian inference based-approaches (MrBayes (Ronquist et al. 2012)). The best-fit models were calculated using the model test from RaxML. For the PhyML, tree topology was inferred using 1000 bootstrap replicates, dependent on sequence alignment size. Consensus trees with estimated bootstrap values were generated. For MrBayes, four parallel Markov chain Monte Carlo runs were carried out for $1x10^7$ generations, trees were sampled every 100th generation, and burn-in was set to 10,000 generations.

Consensus trees with estimated clade posterior probabilities were generated. Details for each alignment and gene tree follows:

i. SIRT3: 42 amino acid SIRT3 protein sequences downloaded from NCBI after PSI-BLAST. The MUSCLE alignment resulted in a consensus length of 284AA. The best substitution model was LG+I+G+F.

ii. HSP90: 43 amino acid HSP90 protein sequences were downloaded. The MUSCLE alignment resulted in a consensus length of 693 AA. The best substitution model was LG+I+G+F.

iii. POU factors: 42 amino acid POU3 factor sequences were downloaded. The Clustal alignment resulted in a consensus length of 108AA. The best substitution model was JTT+G. Three POU5F1 (i.e. Oct4) sequences used in Millane et al. (2011) and the human Oct4 was incorporated into the alignment.

III.3. Results and Discussion

Expression profiling of candidate genes

Gene expression profiling of life cycle stages involved in the reverse development of *Turritopsis dohrnii* can reveal interconnected genetic networks underlying long-lifespans, tissue regeneration and cellular plasticity in metazoans. Super-transcripts have been reported to increase the accuracy of gene expression profiling analyses, and provide distinct advantages over using a single representative sequence for a single gene, often the longest transcript isoforms (Davidson et al. 2017). In the presented work, we used super-transcripts constructed from *T. dohrnii*'s life cycle stages involved in reverse development (colonial polyp, medusa, cyst, reversed polyp) to profile the expression of Sirtuin (SIRT) proteins, telomere and telomerase-related genes, Heat-shock proteins (HSPs), and the Yamanaka Factor gene family (Oct, Sox, Klf, Myc). Only the expression profiles of super-transcripts with consistent expression among all biological replicates of *T. dohrnii* (i.e., present or absent in all replicates) are presented below.

Sirtuin proteins

The overexpression of the SIRT protein Sir2 (silent information regulator 2) has been found to extend lifespan in a variety of taxa such as yeast (Kaeberlein and Powers III 2007), nematodes (Tissenbaum and Guarente 2001) and fruit flies (Rogina and Helfand 2004). Seven human SIRT proteins that have identified (SIRT1-SIRT7). All SIRT proteins have been found in chicken, mouse, frogs, segmented worms and sea anemones (specifically *Nematostella*), but only SIRT1, SIRT4 and SIRT6 have been

identified in *Hydra* (Greiss and Gartner 2009). In *T. dohrnii*, all seven SIRT proteins have been identified among the protein assignments (i.e., annotations) of the super-transcripts, showing sequence similarity to a variety of different metazoan taxa, from sponges to vertebrates. This is likely due to the fact that SIRT proteins, which have been sporadically lost during basal to upper-metazoan evolution (Greiss and Gartner 2009).

Among the *T. dohrnii* super-transcripts, we identified 34 SIRT sequences with expectation value (e-value) e^{-11} to e^{-128} against the NR Metazoa database (Appendix III.A). We identified putative homologs for all seven SIRT proteins, SIRT6 being the most prevalent (Figure III.1A). Notably, a wide range of metazoan taxa was identified as the top-hit species for the SIRT annotated sequences (Figure III.1B-C).

There is evidence of SIRT1, 2, 3 and 6 activity during the reverse development of *T. dohrnii* (Figure III.D-E). SIRT3 was the highest expressed among all stages, exhibiting a sharp peak in activity in the cyst (Figure III.1D, light blue). SIRT3 is involved in mitochondrial energy metabolism and protects cells against oxidative stress, which has been reported to shorten cellular lifespan and accelerate the shortening of telomeres (Lundberg et al. 2000). Furthermore, it is the only SIRT with experimental evidence directly correlated to human longevity (Bellizzi et al. 2005; Kincaid and Bossy-Wetzel 2013). The expression SIRT1, 2 and 6 were low compared to SIRT 3 (Figure III.1E). Both SIRT6 sequences were inactive (Figure III.1E, green) or had low expression in all stages (Figure III.1E, yellow), with an exception in the reversed polyp. SIRT6 has significant implications in increasing the efficiency of DNA repair in long-lived mammals in comparison to its short-lived counterparts (Tian et al. 2019). In

addition to DNA repair, SIRT6 functions to protect telomeres and stabilize the genome (Jia et al. 2012). Two of the SIRT2 sequences showed highest activity in the reversed polyp stage (Figure III.1E, purple/orange), while one peaked at the medusa stage (Figure III.1E, gray). SIRT2 is involved in cell-cycle control and development, and is also known to be a marker of cellular senescence in humans (Grabowska et al. 2017). Lastly, SIRT1 was the only SIRT in which transcriptional activity decreased from the colonial polyp the reverse polyp stage (Figure III.1E, blue). SIRT1 is associated with DNA repair, glucose metabolism, insulin production and promotes cell survival in combination with delaying replicative senescence (Grabowska et al. 2017; Bellizzi et al. 2007).

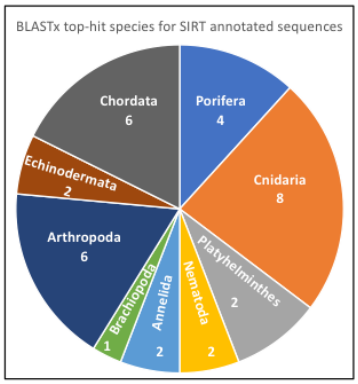
A.

Counts	
SIRT1	2
SIRT2	8
SIRT3	4
SIRT4	3
SIRT5	1
SIRT6	14
SIRT7	2

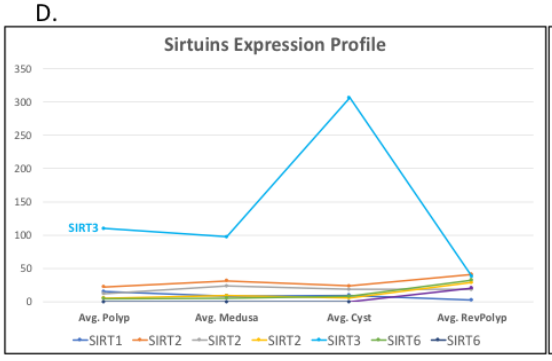
B.

Species	Phylum	Top-hits
Hydra vulgaris	Cnidaria	4
Amphimedon queenslandica	Porifera	4
Macrostomum lignano	Platyhelminthes	3
Pocillopora damicornis	Cnidaria	2
Fulmarus glacialis	Chordata	2
Apteryx rowi	Chordata	2
Acanthaster planci	Echinodermata	1
Lingula anatina	Brachiopoda	1
Pediculus humanus corporis	Arthropoda	1
Capitella teleta	Annelida	1
Anolis carolinensis	Chordata	1
Zootermopsis nevadensis	Arthropoda	1
Acropora millepora	Cnidaria	1
Strongylocentrotus purpuratus	Echinodermata	1
Litomosoides sigmodontis	Nematoda	1
Protobothrops mucrosquamatus	Chordata	1
Soboliphyme baturini	Nematoda	1
Helobdella robusta	Annelida	1
Dinothrombium tinctorium	Arthropoda	1
Penaeus vannamei	Arthropoda	1
Nicrophorus vespilloides	Arthropoda	1
Leptotrombidium deliense	Arthropoda	1
Exaiptasia pallida	Cnidaria	1

C.



D.



E.

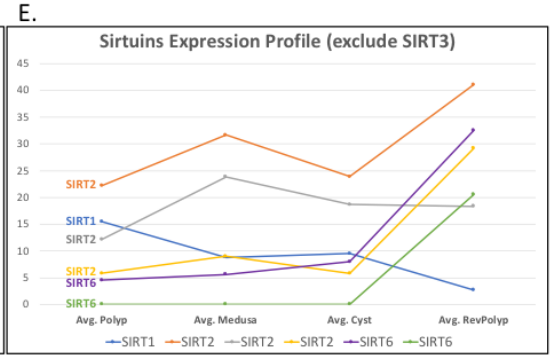


Figure III.1: Sirtuin protein putative homologs in *T. dohrnii.* A) Number of super-transcripts for each SIRT sequence; B) Number of BLASTx top-hits for each species and

their corresponding phylum; C) Pie chart of top-hit species for SIRT annotated sequences; D) Expression profiles of SIRT proteins (most consistent expression patterns across replicates chosen; complete list in Appendix III.A); E) Expression profiles of SIRT proteins excluding highest expressed SIRT3 sequence (D-light blue).

Telomeres and telomerase

The telomerase enzyme prevents the loss of genetic information during DNA replication via the elongation of protective sequences at the end of chromosomes. Telomerase is inactive in most somatic cells in metazoans, with some exceptions such as epidermis cells and malignant and immortal germline cells, where indeterminate cell division occurs (Flores and Blasco 2010; Blasco 2007). Previous work has shown that transcriptional regulation of telomere maintenance genes occurs in the cyst of *T. dohrnii* (Matsumoto et al. 2019).

In the presented work, we identified 56 telomere/telomerase-related super-transcripts (e-value ranges of e^{-06} – 0) (Appendix III.B). Many top-hits corresponded to *Hydra*, other cnidarians, or sponges (Appendix III.B). Putative homologs of TEP1 (telomere-associated protein 1), a subunit of telomerase, was found to be the most numerous (Figure III.2A).

Expression profiling indicates 'Regulator of telomere elongation 1' (RTEL1) was the most active in *T. dohrnii*, showing a sharp increase at the rejuvenated polyp (Figure III.2B, red). RTEL1 plays a crucial role in regulating telomere length, DNA repair, genomic stability in mice and humans (Barber et al. 2008; Uringa et al. 2011). An EST1A (telomerase-binding protein EST1A) sequence exhibited a sharp peak in activity at the cyst stage (Figure III.2C, purple). EST1A binds and cooperates with TERT

(telomerase reverse transcriptase) to elongate telomeres (Snow et al. 2003), and ectopic expression of the TERT alone has been shown to induce telomere lengthening in a variety of human cell cultures (Bodnar et al. 1998; Vaziri and Benchimol 1998). One of the TERT sequences only showed expression in the cyst (Figure III.2C, brown), while the other had higher expression and were active in all stages, particularly in the colonial polyp (Figure III.2C, black). Although more downstream work is necessary to determine function of TERT and EST1A in *T. dohrnii*, our data indicates that the two genes are co-expressed in the cyst and may be working together to elongate telomeres prior to rejuvenation into the earlier polyp stage.

TEP1 was the most abundant among active genes in this category, exhibiting a variety of expression profiles, with the most active in the colonial polyp stage (Figure III.2D, blue). TEP1 is highly active in immortalized murine cell lines, but absent in somatic cells (Harrington et al. 1997). TCAB1(Telomerase Cajal body protein 1) was expressed among all stages but also most active in the colonial polyp (Figure III.2C, light green). TCAB1 is associated with the accumulation of Cajal bodies that deliver RNA sequences to telomerase to elongate and maintain telomere length (Venteicher and Artandi 2009).

Lastly, there was decreased activity of telomerase inhibitors TERF1 (Telomere repeat-binding factor 1) and PINX1 (PIN2/TERF-interacting telomerase inhibitor 1) during the transition from the medusa to the cyst (Figure III.2C, light gray/red). PINX1 has been reported to be a potent telomerase inhibitor (Zhou and Lu 2001), and TERF1 is known to bind and negatively regulate the length of telomeres and cooperates in the

shelterin complex to protect chromosomal ends in mammals (Derevyanko et al. 2017). In combination with the previous reports of telomerase activity in the cyst (Matsumoto et al., 2019), there is high potential for telomerase to be involved in *T. dohrnii* capabilities to reverse its ontogeny using cellular transdifferentiation.

A.

Counts	
TERF1	1
TERF2	1
TERB1	1
TEL2	3
Tel. length silencing	1
RIF1	1
TERT	8
TEP1	29
TCAB1	1
EST1A	3
RTEL1	2
POT1	1
PINX1	4

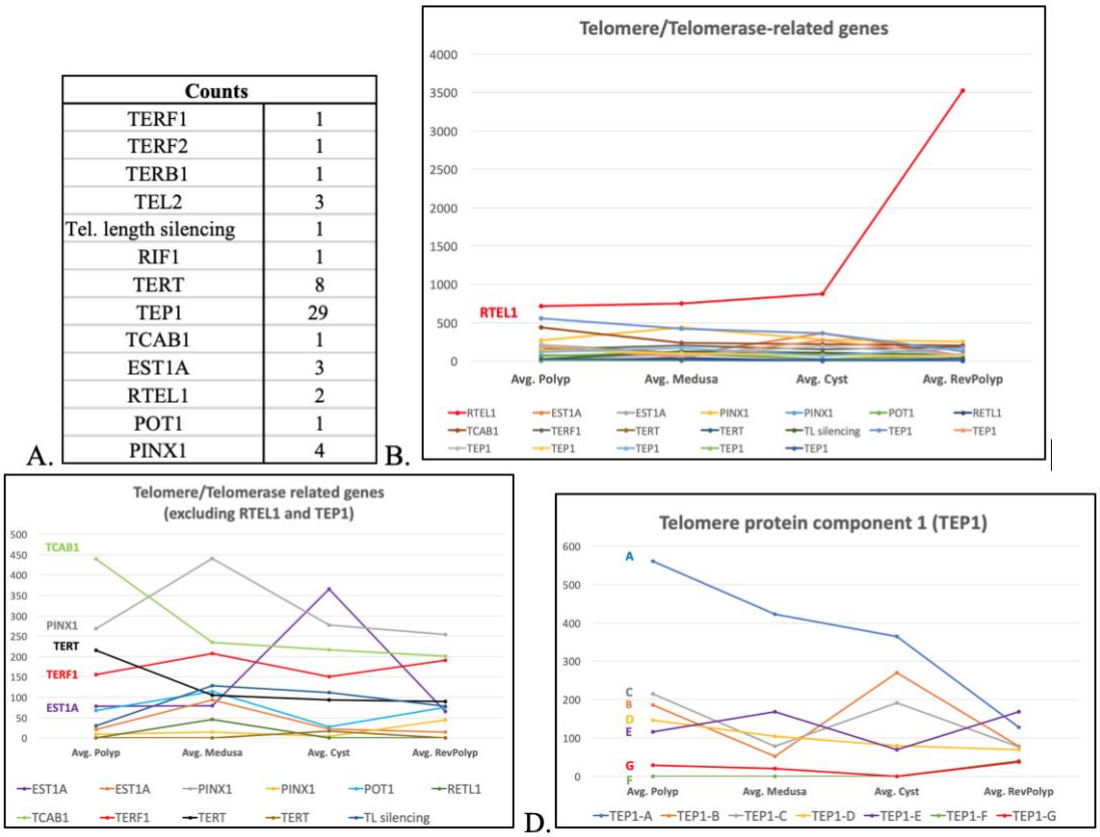


Figure III.2: Telomere/Telomerase-related genes in *T. dohrnii*. A) Number of super-transcripts for each telomere/telomerase-related sequence; B) Expression profiles of all putative homologs (most consistent expression patterns across replicates chosen; complete list in Appendix III.B); C) Expression profiles excluding the RTEL1 and TEP1 sequences; D) Expression profiles of all TEP1 sequences.

Heat-shock protein 70 and 90

HSPs are activated in response to internal and external cellular stress and shock, and promotes the quality control of protein folding processing, preventing misfolding

and repairing or disposing of impaired proteins (i.e. autophagy) (Tower 2011). These molecular chaperones are ubiquitous and present in all domains of life, and have been extensively manipulated in invertebrate models to increase lifespan (Dubrez et al. 2019). As animals age, the production of HSPs and other stress-resistance factors slows down (Hou et al. 2010). The experimental induction of HSPs increased longevity in aged subjects, as the proteins counteracted the accumulation of cellular damage that leads to physiological degeneration, malfunction and disease (Murshid et al. 2013). Within the family, HSP90 and HSP70 are the central facilitators of relieving cellular stress and shock with the aid of smaller HSPs (e.g. HSP27) (Calderwood 2007).

79 HSP70 and 49 HSP90 annotated super-transcripts were found in *T. dohrnii,* with e-values ranging from e-^{06}– 0 (Appendix III.C). Interestingly, *Hydra* was not the species with the most top-hits, and instead belonged to *Seriola lalandi dorsalis*, a ray-finned fish (Figure III.3A). There were more chordate taxa, specifically fish, than cnidarians that correlated to top-hits in HSP putative homologs (Figure III.3B). This may indicate that *T. dohrnii* utilizes some HSPs similar to vertebrate taxa that other Cnidaria do not possess and could be unique to the species.

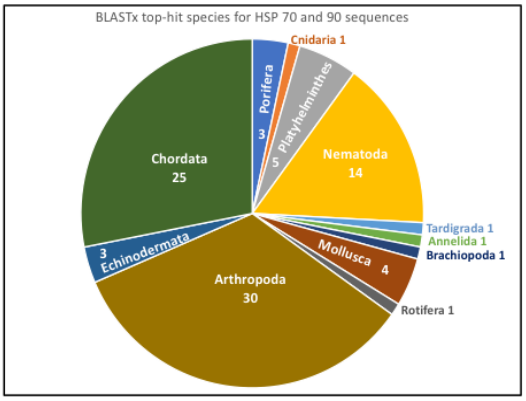
HSP90 and HSP70 sequences were both found highly active in all stages of *T. dohrnii* (Figure III.3). HSP90 top-blast hit belonged to *Hydra* while HSP70 belonged to *Rattus norvegicus*, the brown rat (Appendix III.C). The most active HSP90 peaked at the cyst before dropping back down in the reversed polyp (Figure III.3C, blue). The most active HSP70 had a peak in the medusa, and decreased in activity in the cyst, and in the reversed polyp (Figure III.3C, orange). In addition, three HSP90 super-transcripts were

absent or minimally expressed among all stages except for the reversed polyp (Figure III.3D). The other HSP70 sequences varied in expression patterns, some having a peak in activity at the cyst (Figure III.3D black, light blue), and some in the medusa stage (Figure III.3D light green, dark blue). HSP90 acts as a suppressor of the ubiquitin proteasome system (UPS), while other HPS70 sequence found to be most active in the medusa and cyst stage (Figure III.3E), acts as an activator of the UPS (Pratt et al. 2010). The two HSPs (HPS90 and HPS70) thus cooperate to regulate the UPS, where damaged or misfolded proteins are tagged for proteasomal degradation through ubiquitination (Löw 2011). The UPS plays a large role in slowing down aging, combating stress and damage, and promoting cellular/tissue regeneration in various metazoans (Löw 2011). Additionally, ubiquitin-related genes also were found to be differentially expressed and enriched in the cyst in a previous transcriptomic study of reverse development in *T. dohrnii* (Matsumoto and Miglietta 2021). The coordinated action of HSP70 and HSP90 may be counteracting the cellular stress afflicted to the medusa that triggered ontogeny reversal and promoting the regeneration of polyp structures.

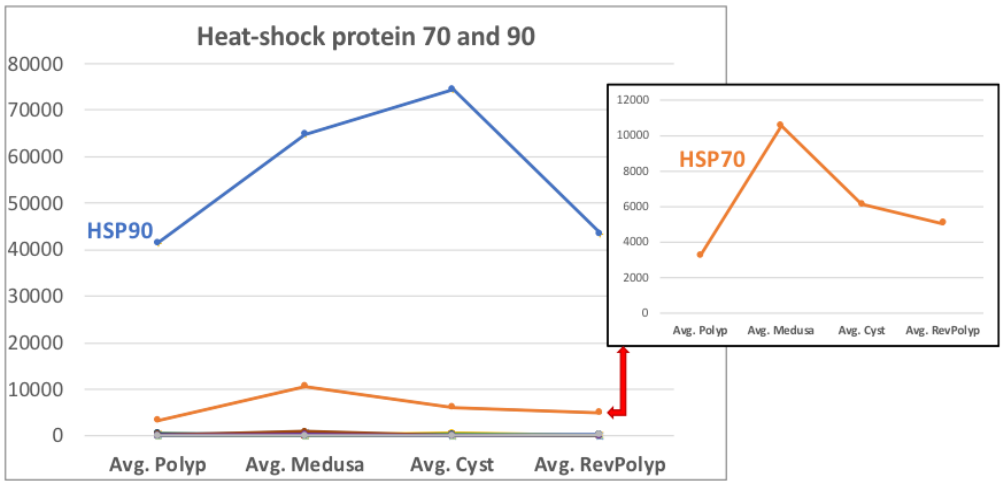
A.

Species	Phylum	Top Hits
Seriola lalandi dorsalis	Chordata	10
Hydra vulgaris	Cnidaria	7
Cotesia chilonis	Arthropoda	7
Siniperca chuatsi	Chordata	6
Mus musculus	Chordata	3
Piliocolobus tephrosceles	Chordata	3
Trichinella patagoniensis	Nematoda	3
Lingula anatina	Brachiopoda	2
Anabas testudineus	Chordata	2
Oncorhynchus tshawytscha	Chordata	2
Botryllus schlosseri	Chordata	2
Trichuris trichiura	Nematoda	2
Crassostrea virginica	Mollusca	2
Crassostrea gigas	Mollusca	2
Colinus virginianus	Chordata	2
Other (1 top hit)		74

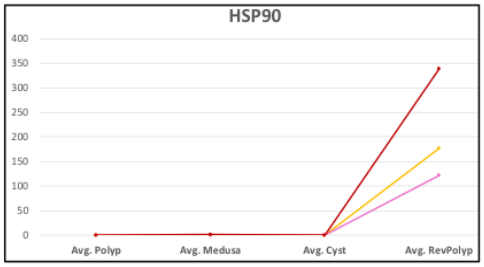
B.



C.



D.



E.

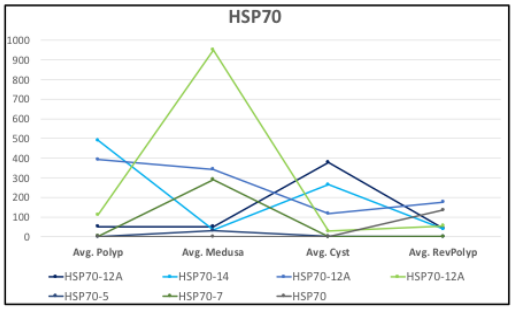


Figure III.3: HSP70 and 90 in *T. dohrnii*. A) Number of BLASTx top-hits for each species and their corresponding phylum (all species with only one hit grouped as 'Other'); B) Pie chart of top-species for HSP70 and 90 annotated sequences; C) Expression profiles of highest expressed HSP70 and 90 (most consistent expression patterns across replicates chosen; complete list in Appendix III.C); D) Expression profiles of HSP90 found only active in the reversed polyp; E) Expression profiles of other HSP70 found among *T. dohrnii* stages.

Yamanaka transcription factors

Oct4 (POU5), Sox2, Klf4 and c-Myc, known as the Yamanaka Transcription Factors have been shown to induce pluripotency in mammals (Takahashi et al. 2007; Takahashi and Yamanaka 2006), with POU5 and Sox2 considered indispensable factors that work cooperatively to induce the cellular pluripotency, and Klf4 and c-Myc considered interchangeable (Takahashi et al. 2007; Takahashi and Yamanaka 2006). The gene families that each of the Yamanaka Factors reside in, namely POU, Sox, Klf and Myc, have various roles in maintaining cellular pluripotency and development in metazoans, often modulated during disease and implicated in biomedical and regenerative research (Lefebvre et al. 2007; Gold et al. 2014; Zhang et al. 2020; Dang 1999). POU5, which Oct4 resides in, has only been reported in vertebrates, whereas other POU classes have undergone diversification prior to the emergence of the eumetazoa (Gold et al. 2014). Sox2 and c-Myc on the other hand, are highly conserved among Metazoa and found universally among both vertebrate and invertebrate taxa (Kondoh and Lovell-Badge 2015; Sarid et al. 1987). Klf4 has yet to be reported in Cnidaria (Steele et al. 2011), but have been found among other invertebrates such as roundworms (Ma et al. 2014; Hsieh et al. 2017).

We first conducted BLAST and alignment-based analyses on the annotated transcriptome and RNA-seq libraries of *T. dohrnii* published in (Matsumoto and Miglietta 2021) (summary in Table III.1, full details of analyses in Appendix III.D). We also included the Thompson Factors (Lin28, Nanog), which are replaceable factors of Klf4 and c-Myc (Yu et al. 2007) to induce pluripotency in mammalian cells. This is the

first step to assess whether the genetic network that enables cellular transdifferentiation in *T. dohrnii* involves pathways used in mammalian stem-cells. Evidence from the *Hydra* genome has shown that c-Myc and Sox2 are the only factors present in this widely used model system (Chapman et al. 2010), bringing the authors to conclude that the stem cell genetic network in *Hydra* probably has an evolutionary origin independent from the network used in mammalian stem cells. Understanding which of these factors are present in *T. dohrnii* is important as it may help us clarify the relationship between cnidarian stem cells and those of other animals, evolution of pluripotency induction in Metazoa, and indicate to which extent *T. dohrnii* can be used as a live model system to study cellular plasticity and pluripotency, at large.

Table III.1: Summary of Yamanaka and Thomson transcription factors homology-based screening on published annotated transcriptome and RNA-seq libraries (Matsumoto and Miglietta, 2019). [*BLAST hit present under <400bps]; SF= Similarity fraction (minimum fraction of sequence identity between the read and reference), LF= length fraction (minimum length fraction that must match the reference sequence). Full details of analyses found in Appendix D.

Factor	tBLASTx (e-value: ≤ 1E-8)	Annotation	RNA-seq mapping analyses			
			SF/LF (0.3)	SF/LF (0.5)	SF/LF (0.7)	SF/LF (0.9)
Oct4	Present (10E-32)	Absent	Present	Present	Present	Present
Sox2	Present (4.7E-42)	Present	Present	Present	Present	Absent
Klf4	Present (6.6E-48)	Absent	Present	Present	Absent	Absent
c-Myc	Present (7.6E-21)	Present	Present	Present	Absent	Absent
Lin28	Absent *	Absent	Present	Present	Present	Present
Nanog	Absent *	Absent	Present	Present	Present	Present

Inference of homology using tBLASTx

We used all assembled transcripts that were filtered for biological contaminants (Matsumoto and Miglietta, 2021). All 12 query sequences, which included all variants of

the six transcription factors reported in GenBank, had hits with regions of high scoring pairs (Table III.1; Appendix III.D). Only transcripts larger than 400 bp were annotated in Matsumoto and Miglietta (2021). Annotation results for tBLASTx hits against Oct4, Sox2, c-Myc, and Klf4 are reported below, while Lin28 and Nanog hits were shorter than 400bp and thus had no annotation (Appendix III.D; Matsumoto and Miglietta, 2021).

The presumed Oct4 homolog in *T. dohrnii* (Td_DN89582_c0_g1_i1) identified in the reciprocal tBLASTx analyses above was annotated (i.e., protein assignment by BLASTx) as 'Brain-specific homeobox/POU domain protein 3' with a e-value of $1.34e^{-177}$ (Appendix III.D). The specific POU domain class 5 transcription factor was not found in the protein annotations. The presumed Sox2 homolog in *T. dohrnii* (Td_DN102764_c1_g1_i4) was annotated as the same protein, 'Sex determining region Y-box 2 protein' with a e-value of $6.39e^{-128}$ (Appendix III.D). The presumed c-Myc (Myc proto-oncogene protein) homolog in *T. dohrnii* (Td_DN99862_c1_g1_i1) was also annotated as the same protein 'Myc proto-oncogene protein', with a e-value of $2.55e^{-23}$ (Appendix III.D). The presumed Klf4 homolog in *T. dohrnii* (Td_DN110215_c0_g1_i7) was annotated as 'Krueppel-like factor 5' with a e-value of 3.48^{e-47} (Appendix III.D). The specific Klf4 factor was not found in among the annotations.

RNA-seq alignment analysis

We conducted RNA-seq mapping analyses using a range of length and similarity fractions (i.e., level of stringency) from 0.3 (least stringent) to 0.9 (most stringent). Read mapping of queried factors with lower stringency was performed due to the evolutionary

distance between hydrozoans and humans. With a low-level stringency of a length fraction and similarity fraction of 0.3, the presence of all factors was recovered, and the consensus length of all of reads mapped to the query was similar to the actual length of all factors, merely 1-4 bp difference, with the exception of c-Myc with a 769 bp disparity (Appendix III.D). All factors were present with the stringency parameter of 0.3 and 0.5, but c-Myc and Klf4 were not present when the stringency parameter was set to 0.7. When a length and similarity fraction was set to 0.9, reads that map only to Oct4, Lin 28 and Nanog were recovered (Appendix III.D).

Expression profiling

Next, we identified thirty super-transcripts annotated as POU, Sox, Klf, and Myc among the super-transcripts in *T. dohrnii* with e-values e^{-08}– e^{-168} (Figure III.4A, Appendix III.E). The majority of top-hits belonged to cnidarian taxa (Appendix III.E).

We found POU3F3 and POU4F3 active in *T. dohrnii*, with the POU3F3 peaking in the medusa and cyst, and the POU4F3 homolog enriched at the medusa and in the reversed polyp (Figure III.4B, red/purple). POU3F3 is involved in neuronal development and works synergistically with Sox factors in mammals (Kuhlbrodt et al. 1998), and POU4F3 is a regulator of cell identity in the nervous system and its dysregulation is a central to hearing loss in mammals (Clough et al. 2004; Hertzano et al. 2004).

Genes in the Sox gene family are involved in cell fate determination during development (Lefebvre et al. 2007). Invertebrates have been reported to only possess a single Sox gene that is representative of the numerous Sox groups found in vertebrates (Bowles et al. 2000). The analyses however, only consisted of *Drosophila*, roundworms

and an incomplete genome of the sea urchin, and thus, may represent a partial view of the evolutionary history of the Sox in invertebrates (Bowles et al. 2000). In *T. dohrnii,* we found Sox2 and Sox10-like putative homologs (Figure III.4). Sox2-like homologs varied in expression profile (Figure III.4CD), and the Sox2 factor was identified in the BLAST homology and alignment-based analyses (Table III.1; Appendix III.D). Activity of Sox10-like was very low among all stages (Figure III.4A, Appendix III.E).

We found homologs of Klf1, 7, 11 and 13 active during *T. dohrnii*'s reversal (Figure III.4). Klf11 was the most active, particularly in polyp stages (Figure III.4C, yellow) and Klf1, 7 and 13 all increased from the medusa stage, peaked at the cyst, and declined after reversal (Figure III.4D, yellow/gray). Klf factors are highly associated to the regulation of the cell cycle in mammals (Zhand et al., 2020; Tetreault et al., 2013). Similarly, transcripts associated with mitotic cell cycle regulation were found to be highly regulated at the cyst stage in previous transcriptomic studies (Matsumoto et al. 2019; Matsumoto and Miglietta In review). In nematodes, Klf factors (specifically Klf1, 2 and 3) are requirements for lifespan extension, where its deficiency reduces lifespan and its over-expression increases longevity (Hsieh et al. 2017).

c-Myc was found active during *T. dohrnii*'s reverse development, along with its binding protein MBP, which stimulates the activation of Myc (Taira et al. 1998) (Figure III.4D). c-Myc is a proto-oncogene with diverse roles that includes regulating the cell cycle, mediation of cell lifespan, and maintenance of pluripotent cell identity in mammals (Miller et al. 2012; Chappell and Dalton 2013). Both c-Myc homologs found in *T. dohrnii* had a gradual increase in expression through reverse development and

peaked at the cyst and reversed polyp stages (Figure III.4D, green/red). c-Myc is one of the Yamanaka Transcription Factors and was also identified using BLAST homology and alignment-based approaches (Table III.1, Appendix III.D).

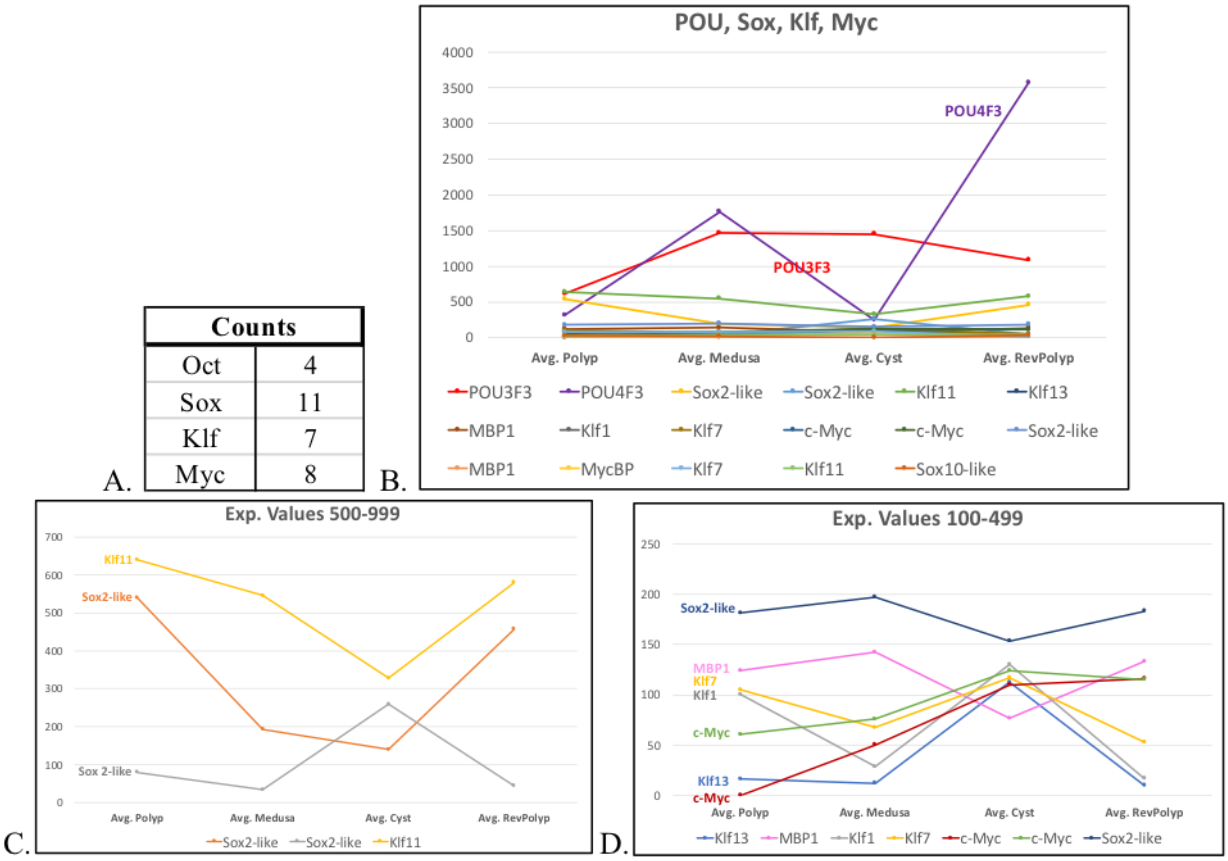


Figure III.4: POU, Sox, Klf, and Myc in *T. dohrnii*. A) Number of super-transcripts for each gene family; B) Expression profiles of highest expressed POU, Sox, Klf, and Myc homologs (most consistent expression patterns across replicates chosen; complete list in Appendix III.E); C) Expression profiles of homologs in the normalized expression value range 500-999; D) Expression profiles of homologs in the normalized expression value range 100-499.

Evolutionary gene tree analyses

We used a phylogenetic approach to confirm the homology or infer the evolutionary history of SIRT3, HSP90 and POU3/5 in *T. dohrnii*, three genes that

showed high activity during the reverse development in our profiling analyses. Phylogenetic hypotheses were constructed using Bayesian and Maximum Likelihood (ML) methods. Placozoa sequences were utilized as the outgroup/root for all three gene trees instead of Porifera (Nosenko et al. 2013), as the latter phyla did not have homologous representatives of SIRT3 and POU3. Details for each alignment and tree reconstruction can be found in the methods.

SIRT3- The SIRT3 dataset consisted of 42 sequences with an alignment length of 284bp and contained 197 phylogenetic informative sites (69.37% of alignment). Bayesian inference (MrBayes) and Maximum Likelihood (PhyML) phylogenetic trees were congruent in showing five major clades, namely Vertebrata, Playhelminthes, Mollusca, Cephalo/Hemichordata/Echinodermata and Cnidaria (Figure III.5). The Cnidaria clade includes only Anthozoa representatives, as SIRT3 has yet to be reported among other classes within the phyla. The *T. dohrnii* SIRT 3 sequence did not form a clade with the rest of Cnidaria (Figure III.5, purple), but resulted as sister to the rest of the Metazoa, indicating SIRT3 in *T. dohrnii* may be the outcome of a gene duplication event.

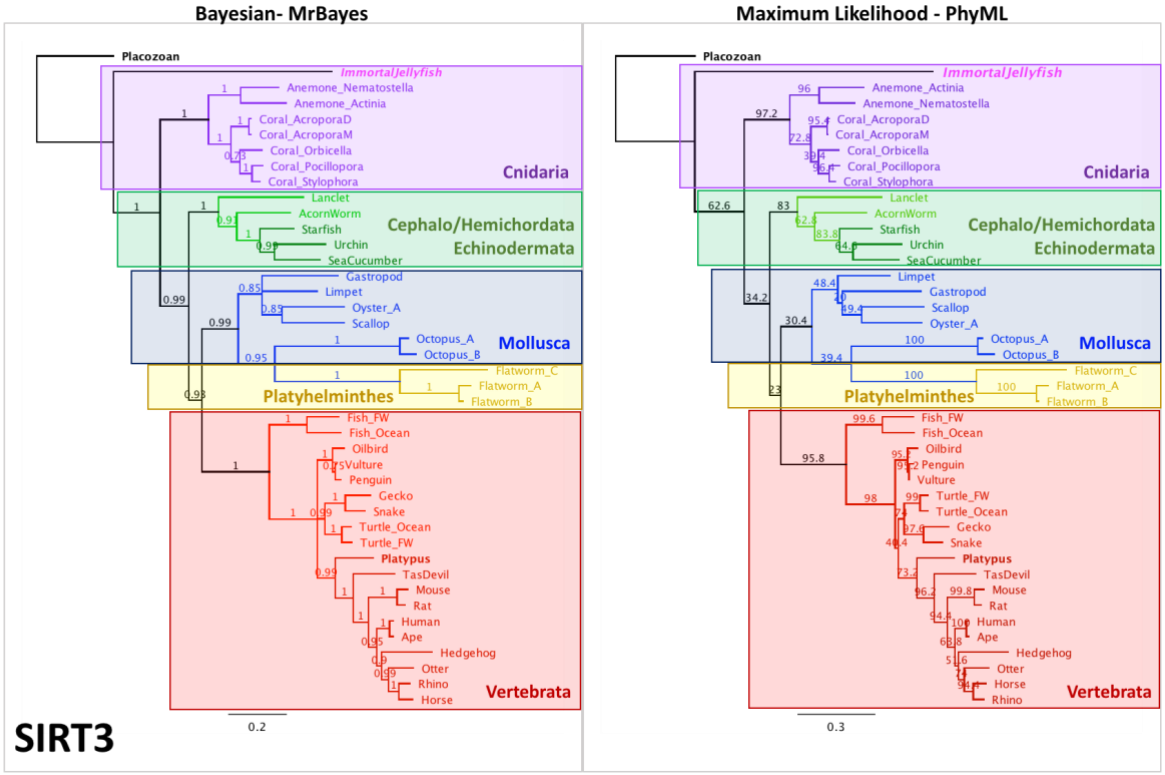


Figure III.5: Evolutionary gene trees of SIRT3. The evolutionary history of SIRT3 in *T. dohrnii* was inferred using Maximum Likelihood (PhyML) and Bayesian inference (MrBayes).

HSP90- The HSP90 dataset consisted of 43 sequences with an alignment length of 693bp and contained 346 phylogenetic informative sites (49.93% of alignment). Both Bayesian and ML were congruent in showing the *T. dohrnii* HSP90 sequence fall within the Cnidaria, and sister to the hydrozoan *Hydra* (Figure III.6, purple).

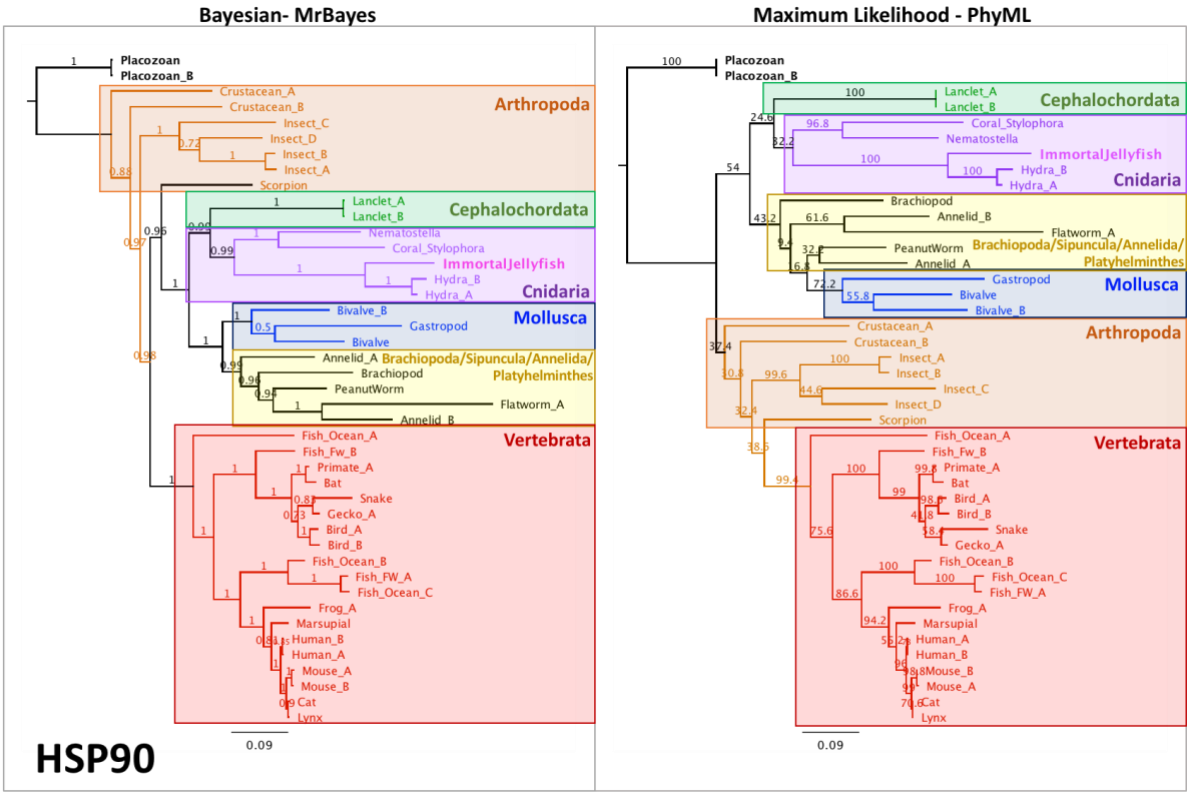


Figure III.6: Evolutionary gene trees of HSP90. The evolutionary history of HSP90 in *T. dohrnii* was inferred using Maximum Likelihood (PhyML) and Bayesian inference (MrBayes).

POU factors- The POU dataset consisted of 42 sequences with an alignment length of 108 bp and contained 57 phylogenetic informative sites (52.78% of alignment). Although members of the POU3 family have been found in invertebrates and POU5 have not, the similarity between the two families and their evolutionary history has been debated in Hydrozoa (Millane et al. 2011). We used POU5F1 (i.e., Oct4) and POU3F3 sequences (Figure III.7, green) to determine the evolutionary relationship between *T. dohrnii*'s POU3 factor with those of other Metazoa and vertebrate. Phylogenetic trees constructed using ML and Bayesian methods resulted in congruent reconstruction

(Figure 7). POU5F1 (from Vertebrates) forms a monophyletic clade sister to a clade that contain POU3F3 sequences from Cnidaria, including those from *T. dohrnii*. More specifically, *T. dohrnii*'s sequence is sister to the POU3F3 factor (refered to as Polynem; GenBank Accession: AEG66930.1) from the closely related *Hydractinia*.

A third clade, containing POU3 from Vertebrates and Invertebrates (annelids, echinoderms, nematodes, mollusks, and arthropods) is sister to the Cnidaria POU3+Vertebrata POU5) (Figure III.7, red). Though more hydrozoan sequences are needed to further resolve the evolutionary history of the POU3/Polynem factor, its inferred from the gene trees that cnidarian POU3 factors show more similarity to POU5 than other metazoan POU3 factors, while the POU3F3/Polynem gene in *T. dohrnii* and *Hydractinia,* and possibly the coral *Stylophora*, have diverged within the clade from their other cnidarian counterparts and show intriguing similarity in their evolutionary path as the vertebrate POU5 factor (Figure III.7).

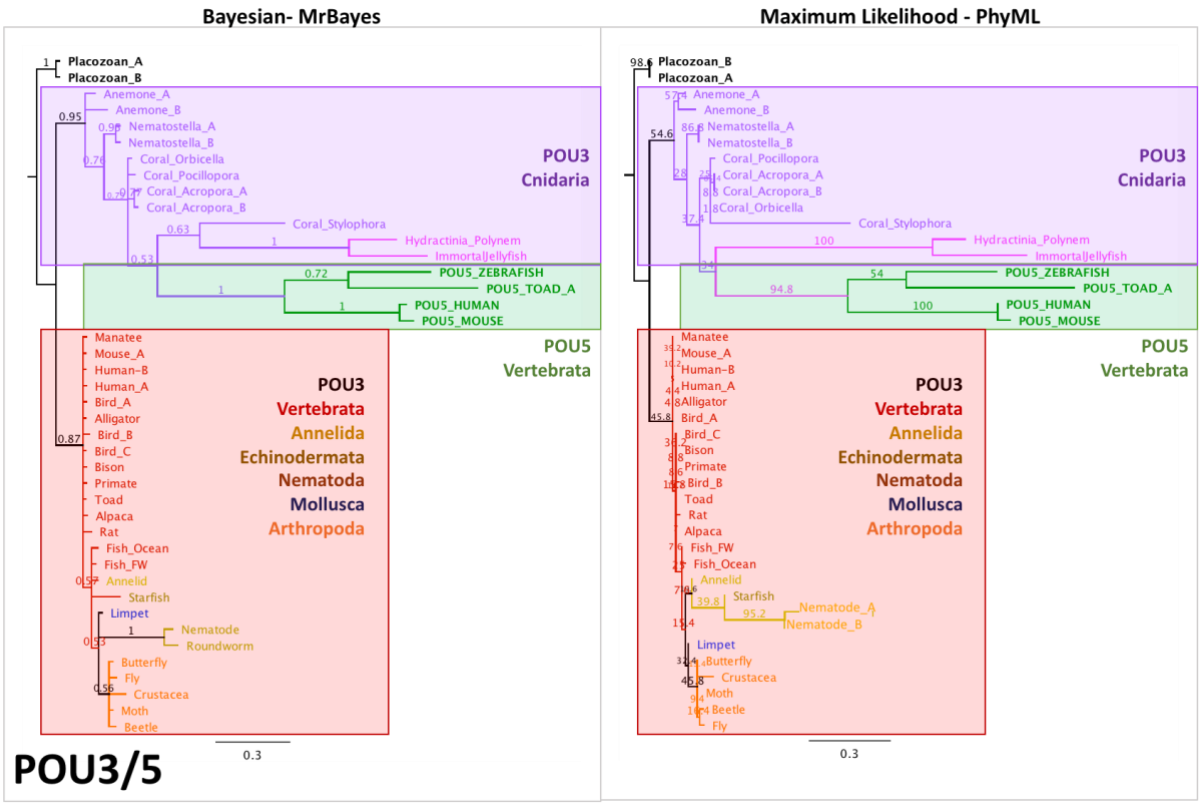


Figure III.7: Evolutionary gene trees of the POU domain. The evolutionary history of the POU in *T. dohrnii* was inferred using Maximum Likelihood (PhyML) and Bayesian inference (MrBayes).

III.4. Conclusion

T. dohrnii offers an unparalleled experimental paradigm to investigate the changes in gene activity that occur during reverse development that induce transdifferentiation, subsequently controlling the directionality of ontogeny, *in vivo*. We profile expressional changes of putative gene homologs recognized as crucial genetic regulators of tissue repair and regeneration, inductors of the pluripotent state, and factors that promote longevity and increase lifespan, during *T. dohrnii*'s reverse development sequence. Our work displays that *T. dohrnii* manipulates genetic networks of high

relevance in biomedical studies in mammals, such as SIRT3, POU factors, RTEL1 and HSP70/90, showcasing the species as a useful research system that can contribute to the further understanding of genetic networks underlying regeneration, pluripotency, and longevity in Metazoa. Furthermore, our analyses indicate that the Yamanaka factors' homologs may be present in *T. dohrnii*. They thus challenge our understanding of how the networks controlling a pluripotent cell state in Cnidaria relate to those of mammals. Albeit not exhaustive, our analyses show a complex scenario and an area in need of further exploration. They also highlight the need to expand the research on the genetics of induction of pluripotency to a variety of species that go beyond classically used model systems.

CHAPTER IV

DRAFT GENOME ASSEMBLY OF *Turritopsis dohrnii* (CNIDARIA, HYDROZOA), FROM BOCAS DEL TORO, PANAMA

IV.1. Introduction

Cnidarians are considered as one of the oldest lineages in the metazoan radiation, phylogenetically placed as the sister-taxa to bilaterians (Gold et al. 2019). The body composition of cnidarians contains much fewer tissue types and cells in comparison to bilaterians (Steele et al. 2011). Despite displaying relatively simple body plans, cnidarians possess a diverse array of life history traits and regeneration strategies that have been recognized in evolutionary, developmental, and medical research sectors (Putnam et al. 2007; Chapman et al. 2010; Fuchs et al. 2014). Comparative genomics of different cnidarians systems can provide insight into how complex life cycles and regenerative capabilities are reflected in the genomic structure, composition, and activity.

Comparative functional genomics can help connect genome composition and structure to the morphology and functionality of an organism, to infer the evolution of inherited traits of different taxa (Koonin et al. 2000; Rubin et al. 2000; Zhang et al. 2014; Gold et al. 2019). The most recognized invertebrate model systems, *C. elegans* and *D. melanogaster*, have provided tremendous knowledge in various research disciplines, such as aging, tissue regeneration and evolutionary development, but have undergone broad divergence and genomic rearrangement from the most recent ancestor

with vertebrates (Kortschak et al. 2003; Chapman et al. 2010; Bosch et al. 2017; Murthy and Ram 2015). Though the divergence of cnidarians from vertebrates occurred earlier than nematodes and fruit flies, many inherited features of the genome have been reported to be conserved in the most documented cnidarian model species, *Hydra vulgaris* and *Nematostella vectensis* (Putnam et al. 2007; Chapman et al. 2010; Martínez and Bridge 2012; Sullivan and Finnerty 2007). Consequently, cnidarians represent a promising group to further examine a variety of cellular and developmental processes in metazoans.

Members of the class Hydrozoa within phylum Cnidaria undergo drastic changes in body form and behavior during their lifecycle, which implies that a single genome has the capacity and flexibility to encode for a variety of developmental forms during its ontogenetic sequence. Unlike most other Hydrozoa, when facing death *Turritopsis dohrnii* (Weismann 1883) can reverse develop into an earlier life cycle stage (juvenile polyp stage) through cellular transdifferentiation (Schmich et al. 2007; Piraino et al. 1996; Miglietta et al. 2018; Matsumoto et al. 2019). Moreover, *T. dohrnii* possesses multiple sister species that do not have the capability to undergo reverse development (Miglietta et al. 2018), providing potentially a significant comparative paradigm to comprehensively assess the differences in genetic networks that underlie the capability to be immortal in *Turritopsis.*

The availability of a genome for *T. dohrnii* would enable comparison with other metazoan taxa to further our understanding of the genetic basis of its capability for ontogeny reversal. We present our efforts towards generating the first draft genome assemblies for *T. dohrnii*, the Immortal Jellyfish. Our work exemplifies the complexity

of genome construction in the species and initiates the first crucial steps towards developing *T. dohrnii* into a purposeful genomic research system.

IV.2. Methods

Specimen collection, rearing and identification

A *T. dohrnii* colony was collected in Bocas del Toro, Panama (Atlantic) in July 2015. The individual colony was rinsed with filtered seawater to reduce the number of fouling species, kept in glassware and starved for more than 24 hours. The colony was preserved in RNAlater stabilization solution (ThermoFischer Scientific) and stored in -80C for further processing.

Before proceeding to high-molecular weight (HMW) DNA extraction, DNA was extracted from spare tissue from the colony using a method provided by Zietara et al. (2000). Proper purity and concentration of the extracted DNA was verified using the Thermo Scientific NanoDrop2000 Spectrophotometer. A fragment of the mitochondrial 16S gene was amplified using forward SHA and reverse SHB primers [Forward (SHA): 5'-TCGACTGTTTACCAAAAACATAGC-3', Reverse (SHB): 5'-ACGGAATGAACTCAAATCATGTAAG-3']. PCR products were digested and visualized through gel electrophoresis, treated with ExoSAP to remove contaminants, and sequenced at the Texas A&M University in Corpus Christi's Genomic Core Lab to confirm the species identification of the samples as *T. dohrnii*. The 16S sequences were uploaded to GenBank under accession #MH029866.

Genomic DNA extraction/library creation/DNA-sequencing

DNA extraction, library creation and sequencing were performed at the Cold Spring Harbor Lab genomic facilities in New York. HMW genomic DNA was extracted from hydranths of a single colony (polyp stage has the largest amount of tissue due to

multiple clonal individuals) using the MagAttract Kit (Cat.# 67563, Qiagen). The quality and quantity of the extracted DNA was assessed using the Agilent Femto Pulse Run (Agilent Technologies Inc.) and a loading concentration of approximately 90-70pM of HMW genomic DNA was used to build two DNA libraries with a 9-13kb fragment size selection. Three cells in the Pacific Biosciences (PacBio) Sequel II sequencing platform were used for the initial library, and one cell was used for the second library. High-fidelity (HiFi) read were produced from the raw subreads using consensus circular sequencing (CCS).

Genome assembly

CCS (HiFi) Assemblies

The following algorithms/software were utilized to assemble the CCS reads using the following parameters:

1) HiCanu (Nurk et al. 2020):
    a. Trial 1- estimated genome size= 385 Mb, stop on low coverage =10
    b. Trial 2- estimated genome size= 500 Mb, stop on low coverage = 4
2) Flye (Kolmogorov et al. 2019):
    a. Trial 1- HiFi error rate = 0.001, estimated genome size = 500 Mb, polishing iterations = 3, and minimum overlap length = 5000 (computed as automatic)
    b. Trial 2- HiFi error rate = 0.01, estimated genome size = 385 Mb, polishing iterations= 3, and minimum overlap length = 2500
3) SPAdes (Bankevich et al. 2012):

a. Trial 1- Coverage cutoff= off, K-mer sizes = 21, 33, 55, phred offset = 33, mismatch careful mode = on
b. Trial 2- Coverage cutoff = 6, K-mer sizes= 21, 33, 55, phred offset =33, mismatch careful mode = on

4) Improved Phased Assembly (IPA) HiFi Genome Assembler: genome size = 0 (down-sampling turned off), coverage = 0 (down-sampling turned off), polishing run = 1.

BUSCO v5.1.2 (Simão et al. 2015) was utilized to evaluate the completeness of the newly scaffolded genome in terms of gene content and fragmentation. The Metazoa database was downloaded from the software website (http://busco.ezlab.org), and the assessment was performed using an expectation value cutoff of 10^{-3}. The assembly with the higher BUSCO scores from each of the trials were presented in Table IV.3 of this dissertation chapter. The N50 values and BUSCO scores for all seven assemblies can be found in Appendix IV.A.

Using N50 length statistics and BUSCO genome completeness analyses, the following statistics were used to find the best genome for subsequent processing: a) most contiguous genome (i.e., fewest number of contigs per overall assembly size); b) within target range of the genome size (genome estimated ~383Mbp); c) highest number of complete and fragmented BUSCO statistic; d) close to estimated GC% of ~37-41% (Matsumoto et al. 2019; Matsumoto and Miglietta, 2021). The chosen genome for further quality analysis (IPA genome assembly) was scaffolded using the LRScaf 1.1.10 software (Qin et al. 2019) using extracted reads that did not pass the previous CCS

process with the following parameters: identity= 0.15, minimum support links = 3, minimum overlap length= 500, minimum ratio of 0.0001. Gap filling was performed using TGS-GapCloser (Xu et al. 2020) using reads that did not pass CCS, with two rounds of polishing using Racon (Vaser et al. 2017).

Microbial contamination of the scaffolds was assessed using the Kraken 2.0.9 taxonomic classifier (Wood and Salzberg 2014) using the RefSeq 2020.07 database (includes archaea, bacteria, fungi, protozoa, viruses, *Homo sapiens, Mus musculus, Rattus norvegicus, Bos taurus, Sus scrofa, Drosophila melanogaster, Arabidopsis thaliana*) using a confidence filter of 0 and BLASTn using NCBI's Non-Redundant Nucleotide (NT) database with an e-value of 10^{-6}.

CLR Assemblies

The raw sequencing subreads (i.e., Continuous long read (CLR)) generated from four cells of the PacBio Sequel II platform were corrected and trimmed using Canu's correction software (Koren et al. 2017), utilizing 40x coverage of the longest subreads. The correction step was performed on all cells combined, as well as individually for Cell #1, #2 and #3 for assembly downstream.

Contamination filtering pre-assembly

The following steps were performed to filter microbial contamination from the corrected dataset of all cells combined:

Step 1) Filtering was performed using Kraken 2 (Wood and Salzberg 2014) and the RefSeq 2020.07 database, and the corrected reads that were unclassified (i.e., unassigned) or classified to originate from eukaryotes were kept.

Step 2) Corrected reads that were classified as bacterial, archaeal, viral, protozoan, and fungal sources were divided into sub-datasets of 200,000 sequences and re-analyzed using BLASTn against NCBI's NT database using a e-value of 10^{-6}, and corrected reads that did not have BLAST hits were kept (Appendix IV.B).

Step 3) Corrected reads with BLAST hits against the NT database were re-analyzed using the NT Metazoa (taxids: 33208) database using a e-value of 10^{-6}. Contigs with hits against the Metazoa database were kept, and contigs without any hits were marked as microbial contamination (Appendix IV.C).

Step 4) Among corrected reads that were marked as contamination, sequences that had hits greater than 90% identity or an e-value of 0.0 were discarded, and the rest of the reads were kept (Appendix IV.D).

Two datasets were created for assembly: A) Assembly dataset (AD)- includes steps above (steps 1 through 4); B) Stringent Assembly dataset (ST)- includes steps 1-3 and excludes all corrected reads that were marked as microbial contamination (Step 4).

Assembly

Genome assembly was performed on the AD dataset using two different assemblers:

1) Canu (Koren et al. 2017) with parameters of estimated genome size set as 500 Mb and stopping assembly on low coverage reads as 10x; 2) Hifiasm (Cheng et al. 2021) using error correction iterations set to 0.

Only the Canu assembler was utilized to assemble the ST dataset and individually corrected PacBio cells (Cell #1, #2, and #4). All six assemblies were

assessed for genome completeness and duplication using BUSCO v5.1.2 (Simão et al. 2015) with the Metazoa database. The ST assembly and Cell #1 (C1) assembly were chosen for further processing based on genome size and BUSCO completeness scores.

Contamination filtering post-assembly

Kraken 2 (Wood and Salzberg 2014) and BLASTn was used to filter contamination post-assembly from the ST and C1 contigs (see contamination filtering pre-assembly steps 1- steps 3), and BUSCO (Simão et al. 2015) with the Metazoa database was used to assess both filtering approaches. After assessment, the ST Kraken (STK) assembly was produced for further processing downstream along with the ST and C1 assembly.

Genome Scaffolding

Scaffolding with genomic sequences

Scaffolding with genomic sequences were performed on the three assemblies (ST, C1 and STK assemblies) using LRScaf 1.1.10 (Qin et al. 2019) based on Minimap2 (Li 2018) alignments. The Error-corrected AD dataset was used for the first scaffolding step. Reads that were less than 10,000 bp were first removed from the AD dataset. Two trials of scaffolding were performed with the following parameters: Trial A) identity= 0.15, minimum support links = 3, minimum overlap length= 500, minimum overlap ratio = 0.0001, and the rest of the parameters as default; Trial B) identity= 0.1, minimum support links = 1, minimum overlap length= 160, and the rest of the parameters on default. The parameters with the higher number of scaffolded contigs (Trial B) was used for the remainder of the iterations and presented in this dissertation chapter (Table IV.5).

The IPA assembly constructed from CCS reads were used for the second scaffolding step, and the SPAdes assembly also constructed from CCS reads were used for the third scaffolding step. Lastly, an additional scaffolding step (Step 4) was included for only the STK assembly using the CLR assemblies of Cell #1, Cell #2, and Cell #4.

Scaffolding with transcriptomic sequences

Scaffolding using transcriptomic data (RNA-seq libraires, assembled transcripts) were also performed on all three assemblies. First, the RNA-seq libraries of the colonial polyp, medusa, cyst and reversed polyp stage (under NCBI BioSample Accession #SAMN13924705- SAMN13924707, SAMN12669945) using P_RNA_scaffolder (Zhu et al. 2018). Two trials were performed based on different alignment software, HISAT2 (Kim et al. 2019) and STAR (Dobin et al. 2013). The aligner with the higher mapping percentages (HISAT2) was used for the scaffolding step and presented in this dissertation chapter (Table IV.6).

Next, the assembled transcripts from Matsumoto and Miglietta (2021) were used with L_RNA_scaffolder (Xue et al. 2013) based on the BLAT aligner (Kent 2002). Two different trials using different parameters were performed. For both trials, the parameters were the following: Alignment length coverage= 0.95, threshold of alignment identity = 0.9, and minimal number of supporting reads =1. For Trial 1, the maximal intron length between two exons was set to 100 kb (as recommended for vertebrates), and for Trial 2 was set as 15kb (as recommended for invertebrates). As both trials produced the same percentage of scaffolded contigs, the output from Trial 2 was utilized for further processing.

Gap-filling

TGS-GapCloser software (Xu et al. 2020) was used to fill the gaps in our scaffolded genomes. Four gap-filling steps were performed using the following datasets: 1) Error corrected-reads from all four cells; 2) Raw CCS reads; 3) IPA assembly generated from CCS reads; and 4) SPAdes assembly generated from CCS reads. Racon-polishing (Vaser et al. 2017) was not utilized with the TGS-GapCloser software as the all reads used in this step were already error-corrected or from CCS reads.

IV.3. Results and Discussion

Isolation of HMW genomic DNA

HMW genomic DNA was extracted from a single polyp colony, barcoded and confirmed as *T. dohrnii* (GenBank Accession #MH029866). Although the medusa stage is the least likely to possess biological contaminants due to its planktonic behavior, polyp hydranths were chosen as the extraction material due to the following:

1. The highest concentration of RNA was isolated from a single specimen from previous transcriptomic experiments (Table IV.1) and large required amount of genomic DNA necessary for PacBio library preparation and sequencing.
2. Most accurate isolation of tissue from the same individual (physically attached clonal individuals) to reduces haplotypes.
3. BLASTx analyses of the preliminary transcriptome assessment of the polyp hydranth (GenBank Accession #SAMN13924705) reported no major contaminants among the species with the most top hits (Figure IV.1). Regardless, concerns of contamination

are addressed downstream using a homology-based approach to find and filter out genomic contigs/scaffolds from non-target organisms.

Table IV.1: The top five highest concentrations of RNA isolated from medusa and polyp individuals.

Lifecycle Stage	Sample ID	Concentration (ng/ul)	Avg. (ng/ul)
Medusa	M1	4.19	3.47386
	M2	4.01	
	M3	3.34	
	M4	2.95	
	M5	2.87	
Polyp	P1	6.12	5.2495
	P2	5.38	
	P3	5.19	
	P4	5.18	
	P5	4.37	

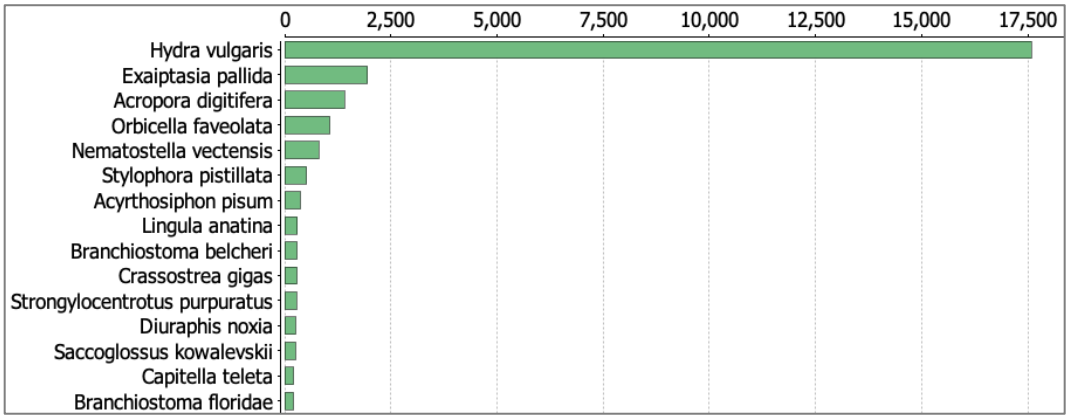


Figure IV.1: Top 15 species with the highest number of top hits in the polyp (hydranth) transcriptome.

Assembly trials with CCS and CLR datasets

The genome size of 27 cnidarian species from Japan were estimated based on flow cytometry (Adachi et al. 2017). A *Turritopsis* species from North Japan, the only species from family Oceaniidae, was reported to have an estimated genome size of 383

(+/- 4.90) Mbp. In comparison, other reported hydrozoan species had a genome size range of 323-681 Mbp (Adachi et al. 2017). However, the paper excludes *Hydra* genome size estimation, which ranges from ~380GB–1.5GB among different closely related species (Zacharias et al. 2004). The reported large range in genome size particularly among different hydrozoan groups is inferred to be a consequence of substantial evolutionary gene losses and gains (Adachi et al. 2017).

Three cells (Cells #1-3, Library 1) were initially sequenced on the PacBio Sequel II platform, but the genome coverage for the HiFi reads was much lower than anticipated (~20x coverage combined) due to an extremely low CCS conversion rate (~7-9%) among sequenced reads (Table IV.2). The majority of the generated zero-mode waveguides (ZMWs) were filtered out (~93-98%) during CCS due to now reaching the minimum number of three passes among the three cells (Table IV.2). Therefore, to determine whether there was an error at the extraction or library preparation step, a new HMW extraction and library was generated and sequenced (Cell #4-Library 2). Although the CCS conversion rate increased (~25%), a large portion of the ZMWs were still filtered out due not reaching three passes (Table IV.2).

Table IV.2: CCS processing and quality assessment of PacBio Sequel II subreads generated from two libraries.

	SMRT Cell			
	Cell 1 (Library 1)	Cell 2 (Library 1)	Cell 3 (Library 1)	Cell 4 (Library 2)
ZMWs input	3126260	3643159	991699	3250422
ZMWs generating CCS	**272319 (8.71%)**	**323295 (8.87%)**	**69282 (6.99%)**	**803129 (24.71%)**
ZMWs filtered	2853941 (91.29%)	3319864 (91.13%)	922417 (93.01%)	2447293 (75.29%)
Median length filter	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Below SNR threshold	86035 (3.01%)	134188 (4.04%)	8112 (0.88%)	9580 (0.39%)
Lacking full passes	2708502 (94.90%)	3097972 (93.32%)	902900 (97.88%)	2331843 (95.28%)
Heteroduplex insertions	6527 (0.23%)	9693 (0.29%)	1360 (0.15%)	14225 (0.58%)
Coverage drops	456 (0.02%)	670 (0.02%)	115 (0.01%)	1586 (0.06%)
Insufficient draft coverage	9964 (0.35%)	14569 (0.44%)	2997 (0.32%)	9949 (0.41%)
Draft too different	446 (0.02%)	751 (0.02%)	92 (0.01%)	1175 (0.05%)
Draft generation error	5092 (0.18%)	8483 (0.26%)	962 (0.10%)	5876 (0.24%)
Draft above maximum length	40 (0.00%)	63 (0.00%)	10 (0.00%)	145 (0.01%)
Draft below minimum length	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Reads failed polishing	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
Empty coverage windos	1126 (0.04%)	1532 (0.05%)	285 (0.03%)	2154 (0.09%)
CCS did not converge	8 (0.00%)	13 (0.00%)	3 (0.00%)	84 (0.00%)
CCS below minimum RQ	35785 (1.25%)	51993 (1.57%)	5591 (0.61%)	70821 (2.89%)
Unknown error	0 (0.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)

Preliminary HiFi assembly

Although the conversion rates and genome coverage of the four cells were not ideal, assembly trials of the ZMWs that passed CCS (i.e., HiFi reads) were performed using four different software: HiCanu (Nurk et al. 2020), Flye (Kolmogorov et al. 2019), SPAdes (Bankevich et al. 2012)and IPA assembler (Table IV.3). Using N50 length statistics and BUSCO genome completeness analyses, the following statistics were used to find the best genome for subsequent scaffolding and polishing: a) most contiguous genome (i.e., fewest number of contigs per overall assembly size); b) within target range of the genome (genome estimated ~383Mbp); c) highest number of complete and fragmented BUSCOs; d) close to estimated GC% of ~37-39% (Matsumoto et al., 2019; Matsumoto and Miglietta, 2021) (Table 3, labeled in green), where the best genome will

be used for further processing. Unfortunately, despite originating from an individual *T. dohrnii* colony, all of the CCS assemblies had very low completeness (~12%-18%) with an unusually high percentage of duplicated reads with similar percentages to single copy BUSCOs (Table IV.3). This may be due to a large number of biological contaminants within the draft genome and/or the result of poor CCS conversion rates and genome coverage of the HiFi reads (Table IV.2). Although the IPA assembler had the best assembly in terms of length statistic, the assembly had the furthest GC% from estimation, as well as the lowest BUSCO completeness. On the other hand, the SPAdes assembler had the highest BUSCO completeness, but had the largest and least contiguous assembly, while the Flye assembly had the closest GC% as estimated (Table IV.3).

Table IV.3: Assembly of the genomic HiFi reads using different algorithms (HiCanu, Flye, SPAdes, IPA), along with the BUSCO genome completeness assessment. (Green: statistics for highest quality genome; C+F- Complete and Fragmented, C- Complete, S- Single, D- Duplicated, F- Fragmented, M- Missing)

Genomic HiFi long-reads assembly				
	HiCanu	**Flye**	**SPAdes**	**IPA Assembler**
#Contigs	9,380	3,090	208,726	**1,723**
Total Size (bp)	328,234,377	231,540,628	714,570,551	**348,119,487**
N50	64,429	170,348	42,010	**347,125**
Mean	34,993	74,811	3,423	**202,043**
Median	16,211	35,896	258	**98,823**
GC%	39.88	**39.27**	40.90	42.21
Largest	5,108,046	5,116,225	1,415,270	**5,127,597**
Shortest	1,060	204	128	**15,656**
BUSCO	**C+F: 14.8% (141)**	**C+F: 12.8% (122)**	**C+F: 17.7% (168)**	**C+F: 12.4% (118)**
	C: 12.1% (115)	C: 11.1% (106)	C: 13.6% (129)	C: 11.0% (105)
	S: 6.1% (58)	S: 5.8% (55)	S: 7.7% (73)	S: 5.0 (48)
	D: 6.0% (57)	D: 5.3% (51)	D: 5.9% (56)	D: 6.0% (57)
	F: 2.7% (26)	F: 1.7% (16)	F: 4.1% (39)	F: 1.4% (13)
	M: 85.2% (813)	M: 87.2% (832)	M: 82.3% (786)	M: 87.6% (836)

To investigate the issues within the HiFi reads and assembly, the IPA assembly was polished and scaffolded with reads that failed CCS, and further examined using a taxonomic classifier software Kraken 2.0.9 (Wood et al. 2019) using the RefSeq 2020.07 database which includes micro-organisms, vertebrate model systems (human, mouse, rat, boar, cow and boar), the invertebrate model system *Drosophila*, and the plant model system *Arabidopsis* (for all specific taxa, see methods). Using a base-Kraken confidence filter of 0 (precision: 95.43, sensitivity 77.32), the analysis revealed that the majority of the HiFi scaffolds belonged to bacterial taxa (98.3%), and the unassigned (i.e., likely belongs to *T. dohrnii*) and metazoan species only accounted for 1.65% of the scaffolds (27 scaffolds). A BLASTn analysis of the scaffolded genome revealed 27 scaffolds with no hits against the NCBI Non-Redundant Nucleotide (NT) database, showing that the confidence filter of 0 is an acceptable parameter to filter Metazoa and unassigned reads from the genome downstream.

Preliminary CLR assembly

Due to similar problems faced in rare occasions among researchers sequencing Hydrozoa and Ctenophora species on PacBio platforms, particularly during library creation and CCS (Personal communication with Dr. Jessie Lie and Dr. Lutz Froenicke of UC Davis Sequencing and Bioinformaticss Core; Dr. Sara Goodwin and Dr. Yin Jing of CSHL Sequencing Core), the subreads (i.e., raw CLR reads) were used to for assembly trials. Due to the large size of the subread datasets (~128GB Cell #1, ~152GB

Cell #2, ~38GB Cell #3, ~305GB Cell #4), the initial assembly trial was performed only on subreads from Cell #1.

Unlike the HiFi reads that are corrected during CCS, CLR subreads generated from PacBio have much lower quality reads that need to be error corrected prior to assembly. 40x coverage of the longest subreads were corrected and trimmed using the Canu (Koren et al. 2017) correction tool, generating ~2.3 million output reads (Table IV.4B). This included shorter rescued reads that were not well represented in the 40x coverage correction, and in total covered ~70x of the estimated genome size (Table IV.4B). The assembled contigs were filtered for biological contaminants (i.e., only kept Eukaryota/unassigned contigs from Kraken2 output), and the resulting draft genome consisted of ~283Mbp among 21,774 contigs, with a GC content of 39% (Table IV.4C). Despite using only one CLR dataset (Cell #1 from Library 1) and assuming all classified contigs (e.g., not classified as Eukaryota) needed to be filtered out, the BUSCO analysis shows that the genome has substantially higher completeness (~54.7%) and lower duplication rates (~4.4%) than the HiFi assembles (Table IV.4C). Therefore, the HiFi datasets were excluded from the assembly dataset.

Table IV.4: Correction, assembly and genome quality assessment of CLR datasets from Cell #1 (Library 1) only.

A. CLR Cell #1		C. Assembly	
Total subreads	6,476,483	#Contigs	21,774
Raw reads w/ overlaps	5,468,845	Total Size (bp)	283,082,901
Coverage	157.039 (94.90%)	N50	13,403
Raw reads w/o overkaps	1,007,638	Mean	13,001
Coverage	8.443 (5.10%)	Median	11,381
B. 40x Correction of Longest Reads		GC%	0.39
Candidate for correction	3,180,515	Largest	324,627
Coverage	104.87	Shortest	1,022
Corrected	660,743	BUSCO	**C+F: 54.7% (522)**
Rescued	1,816,020		C: 43.8% (418)
Total corrected output	2,335,790		S: 39.9% (376)
Post-trimming	2,329,427		D: 4.4% (42)
Basepairs	26,744,765,502		F: 10.9% (104)
Coverage	69.46		M: 45.3% (432)

Genome assembly of *T. dohrnii*

Contamination filtering prior to assembly

CLR datasets from all generated cells (both libraries) were error-corrected using Canu (Koren et al. 2017), and the generated corrected reads were taxonomically classified using Kraken2 (Wood et al. 2019). As it is crucial not to discard reads that belong in *T. dohrnii*'s genome, a subset of the classified corrected dataset was preliminarily analyzed using BLASTn. Two different datasets were created, one with a stringent contamination filtering criterion (i.e., ST Dataset) and another with a more lenient filtering criterion (i.e., AD Dataset), elaborated in detail below.

Among classified 'Eukaryota' reads, the *Arabidopsis* (Thale Cress plant) category is the only non-metazoan species that could potentially include epibiotic algal species that often reside inconspicuously on *T. dohrnii* colonies. BLASTn results show

that among 6,111 reads classified as *Arabidopsis*, 898 (~15%) had hits. Among the reads with hits, only 10 sequences had >90% identity against all nucleotide sequences among NCBI, none of which were from algal species, while four aligned to metazoan species (fish, starfish, oyster and *Drosophila*). Therefore, the analyses show that our data has little contamination from algal species, reaffirming the preliminary analysis of the polyp transcriptome prior to DNA extraction (Figure IV.1). The reads classified as *Arabidopsis* are presumed to belong to parts of *T. dohrnii*'s genome that have comparable *k-mer* distributions found in plants and other eukaryotes. Therefore, error-corrected reads that classified as 'Eukaryota' and 'Unassigned' were extracted and used for both the ST and AD assembly dataset.

All other corrected reads that were categorized as biological contaminants (i.e., classified as bacteria, fungi, protozoa, virus) were processed through BLASTn to find sequences that do not align to microbial genomes. Approximately 19.5% (880,830 out of 4,517,631) of the reads from predicted contaminants had no BLAST hits and thus, were placed into both of the assembly datasets as they were presumed to originate from *T. dohrnii* (Appendix IV.A). Although a large portion of the hits did belong to microbial genomes, there were a number of sequences that showed top hits to a variety of metazoan taxa, such as insects (e.g., *Drosophila*, *Aethina*), other Cnidaria (*Actinia, Acropora*), and fish (*Acipenser*, *Acanthochromis*), to name a few. Additionally, sequences with hits against both microbial and metazoan genomes may be the outcome of horizontal gene transfer and cannot be ruled out. Therefore, all sequences that had hits were re-BLASTed against the NCBI's Nt Metazoa database (taxids: 33208).

Approximately 25% (895,973 out of 3,636,801) of the dataset had hits against metazoan genomes were presumed to belong to *T. dohrnii*'s genome (Appendix IV.B), and were incorporated into both the ST and AD datasets. The ST dataset with higher stringency criteria ultimately was composed of 3,228,022 corrected filtered reads. Lastly, the sequences with no hits against metazoan genomes (Appendix IV.B) were re-evaluated using the original BLASTn results to add to the AD dataset with less stringent criteria (i.e., all taxa). All sequences with hits against microbial genomes with >90% identity and e-value of 0.0 were removed (Appendix IV.C), and the remaining sequences (1,622,274 out of 2,740,828) were kept as part of the AD dataset. Ultimately, a total of 4,851,203 corrected reads were selected for the AD dataset. Remaining contaminating microbial sequences (i.e., contigs) in the dataset will be filtered out after the assembly prior to subsequent steps such as scaffolding and gap-filling.

Assembly trials using all cells

The AD dataset was constructed into a genome using the Canu assembler (Koren et al. 2017) and Hifiasm assembler (Cheng et al. 2021). The two genomes resulted in substantially different sizes, number of contigs, and BUSCO completeness (Table IV.5). The Hifiasm assembler produced a smaller (441Mbp) and more contiguous genome (N50: ~28Mbp, mean: ~21Mbp, median: ~12.5Mbp) than the Canu assembler (size: 1,737Mbp, N50: ~16.6Mbp, mean: 15.7Mbp, median: 11.3 Mbp) (Table IV.5). However, the Hifiasm assembly had a considerably lower percentage of BUSCO completeness at 36.9% in comparison to the Canu assembler, which resulted in 78.5%

completeness (Table IV.5). Therefore, only the Canu assembler was utilized for subsequent assembly trials. In contrast to the genome produced from the AD dataset (Canu assembler), the ST dataset generated a genome that was smaller in size (~1075Mbp) and number of contigs (~72k), but had similar BUSCO completeness at 77.7% despite the large size difference (Table IV.5). Therefore, genome generated from the ST dataset (ST Assembly) was chosen for further contamination filtering using Kraken and BLASTn and scaffolding, while the other two genomes (produced from AD dataset) was discarded.

Assembly trials using individual cells

In addition to the assembly of the two contamination filtered datasets (ST and AD dataset), each of the CLR datasets for cell 1, 2 and 4 were corrected and assembled individually (Table IV.5), as more sequencing reads and coverage does not necessarily result in a better assembly and may include more background noise sequences. Cell 3 was not processed due to the low number of generated sequencing reads (Table IV.2). Contamination filtering was not performed on the individual cell datasets, and will be performed on the contigs after assembly instead to conserve the total amount of needed computational resources. Among the assemblies built from individual cells, Cell #1 resulted in largest (1,331 Mbp, ~92k contigs) but highest BUSCO completeness (69.4%), while Cell #2 resulted in the most contiguous assembly with lower BUSCO completeness (58.2%) (Table IV.5). Ultimately, the assembly from Cell #1 (C1 Assembly) was chosen for subsequent contamination filtering and scaffolding due to the

highest genome completeness percentage, while Cell #2 and Cell #4 assemblies were discarded.

Table IV.5: Genomic CLR dataset assembly trials. * Sequencing reads for each individual cells were not filtered for contamination prior to assembly.

Genomic CLR assembly						
	AD Dataset (Canu Assembler)	**AD Dataset (Hifiasm assembler)**	**ST Dataset**	**Cell 1 Only***	**Cell 2 Only***	**Cell 4 Only***
#Contigs	111,010	21,038	72,023	92,053	54,809	73,456
Total Size (bp)	1,737 Mbp	441 Mbp	1,074 Mbp	1,331 Mbp	901 Mbp	1,068 Mbp
N50	16,677 bp	27,930	16,016	14,513	16,517	15,517
Mean	15,652	20,964	14,914	14,461	16,442	14,544
Median	11,299	12,487	11,488	11,797	12,213	10,248
GC%	45.46	43.67	43.68	43.51	41.33	47.20
Largest	1,849,331 bp	1,852,332	528,650	1,670,147	1,889,271	1,863,357
Shortest	1,743 bp	3,887	2,034	1,022	1,033	1,024
BUSCO	**C+F: 78.5% (748)**	**C+F: 36.9% (352)**	**C+F: 77.7% (741)**	**C+F: 69.4% (662)**	**C+F: 58.2% (556)**	**C+F: 51.7% (493)**
	C: 70.3% (670)	C: 31.1% (297)	C: 69.2% (660)	C: 58.4% (557)	C: 49.2% (470)	C: 42.3% (403)
	S: 51.5% (491)	S: 22.4% (214)	52.5% (501)	S: 44.2% (422)	S: 29.6% (378)	S: 27.8% (265)
	D: 18.8% (179)	D: 8.7% (83)	D: 159 (16.7%)	D: 14.2% (135)	D: 9.6% (92)	D: 14.5% (138)
	F: 8.2% (78)	F: 5.8% (55)	F: 8.5% (81)	F: 11% (105)	F: 9.0% (86)	F: 9.4% (90)
	M: 21.5% (206)	M: 63.1% (602)	M: 22.3% (213)	M: 30.6% (292)	M: 41.8% (398)	M: 48.3% (461)

Contamination filtering post-assembly

The two chosen genomes (ST and C1 Assembly) were filtered through Kraken2 and BLASTn to reduce microbial contamination among assembled contigs. Kraken filtering was able to capture a large portion of the genome in terms of BUSCO completeness from prior to filtering for both the ST assembly (622 out of 748 complete or fragmented BUSCOs) and C1 Assembly (522 out of 662 or complete or fragmented BUSCOs), despite being significantly smaller in total genome size (ST: 385 Mbp; C1: 285 Mbp) (Table IV.6AB). In addition, the rate of duplication decreased substantially after Kraken filtering for both assemblies, from approximately 14-17% to about 4-8% (Table IV.6AB).

For the BLASTn contamination filtering process, only 1,852 contigs were identified as microbial contamination within the ST Assembly, presumed so as the corrected-read dataset was pre-filtered for contamination using Kraken and BLASTn prior to assembly (Table IV.6A; Appendix IV.A-B). On the other hand, for the C1 Assembly, which was not filtered for contamination prior to the assembly, identified 41,543 contigs as microbial contaminants (Table IV.6B). After BLASTn filtering, all but 12 BUSCOs were recovered for the ST Assembly (Table IV.6A) and 19 for the C1 Assembly (Table IV.6B). However, the level of duplication was recaptured as well for both assemblies, to approximately 13-16% (Table IV.6AB).

The ST Assembly after only Kraken filtering reported similar BUSCO completeness to the C1 Assembly (65.2% vs. 67.4% completeness, 21 less complete or fragmented BUSCOs) despite being about half the genome size and number of contigs (385 Mbp among ~35.6k contigs vs. 761 Mbp among ~50.5k contigs) (Table IV.6AB). This could indicate that the BLASTn filtering step may include microbial contamination that reside in other published animal genomes and resulting in hits against the Metazoa genome. In addition to the two assemblies (ST and C1 Assembly), the Kraken filtered ST assembly (STK assembly) was also further processed (i.e., scaffolded, gap-filling, etc.) as a conservative draft genome option (Table IV.6C).

Table IV.6: Contamination filtering using Kraken and BLASTn, followed with scaffolding with genomic reads using LRScaf for assemblies: A) ST Assembly, B) C1 Assembly, C) STK Assembly.

A) ST Assembly					
	Contamination filtering		Scaffolding with Genomic Sequences		
	Kraken Filtering	BLASTn Filtering	Scaffolding #1 (AD)	Scaffolding #2 (CCS-IPA)	Scaffolding #3 (CCS-SPAdes)
# Contigs/Scaffolds	26,653	70,171	66,433	64,409	63,800
Total Size (bp)	385 Mbp	1,044 Mbp	1,053Mbp	1174 Mbp	1189 Mbp
N50	15,490	15,954	17,041	19,667	20,388
Mean	14,389	14,881	15,846	18,223	18,643
GC%	39.46	43.66	43.20	38.75	38.25
% N	-	-	0.97	11.18	12.32
% of scaffolded contigs	-	-	11.40	27.50	30.40
Largest sequence	528,650	528,650	528,650	5,060,257	5,060,257
Shortest sequence	2,034	2,034	2,034	2,034	2,034
BUSCO	**C+F: 65.2% (622)**	**C+F: 77.2% (736)**			
	C: 56.7% (541)	C: 68.8% (656)			
	S: 48.7% (465)	S: 52.4% (500)			
	D: 8.0% (76)	D: 16.4% (156)			
	F: 8.5% (81)	F: 8.4% (80)			
	M: 34.8% (332)	M: 22.8% (218)			

B) Cell 1					
	Contamination filtering		Scaffolding with Genomic sequences		
	Kraken Filtering	BLASTn Filtering	Scaffolding #1 (AD)	Scaffolding #2 (CCS-IPA)	Scaffolding #3 (CCS-SPAdes)
# Contigs/Scaffolds	21,774	50,510	47, 061	45,809	45,295
Total Size (bp)	283 Mbp	761 Mbp	772 Mbp	880 Mbp	894 Mbp
N50	13,403	14,969	16,080	18,404	19,204
Mean	13,001	15,070	16,416	19,221	19,747
GC%	39.87	41.78	41.18	36.14	35.61
% N	-	-	1.43	13.52	14.73
% of scaffolded contigs	-	-	15.70	31.70	34.70
Largest sequence	324,627	1,670,147	1,670,147	5,090,353	5,090,353
Shortest sequence	1,022	1,022	1,022	1,022	1,022
BUSCO	**C+F: 54.7% (522)**	**C+F: 67.4% (643)**			
	C: 43.8% (418)	C: 56.6% (540)			
	S: 39.9% (376)	S: 43.4% (414)			
	D: 4.4% (42)	D: 13.2% (126)			
	F: 10.9% (104)	F: 10.8% (103)			
	M: 45.3% (432)	M: 32.6% (311)			

C) ST Kraken					
	Contamination filtering	Scaffolding with Genomic Sequences			
	Kraken Filtering	Scaffolding #1 (AD)	Scaffolding #2 (CCS-IPA)	Scaffolding #3 (CCS-SPAdes)	Scaffolding #4 (Cell 1, Cell 2, Cell 4)
# Contigs/Scaffolds	26,653	25,651	25,505	25,388	25,131
Total Size (bp)	385 Mbp	388 Mbp	424 Mbp	425 Mbp	427 Mbp
N50	15,490	16,285	17,697	17,849	18,253
Mean	14,389	15,134	16,623	16,740	16,986
GC%	39.46	39.24	35.92	35.83	35.57
% N	-	0.61	9.03	9.26	9.90
% of scaffolded contigs	-	7.10	16.40	17.40	19.20
Largest sequence	528,650	528,650	4,832,215	4,832,215	4,832,215
Shortest sequence	2,034	2,034	2,034	2,034	2,034
BUSCO	**C+F: 65.2% (622)**				
	C: 56.7% (541)				
	S: 48.7% (465)				
	D: 8.0% (76)				
	F: 8.5% (81)				
	M: 34.8% (332)				

Genome scaffolding with genomic and transcriptomic sequences

Scaffolding with genomic sequences

Multiple scaffolding iterations were performed on each of the three assemblies to bridge contigs with gap regions (i.e., Ns), corresponding to the alignment overlaps of reads and assembled sequences. Scaffolding was first performed with genomic reads using the LRScaf software (Qin et al., 2019). Scaffolding performed with the contaminant-filtered AD dataset (refer to Appendix IV. A-C) resulted in 11.4%, 15.7% and 7.1% of contigs in scaffolded contigs for the ST, C1 and STK Assembly, respectively (Table IV.6). Additionally, the two assemblies produced from the IPA Assembler and SPAdes Assembler using CCS reads were used to scaffold the three genomes, resulting in a marked increase of the percentage of contigs in scaffolds, 27.5%, 31.7% and 17.4% for the ST, C1 and STK Assembly, respectively (Table IV.6). The IPA and SPAdes assemblies were utilized for scaffolding due to having the highest contiguity (IPA assembly) and BUSCO completeness (SPAdes assembly) among the CCS assembly trials (Table IV.3). For the STK Assembly, an additional scaffolding iteration was performed using Cell 1, Cell 2 and Cell 4 assemblies using CLR datasets (Table IV.6). There was just a 1.8% increase (resulting in 19.2% total) in the percentage of contigs scaffolded in the STK Assembly as most sequence bridges were incorporated during the first iteration of scaffolding with reads from the AD dataset.

Scaffolding with transcripts and RNA-seq libraries

Algorithms and software have been developed to use assembled transcriptomes and RNA-sequencing (RNA-seq) reads to further scaffold, polish and improve genome

assemblies in metazoans (Xue et al., 2013; Zhu et al., 2018). Prior to scaffolding iterations with previously sequenced RNA-seq libraries and assembled transcripts (Matsumoto and Miglietta, 2021), the coverage of the assembled transcripts of *T. dohrnii'*s genome (original, pre-filtered transcriptome from Matsumoto and Miglietta, 2021 of colonial polyp, medusa, cyst and reversed polyp stages) was assessed using the BUSCO (Simão et al. 2015) tool in genome mode. This provided a broad overview on how much of the genome our transcripts and RNA-seq reads covered. Among the 416,169 transcripts with a total size of 323.124Mb with the longest contig as 27,093bp, there was 81.6% BUSCO completeness (complete and fragmented), indicating that the transcriptomic sequences covered a substantial portion of the genome and could provide more connecting bridges among the contigs in our draft genome assemblies. Furthermore, the raw RNA-seq libraries that were used to generate the transcriptome may be able to decipher additional contaminating sequences through alignment analyses against the draft genomes downstream.

Scaffolding iterations with RNA-seq libraries and assembled transcripts using P_RNA_scaffolder (Zhu et al., 2018) and L_RNA_scaffolder (Xue et al., 2013), respectively, resulted in an increase in the number of scaffolded contigs in all three assemblies, specifically 4.1% for the ST Assembly, 3.9% for the C1 Assembly, and 10.6% in the STK Assembly (Table IV.7). Ultimately, the assemblies resulted in 62,042 (from 70,171) contigs, 43,508 (from 50,510) contigs, and 23,544 (from 26,653) contigs for the ST, C1, and STK Assembly, respectively (Table IV.7). There was a minuscule dip in the BUSCO completeness for each of the assemblies (0.2 - 0.7%) after genomic

and transcriptomic scaffolding iterations, resulting from N's that are inserted in the bridged contigs (i.e., 100N's inserted to connect scaffolding contigs; see methods) and not being considered as a fragmented BUSCO as prior to the scaffolding (Table IV.7).

Table IV.7: Scaffolding with RNA-seq libraries and assembled transcripts using P/L_RNA_scaffolder, followed by gap-filling using genomic sequences of the three draft genome assemblies using TGSGapCloser: A) ST Assembly, B) C1 Assembly, C) STK Assembly.

A) ST Assembly

	Scaffolding with RNA sequences		Gap-filling			
	RNA-seq	**Transcripts**	**All Corrected Reads**	**CCS Raw Reads**	**CCS Assembly (IPA)**	**CCS Assembly (SPAdes)**
# Scaffolds	63,180	62,042	62,042	62,042	62,042	62,042
Total Size (bp)	1189.506 Mbp	1189.621 Mbp	1188.638 Mbp	1188.647 Mbp	1191.141 Mbp	1191.182 Mbp
N50	20,924	22,008	21,983	21,980	22,058	22,061
Mean	18,827	19,174	19,159	19,159	19,199	19,200
GC%	38.24	38.24	38.72	38.76	43.04	43.39
% N	12.33	12.34	11.18	11.12	0.98	0.11
% of scaffolded contigs	32.00	34.50	26.00	25.50	7.9	5.2
Largest sequence (bp)	5,060,257	5,060,257	5,057,075	5,057,075	5,060,902	5,060,902
Shortest sequence (bp)	2,034	2,034	2,034	2,034	2,034	2,034
	Before Scaffolding	**After Scaffolding**				**After Gapfilling**
BUSCO	**C+F: 77.2% (736)**	**C+F: 77.0% (734)**				**C+F: 77.0% (734)**
	C: 68.8% (656)	C: 68.7% (655)				C: 68.7% (655)
	S: 52.4% (500)	S: 52.0% (496)				S: 52.0% (496)
	D: 16.4% (156)	D: 16.7% (159)				D: 16.7% (159)
	F: 8.4% (80)	F: 8.3% (79)				F: 8.3% (79)
	M: 22.8% (218)	M: 23.0% (220)				M: 23.0% (220)

B) C1 Assembly

	Scaffolding with RNA sequences		Gap-filling			
	RNA-seq	**Transcripts**	**All Corrected Reads**	**CCS Raw Reads**	**CCS Assembly (IPA)**	**CCS Assembly (SPAdes)**
# Scaffolds	44,674	43,508	43,508	43,508	43,508	43,508
Total Size (bp)	894.520 Mbp	894.636 Mbp	893.482 Mbp	893.408 Mbp	895.721 Mbp	896.330 Mbp
N50	19,779	21,055	21,000	21,000	21,143	21,169
Mean	20,023	20,563	20,536	20,534	20,588	20,602
GC%	35.61	35.61	36.19	36.25	41.39	41.75
% N	14.73	14.74	13.27	13.19	0.98	0.1
% of scaffolded contigs	36.20	38.60	28.00	27.50	7.1	4.9
Largest sequence	5,090,353	5,090,353	5,090,363	5,090,363	5,093,939	5,093,939
Shortest sequence	1,022	1,022	1,022	1,022	1,022	1,022
	Before Scaffolding	**After Scaffolding**				**After Gapfilling**
BUSCO	**C+F: 67.4% (643)**	**C+F: 67.0% (639)**				**C+F: 67.9 (648)**
	C: 56.6% (540)	C: 56.2% (536)				C: 57.5% (549)
	S: 43.4% (414)	S: 43.2% (412)				S: 44.1% (421)
	D: 13.2% (126)	D: 13.0 (124)				D: 13.4% (128)
	F: 10.8% (103)	F: 10.8% (103)				F: 10.4% (99)
	M: 32.6% (311)	M: 33.0% (315)				M: 32.1% (306)

Table IV.7 Continued: Scaffolding with RNA-seq libraries and assembled transcripts using P/L_RNA_scaffolder, followed by gap-filling using genomic sequences of the three draft genome assemblies using TGSGapCloser: A) ST Assembly, B) C1 Assembly, C) STK Assembly.

C) STK Assembly						
	Scaffolding with RNA sequences		Gap-filling			
	RNA-seq	Transcripts	All Corrected Reads	CCS Raw Reads	CCS Assembly (IPA)	CCS Assembly (SPAdes)
# Scaffolds	24,429	23,544	23,544	23,544	23,544	23,544
Total Size (bp)	426.938 Mbp	427.037 Mbp	426.766 Mbp	426.860 Mbp	426.879 Mbp	426.947 Mbp
N50	19,010	20,477	20,474	20,477	20,477	20,480
Mean	17,477	18,137	19,159	19,159	19,199	19,200
GC%	35.57	35.56	35.85	35.87	39.05	39.11
% N	9.92	9.93	9.17	9.08	1.00	0.83
% of scaffolded contigs	24.40	29.80	23.20	22.30	13.3	12.7
Largest sequence (bp)	4,832,215	4,832,215	4,832,215	4,832,215	4,834,034	4,834,034
Shortest sequence (bp)	2,034	2,034	2,034	2,034	2,034	2,034
BUSCO	**Before Scaffolding**	**After Scaffolding**				**After Gapfilling**
	C+F: 65.2% (622)	**C+F: 64.9% (618)**				**C+F: 66.1% (630)**
	C: 56.7% (541)	C: 56.7% (540)				C: 58.3% (556)
	S: 48.7% (465)	S: 49.0% (467)				S: 47.5% (453)
	D: 8.0% (76)	D: 7.7% (73)				D: 10.8% (103)
	F: 8.5% (81)	F: 8.2% (78)				F: 7.8% (74)
	M: 34.8% (332)	M: 35.1% (336)				M: 33.9% (324)

Genome gap-filling

Gap-filling was performed using the TGS-GapCloser software (Xu et al. 2020), designed specifically for third-generation sequencing reads and assembled contigs to fill gap regions (i.e., N's) among scaffolds. Read datasets (Corrected CLR and CCS) and the two CCS assemblies used for scaffolding (IPA and SPAdes assembly) were used for the gap-filling on all three draft assemblies. The corrected CLR read dataset was able to fill 8.5% of the 34.5% of scaffolded contigs in the ST Assembly, 10.6% of the 38.6% in the C1 Assembly, and 6.6% of the 29.8% in the STK Assembly, while the raw CCS reads were unable to fill much of the remaining scaffolds, only about 0.5% - 0.9% among the assemblies (Table IV.7). Lastly, the two CCS assemblies, IPA and SPAdes assembly were able to fill the gaps of 20.3%, 22.6% and 9.6% of the scaffolded contigs for the three assemblies, respectively (Table IV.7). After the gap-filling iterations, the C1 and STK assembly had 5 and 8 additional recovered BUSCOs, respectively, in comparison

to the pre-scaffolded assemblies (Table IV.7). On the other hand, the ST assembly resulted in 2 less BUSCOs that were not able to be recovered in the gap-filling step (Table IV.7).

Synopsis of the draft genome assemblies

The final (i.e., after gap-filling step) ST Assembly resulted in a total genome size of approximately 1,191 Mbp consisting of 62,042 scaffolds with a N50 of 22,061 bp and a GC% of 43.39%, with the largest scaffold as 5,060,902 bp (Table IV.7). The assembly had 77% BUSCO completeness with a duplication rate of 16.7% (Table IV.7). The C1 Assembly resulted in a total genome size of approximately 896 Mbp consisting of 43,508 scaffolds with a N50 of 21,169 bp and a GC% of 41.75%, with the largest scaffold as 5,093,939 bp (Table IV.7). The assembly had 67.9% BUSCO completeness with a duplication rate of 13.4% (Table IV.7). Lastly, the STK Assembly resulted in a total genome size of approximately 427Mbp consisting of 23,544 scaffolds with a N50 of 20,480 bp and a GC% of 39.11%, with the largest scaffold as 4,834,034 bp (Table IV.7). The assembly had 66.1% BUSCO completeness with a duplication rate of 10.8% (Table IV.7).

The ST Assembly had the highest BUSCO completeness (77%), but was much larger in genome size (1191Mbp) than estimated with a related *Turritopsis* species and most other Hydrozoa species (Adachi et al. 2017) and had the highest duplication rates among the three draft genomes. However, there are *Hydra* species that report genomes that are up to 1500 Mbp (Zacharias et al. 2004) and thus, *T. dohrnii* having a genome

size much larger than the related *Turritopsis* species cannot be ruled out. On the other hand, the STK Assembly was closer to the estimated genome size (427Mbp) to the related *Turritopsis* species and other hydrozoan species excluding *Hydra* (Adachi et al. 2017) and reported the lower duplication rates (~6% less), despite possessing BUSCO completeness about 11% lower than the ST Assembly (66.1%). The C11 Assembly was discarded, as the genome size was almost double the size of the STK Assembly despite having similar BUSCO completeness (1.8% difference in completeness). Ultimately, the ST Assembly and STK Assembly was kept as the two final draft genomes, where the ST Assembly representing the less conservative draft genome with more lenient filtering parameters and higher genome completeness, and ST Kraken assembly representing the conservative draft genome with strict filtering parameters and lower genome completeness and duplication levels.

Draft genome comparison with other Cnidaria

In comparison to the published genomes of other Cnidaria, the draft assemblies produced for *T. dohrnii* (ST and STK Assembly) have much larger number of assembled scaffolds (~62k and 23.5k), being much higher than the published counterparts (~10.5k - ~10.8k scaffolds) (Table IV.8). In addition, the scaffold N50 for *T. dohrnii* is smaller than most of the reported Cnidaria other than *Stylophora pistillata*, indicating that the draft assemblies are highly fragmented, particularly the ST Assembly (lenient) (Table IV.8). The GC% for the conservative draft assembly (STK Assembly at 39.1%) was similar to a number of the reported cnidarians, while the more lenient assembly (ST Assembly at 43.4%) showed elevated percentages in comparison (Table IV.8). This may

be in indication that the ST Assembly still contains moderate levels of microbial contamination that needs to be identified and removed, as the GC% of pre-filtered assemblies and Kraken-filtered assemblies did show a decrease in GC% (Table IV.5 and IV.6). Lastly, the BUSCO completeness was in the range of published Cnidaria (74.4% - 91.4%) for the ST Assembly (lenient) at 77%, but was lower than the reported range for the STK Assembly (conservative) at 66.1% (Table IV.8).

Table IV.8: Genome assembly comparison among different cnidarian species in Hydrozoa, Anthozoa, Scyphozoa and Cubozoa.

	Hydrozoa				Anthozoa				Scyphozoa		Cubozoa
Species	*Turritopsis dohrnii*		*Hydra vulgaris* (V.2)	*Clytia hemisphaerica*	*Nematostella vectensis*	*Stylophora pistillata*	*Exaiptasia pallida*	*Acropora digitifera*	*Aurelia aurita*	*Chrysora quinquecirrha*	*Morbakke virulenta*
	ST Assembly (lenient)	ST Kraken Assembly (conservative)									
Genome Size (Mbp)	1191	427	854	445	457	400	256	420	377	331	952
# of Scaffolds	62,042	23,544	5,525	7,644	10,804	5,688	4,312	10,804	2,710	2,496	4538
Scaffold N50 (Mbp)	0.2	0.2	1.0	0.4	0.5	0.2	0.4	0.5	1.0	0.7	2.2
GC %	43.4	39.1	25.4	35.0	39.0	38.5	29.8	39.0	34.7	37.5	31.4
Gap base (N) %	0.11	0.83	8.0	16.6	16.6	10.5	17.7	16.6	6.6	4.9	11.9
BUSCO completeness	77	66.1	80.2	86.4	91.4	88	87.1	74.4	79.8	78.8	81.5

IV.4. Conclusion

Further sequencing of high quality long reads that come from less contamination-prone life cycle stage, such as the planktonic medusa stage, or approaches such as single-cell sequencing, will be necessary to reduce the fragmentation and increase the completeness of draft assemblies through scaffolding and gap-filling iterations. Furthermore, the less fragmented genomes with more contiguous and longer scaffolds can be further analyzed using RNA-seq alignment analyses to identify and remove microbial contamination that may still reside in the draft assemblies. Ultimately, a more complete genome can be annotated for gene models and repeats, and whole-genome comparison with other cnidarian taxa may reveal *T. dohrnii* specific genes that may

provide more insight into the genetic networks and constituents that underlie the species' unique ability to undergo reverse development through cell transdifferentiation.

CHAPTER V

GENERAL CONCLUSION

The three combined projects presented in this dissertation developed genomic tools for *T. dohrnii* as a non-traditional research system to further understand the genetics of cellular plasticity and transdifferentiation *in vivo*. This dissertation consists of the following inter-connected research projects:

Project 1) Chapter II produced a high-quality annotated transcriptome for *T. dohrnii* of life cycle stages involved in the reverse development sequence (colonial polyp, medusa, cyst, reversed polyp), and provided insight into genetic constituents and networks associated with each of its life cycle stages and the ontogeny reversal of *T. dohrnii*. The sequential DGE analysis revealed that the cyst stage of *T. dohrnii* is enriched in genes associated with aging/lifespan, regulation of transposable elements, DNA repair and damage response, and cancer and tumor-related genes. The pair-wise DGE analyses revealed that the polyps developed from different trajectories, reverse development and asexual budding, show significant differences in transcriptional profiles, with processes of chromatin remodeling, matrix metalloproteinases, and embryonic development being highly active in the polyp that underwent ontogeny reversal. The medusa and colonial polyp also showed significant differences in gene activity, with transmembrane transport, nervous system, components of the mesoglea, and muscle contraction-related categories enriched in the medusa, while categories

related to chitin metabolism and the formation of primary germ layers are enriched in the colonial polyp.

Project 2) Chapter III identified and profiled the expression of putative gene homologs recognized as crucial regulators of tissue regeneration, cellular plasticity, and longevity in mammalian systems during the ontogeny reversal of *T. dohrnii*. The presented research showcases the relevancy of *T. dohrnii* as a non-traditional research system that may be impactful for furthering our understanding of human health, and the need of expansion on exploratory studies on species that go beyond classically used systems. Genes with high relevance in the medical sector, such as SIRTs, POU factors, HSPs and telomerase activating factors were found to be active during *T. dohrnii*'s reverse development. Furthermore, the presented work provided evidence of Yamanaka Factor homologs being potentially present in *T. dohrnii*, challenge our understanding of how the networks underlying pluripotency in Cnidaria relate to those of mammals.

Project 3) Chapter IV provides the first available draft genome assembly for *T. dohrnii*, albeit not the highest quality due the incorporation of microbial contaminants, providing the quintessential initial step to developing the species as a non-traditional research system, while demonstrating the challenges and complexity of genome construction for the species. The presented work compares different contamination filtering, assembly software, and scaffolding approaches to find the best avenue to genome construction for *T. dohrnii* and offers suggestions for sequencing approaches to further increase the contiguity, gene completeness, and overall quality of the draft genome in future projects.

Ultimately, the production of genomic tools and research studies presented on *T. dohrnii* and its ability to undergo ontogeny reversal have opened new avenues of exploration that can further contribute to our understanding of tissue regeneration, cell plasticity and aging in metazoans. Future research in the long-term project to advance *T. dohrnii* as a valuable research system include generating a more complete and high-quality draft of its genome, investigating the intermediate transitional stages (i.e., shorter increments of time between the medusa, cyst and reversed polyp) that are included in the medusa-to-polyp reversion process, and performing fluorescent *in situ* tissue (e.g., hybridization assays) analyses to visualize changes in the expression of genes that were identified in this dissertation to have the potential to underlie ontogeny reversal and cell transdifferentiation in *T. dohrnii*. In addition, the results of the presented work can be compared to other Cnidaria that do not have capabilities to undergo life cycle reversal to further decipher genetic networks that differentiate *T. dohrnii* to other taxa as an 'immortal' species, or to other highly regenerative taxa to identify networks that are crucial to regenerative strategies in metazoans.

REFERENCES

Abad, M., L. Mosteiro, C. Pantoja, M. Canamero, T. Rayon *et al.*, 2013 Reprogramming in vivo produces teratomas and iPS cells with totipotency features. *Nature* 502 (7471):340-345.

Adachi, K., H. Miyake, T. Kuramochi, K. Mizusawa, and S.-i. Okumura, 2017 Genome size distribution in phylum Cnidaria. *Fisheries science* 83 (1):107-112.

Adams, J.C., and A. Brancaccio, 2015 The evolution of the dystroglycan complex, a major mediator of muscle integrity. *Biology open* 4 (9):1163-1179.

Adler, A.S., S. Sinha, T.L. Kawahara, J.Y. Zhang, E. Segal *et al.*, 2007 Motif module map reveals enforcement of aging by continual NF-κB activity. *Genes & development* 21 (24):3244-3257.

Al-Shahrour, F., R. Díaz-Uriarte, and J. Dopazo, 2004 FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformaticss* 20 (4):578-580.

Alvarado, A.S., and S. Yamanaka, 2014 Rethinking differentiation: stem cells, regeneration, and plasticity. *Cell* 157 (1):110-119.

Amano, H., and E. Sahin, 2019 Telomeres and sirtuins: at the end we meet again. *Molecular & cellular oncology* 6 (5):e1632613.

Anandarajan, M., S. Paulraj, and R. Tubman, 2009 ABCA3 Deficiency: an unusual cause of respiratory distress in the newborn. *The Ulster medical journal* 78 (1):51.

Babicki, S., D. Arndt, A. Marcu, Y. Liang, J.R. Grant *et al.*, 2016 Heatmapper: web-enabled heat mapping for all. *Nucleic acids research* 44 (W1):W147-W153.

Bai, S., R. Thummel, A.R. Godwin, H. Nagase, Y. Itoh *et al.*, 2005 Matrix metalloproteinase expression and function during fin regeneration in zebrafish: analysis of MT1-MMP, MMP2 and TIMP2. *Matrix biology* 24 (4):247-260.

Barber, L.J., J.L. Youds, J.D. Ward, M.J. McIlwraith, N.J. O'Neil *et al.*, 2008 RTEL1 maintains genomic stability by suppressing homologous recombination. *Cell* 135 (2):261-271.

Bavestrello, G., C. Sommer, and M. Sarà, 1992 Bi-directional conversion in Turritopsis nutricula (Hydrozoa). *Scientia Marina* 56 (2-3):137-140.

Beausejour, C.M., and J. Campisi, 2006 Ageing: balancing regeneration and cancer. *Nature* 443 (7110):404.

Bellizzi, D., S. Dato, P. Cavalcante, G. Covello, F. Di Cianni *et al.*, 2007 Characterization of a bidirectional promoter shared between two human genes related to aging: SIRT3 and PSMD13. *Genomics* 89 (1):143-150.

Bellizzi, D., G. Rose, P. Cavalcante, G. Covello, S. Dato *et al.*, 2005 A novel VNTR enhancer within the SIRT3 gene, a human homologue of SIR2, is associated with survival at oldest ages. *Genomics* 85 (2):258-263.

Bessereau, J.-L., 2006 Transposons in C. elegans. *WormBook*:1.

Blasco, M.A., 2007 Telomere length, stem cells and aging. *Nature chemical biology* 3 (10):640.

Bodnar, A.G., M. Ouellette, M. Frolkis, S.E. Holt, C.-P. Chiu *et al.*, 1998 Extension of life-span by introduction of telomerase into normal human cells. *science* 279 (5349):349-352.

Borradaile, N.M., A. Watson, and J.G. Pickering, 2011 Regeneration and Aging: Regulation by Sirtuins and the NAD+ Salvage Pathway, pp. 289-298 in *Regenerative Nephrology*. Elsevier.

Bosch, T.C., A. Klimovich, T. Domazet-Lošo, S. Gründer, T.W. Holstein *et al.*, 2017 Back to the basics: cnidarians start to fire. *Trends in neurosciences* 40 (2):92-105.

Bowles, J., G. Schepers, and P. Koopman, 2000 Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. *Developmental biology* 227 (2):239-255.

Calderwood, S., 2007 Molecular chaperones and the ubiquitin proteasome system in aging. *The Ubiquitin Proteasome System in the Central Nervous System. Nova*:537-552.

Calvo, R., and H.A. Drabkin, 2000 Embryonic genes in cancer. *Annals of oncology* 11:207-218.

Chapman, J.A., E.F. Kirkness, O. Simakov, S.E. Hampson, T. Mitros *et al.*, 2010 The dynamic genome of Hydra. *Nature* 464 (7288):592.

Chen, S., and G. Parmigiani, 2007 Meta-analysis of BRCA1 and BRCA2 penetrance. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* 25 (11):1329.

Clough, R.L., R. Sud, N. Davis-Silberman, R. Hertzano, K.B. Avraham *et al.*, 2004 Brn-3c (POU4F3) regulates BDNF and NT-3 promoter activity. *Biochemical and biophysical research communications* 324 (1):372-381.

Collas, P., and A.-M. Håkelien, 2003 Reprogramming somatic cells for therapeutic applications. *e-biomed: the journal of regenerative medicine* 4 (2):7-13.

Conesa, A., S. Götz, J.M. García-Gómez, J. Terol, M. Talón *et al.*, 2005 Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformaticss* 21 (18):3674-3676.

Dang, C.V., 1999 c-Myc target genes involved in cell growth, apoptosis, and metabolism. *Molecular and cellular biology* 19 (1):1-11.

Davidson, N.M., A.D. Hawkins, and A. Oshlack, 2017 SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes. *Genome biology* 18 (1):148.

De Vito, D., S. Piraino, J. Schmich, J. Bouillon, and F. Boero, 2006 Evidence of reverse development in Leptomedusae (Cnidaria, Hydrozoa): the case of Laodicea undulata (Forbes and Goodsir 1851). *Marine Biology* 149 (2):339.

Derevyanko, A., K. Whittemore, R.P. Schneider, V. Jiménez, F. Bosch *et al.*, 2017 Gene therapy with the TRF 1 telomere gene rescues decreased TRF 1 levels with aging and prolongs mouse health span. *Aging Cell* 16 (6):1353-1368.

Dobrovolskaia, M.A., A.E. Medvedev, K.E. Thomas, N. Cuesta, V. Toshchakov *et al.*, 2003 Induction of in vitro reprogramming by Toll-like receptor (TLR) 2 and TLR4 agonists in murine macrophages: effects of TLR "homotolerance" versus "heterotolerance" on NF-κB signaling pathway components. *The Journal of Immunology* 170 (1):508-519.

Drolet, D.W., K.M. Scully, D.M. Simmons, M. Wegner, K. Chu *et al.*, 1991 TEF, a transcription factor expressed specifically in the anterior pituitary during embryogenesis, defines a new class of leucine zipper proteins. *Genes & development* 5 (10):1739-1753.

Dubrez, L., S. Causse, N.B. Bonan, B. Dumetier, and C. Garrido, 2019 Heat-shock proteins: chaperoning DNA repair. *Oncogene*:1-14.

DuBuc, T.Q., N. Traylor-Knowles, and M.Q. Martindale, 2014 Initiating a regenerative response; cellular and molecular features of wound healing in the cnidarian Nematostella vectensis. *BMC biology* 12 (1):24.

Edgar, R.C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 32 (5):1792-1797.

Flores, I., and M.A. Blasco, 2010 The role of telomeres and telomerase in stem cell aging. *FEBS letters* 584 (17):3826-3830.

Fuchs, B., W. Wang, S. Graspeuntner, Y. Li, S. Insua *et al.*, 2014 Regulation of polyp-to-jellyfish transition in Aurelia aurita. *Current Biology* 24 (3):263-273.

Gauchat, D., F. Mazet, C. Berney, M. Schummer, S. Kreger *et al.*, 2000 Evolution of Antp-class genes and differential expression of Hydra Hox/paraHox genes in anterior patterning. *Proceedings of the National Academy of Sciences* 97 (9):4493-4498.

Gold, D.A., R.D. Gates, and D.K. Jacobs, 2014 The early expansion and evolutionary dynamics of POU class genes. *Molecular biology and evolution* 31 (12):3136-3147.

Gold, D.A., T. Katsuki, Y. Li, X. Yan, M. Regulski *et al.*, 2019 The genome of the jellyfish Aurelia and the evolution of animal complexity. *Nature ecology & evolution* 3 (1):96.

Götz, S., J.M. García-Gómez, J. Terol, T.D. Williams, S.H. Nagaraj *et al.*, 2008 High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic acids research* 36 (10):3420-3435.

Grabowska, W., E. Sikora, and A. Bielak-Zmijewska, 2017 Sirtuins, a promising target in slowing down the ageing process. *Biogerontology* 18 (4):447-476.

Greiss, S., and A. Gartner, 2009 Sirtuin/Sir2 phylogeny, evolutionary considerations and structural conservation. *Molecules and cells* 28 (5):407.

Guindon, S., and O. Gascuel, 2003 A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology* 52 (5):696-704.

Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler, 2013 QUAST: quality assessment tool for genome assemblies. *Bioinformaticss* 29 (8):1072-1075.

Haas, B.J., A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood *et al.*, 2013 De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* 8 (8):1494.

Harrington, L., T. McPhail, V. Mar, W. Zhou, R. Oulton *et al.*, 1997 A mammalian telomerase-associated protein. *science* 275 (5302):973-977.

He, J., L. Zheng, W. Zhang, and Y. Lin, 2015 Life cycle reversal in Aurelia sp. 1 (Cnidaria, Scyphozoa). *PLoS One* 10 (12):e0145314.

Hertzano, R., M. Montcouquiol, S. Rashi-Elkeles, R. Elkon, R. Yücel *et al.*, 2004 Transcription profiling of inner ears from Pou4f3 ddl/ddl identifies Gfi1 as a target of the Pou4f3 deafness gene. *Human molecular genetics* 13 (18):2143-2153.

Higgins, D.G., A.J. Bleasby, and R. Fuchs, 1992 CLUSTAL V: improved software for multiple sequence alignment. *Bioinformaticss* 8 (2):189-191.

Hoesel, B., and J.A. Schmid, 2013 The complexity of NF-κB signaling in inflammation and cancer. *Molecular cancer* 12 (1):86.

Holtfreter, J., 1946 Structure, motility and locomotion in isolated embryonic amphibian cells. *Journal of morphology* 79 (1):27-62.

Hou, Y., H. Wei, Y. Luo, and G. Liu, 2010 Modulating expression of brain heat shock proteins by estrogen in ovariectomized mice model of aging. *Experimental gerontology* 45 (5):323-330.

Hsieh, P.N., G. Zhou, Y. Yuan, R. Zhang, D.A. Prosdocimo *et al.*, 2017 A conserved KLF-autophagy pathway modulates nematode lifespan and mammalian age-associated vascular dysfunction. *Nature communications* 8 (1):1-12.

Hsu, A.-L., C.T. Murphy, and C. Kenyon, 2003 Regulation of aging and age-related disease by DAF-16 and heat-shock factor. *science* 300 (5622):1142-1145.

Jia, G., L. Su, S. Singhal, and X. Liu, 2012 Emerging roles of SIRT6 on telomere maintenance, DNA repair, metabolism and mammalian aging. *Molecular and cellular biochemistry* 364 (1-2):345-350.

Jopling, C., S. Boue, and J.C.I. Belmonte, 2011 Dedifferentiation, transdifferentiation and reprogramming: three routes to regeneration. *Nature reviews Molecular cell biology* 12 (2):79.

Kaeberlein, M., and R.W. Powers III, 2007 Sir2 and calorie restriction in yeast: a skeptical perspective. *Ageing research reviews* 6 (2):128-140.

Kaity, B., R. Sarkar, B. Chakrabarti, and M.K. Mitra, 2018 Reprogramming, oscillations and transdifferentiation in epigenetic landscapes. *Scientific reports* 8 (1):1-12.

Kincaid, B., and E. Bossy-Wetzel, 2013 Forever young: SIRT3 a shield against mitochondrial meltdown, aging, and neurodegeneration. *Frontiers in aging neuroscience* 5:48.

Koç, A., and V.N. Gladyshev, 2007 Methionine sulfoxide reduction and the aging process.

Kondoh, H., and R. Lovell-Badge, 2015 *Sox2: biology and role in development and disease*: Academic Press.

Koonin, E.V., L. Aravind, and A.S. Kondrashov, 2000 The impact of comparative genomics on our understanding of evolution. *Cell* 101 (6):573-576.

Koren, S., B.P. Walenz, K. Berlin, J.R. Miller, N.H. Bergman *et al.*, 2017 Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research* 27 (5):722-736.

Kortschak, R.D., G. Samuel, R. Saint, and D.J. Miller, 2003 EST analysis of the cnidarian Acropora millepora reveals extensive gene loss and rapid sequence divergence in the model invertebrates. *Current Biology* 13 (24):2190-2195.

Kubota, S., 2005 Distinction of two morphotypes of Turritopsis nutricula medusae (Cnidaria, Hydrozoa, Anthomedusae) in Japan, with reference to their different abilities to revert to the hydroid stage and their distinct geographical distributions.

Kubota, S., 2011 Repeating rejuvenation in Turritopsis, an immortal hydrozoan (Cnidaria, Hydrozoa).

Kuhlbrodt, K., B. Herbarth, E. Sock, J. Enderich, I. Hermans-Borgmeyer *et al.*, 1998 Cooperative function of POU proteins and SOX proteins in glial cells. *Journal of Biological Chemistry* 273 (26):16050-16057.

Lefebvre, V., B. Dumitriu, A. Penzo-Méndez, Y. Han, and B. Pallavi, 2007 Control of cell fate and differentiation by Sry-related high-mobility-group box (Sox) transcription factors. *The international journal of biochemistry & cell biology* 39 (12):2195-2214.

Li, B., and C.N. Dewey, 2011 RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformaticss* 12 (1):323.

Li, J.-y., D.-h. Guo, P.-c. Wu, and L.-s. He, 2018 Ontogeny reversal and phylogenetic analysis of Turritopsis sp. 5 (Cnidaria, Hydrozoa, Oceaniidae), a possible new species endemic to Xiamen, China. *PeerJ* 6:e4225.

Limbert, C., G. Päth, R. Ebert, V. Rothhammer, M. Kassem *et al.*, 2011 PDX1-and NGN3-mediated in vitro reprogramming of human bone marrow-derived mesenchymal stromal cells into pancreatic endocrine lineages. *Cytotherapy* 13 (7):802-813.

Liu, H., Y. Yang, Y. Ge, J. Liu, and Y. Zhao, 2019 TERC promotes cellular inflammatory response independent of telomerase. *Nucleic acids research* 47 (15):8084-8095.

Lorian, V., 1989 In vitro simulation of in vivo conditions: physical state of the culture medium. *Journal of clinical microbiology* 27 (11):2403.

Löw, P., 2011 The role of ubiquitin–proteasome system in ageing. *General and comparative endocrinology* 172 (1):39-43.

Lundberg, A.S., W.C. Hahn, P. Gupta, and R.A. Weinberg, 2000 Genes involved in senescence and immortalization. *Current opinion in cell biology* 12 (6):705-709.

Lynch, M., and J.S. Conery, 2003 The origins of genome complexity. *science* 302 (5649):1401-1404.

Lynch, M., and B. Walsh, 2007 *The origins of genome architecture*: Sinauer Associates Sunderland, MA.

Ma, Y., X. Zhang, H. Ma, Y. Ren, Y. Sun *et al.*, 2014 Bioinformatics analysis of the four transcription factors used to induce pluripotent stem cells. *Cytotechnology* 66 (6):967-978.

Martínez, D.E., and D. Bridge, 2012 Hydra, the everlasting embryo, confronts aging. *International Journal of Developmental Biology* 56 (6-7-8):479-487.

Matsumoto, Y., and M.P. Miglietta, 2021 Cellular reprogramming and immortality: Expression profiling reveals putative genes involved in Turritopsis dohrnii's life cycle reversal. *Genome biology and evolution* 13 (7):evab136.

Matsumoto, Y., and M.P. Miglietta, In review Cellular reprogramming and immortality: Expression profiling reveals putative genes involved in Turritopsis dohrnii's life cycle reversal.

Matsumoto, Y., S. Piraino, and M.P. Miglietta, 2019 Transcriptome characterization of reverse development in Turritopsis dohrnii (Hydrozoa, Cnidaria). *G3: Genes, genomes, genetics* 9 (12):4127-4138.

Matsunaga, A., M. Tsugawa, and J. Fortes, 2008 Cloudblast: Combining mapreduce and virtualization on distributed resources for bioinformaticss applications, pp. 222-229 in *eScience, 2008. eScience'08. IEEE Fourth International Conference on*. IEEE.

Merrell, A.J., and B.Z. Stanger, 2016 Adult cell plasticity in vivo: de-differentiation and transdifferentiation are back in style. *Nature reviews Molecular cell biology* 17 (7):413.

Miglietta, M.P., and H.A. Lessios, 2009 A silent invasion. *Biological Invasions* 11 (4):825-834.

Miglietta, M.P., D. Maggioni, and Y. Matsumoto, 2018 Phylogenetics and species delimitation of two hydrozoa (phylum Cnidaria): Turritopsis (McCrady, 1857) and Pennaria (Goldfuss, 1820). *Marine Biodiversity*:1-16.

Millane, R.C., J. Kanska, D.J. Duffy, C. Seoighe, S. Cunningham *et al.*, 2011 Induced stem cell neoplasia in a cnidarian by ectopic expression of a POU domain transcription factor. *Development* 138 (12):2429-2439.

Miller, D.J., D.C. Hayward, J.S. Reece-Hoyes, I. Scholten, J. Catmull *et al.*, 2000 Pax gene diversity in the basal cnidarian Acropora millepora (Cnidaria, Anthozoa): implications for the evolution of the Pax gene family. *Proceedings of the National Academy of Sciences* 97 (9):4475-4480.

Monk, M., and C. Holding, 2001 Human embryonic genes re-expressed in cancer cells. *Oncogene* 20 (56):8085.

Moskovitz, J., S. Bar-Noy, W.M. Williams, J. Requena, B.S. Berlett *et al.*, 2001 Methionine sulfoxide reductase (MsrA) is a regulator of antioxidant defense and lifespan in mammals. *Proceedings of the National Academy of Sciences* 98 (23):12920-12925.

Motoyama, H., S. Ogawa, A. Kubo, S. Miwa, J. Nakayama *et al.*, 2009 In vitro reprogramming of adult hepatocytes into insulin-producing cells without viral vectors. *Biochemical and biophysical research communications* 385 (1):123-128.

Murshid, A., T. Eguchi, and S.K. Calderwood, 2013 Stress proteins in aging and life span. *International Journal of Hyperthermia* 29 (5):442-447.

Murthy, M., and J.L. Ram, 2015 Invertebrates as model organisms for research on aging biology. Taylor & Francis.

Neves, S.R., P.T. Ram, and R. Iyengar, 2002 G protein pathways. *science* 296 (5573):1636-1639.

Nosenko, T., F. Schreiber, M. Adamska, M. Adamski, M. Eitel *et al.*, 2013 Deep metazoan phylogeny: when different genes tell different stories. *Molecular phylogenetics and evolution* 67 (1):223-233.

Nowak, T., D. Januszkiewicz, M. Zawada, M. Pernak, K. Lewandowski *et al.*, 2006 Amplification of hTERT and hTERC genes in leukemic cells with high expression and activity of telomerase. *Oncology reports* 16 (2):301-305.

Nueda, M.J., S. Tarazona, and A. Conesa, 2014 Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. *Bioinformaticss* 30 (18):2598-2602.

Okada, T., 1991 *Transdifferentiation: flexibility in cell differentiation*: Oxford University Press, USA.

Oviedo, N.J., and W.S. Beane, 2009 Regeneration: The origin of cancer or a possible cure?, pp. 557-564 in *Seminars in cell & developmental biology*. Elsevier.

Patruno, M., M.C. Thorndyke, M.C. Carnevali, F. Bonasoro, and P.W. Beesley, 2001 Growth factors, heat-shock proteins and regeneration in echinoderms. *Journal of Experimental Biology* 204 (5):843-848.

Petrou, G., and T. Crouzier, 2018 Mucins as multifunctional building blocks of biomaterials. *Biomaterials science* 6 (9):2282-2297.

Piraino, S., F. Boero, B. Aeschbach, and V. Schmid, 1996 Reversing the life cycle: medusae transforming into polyps and cell transdifferentiation in Turritopsis nutricula (Cnidaria, Hydrozoa). *The Biological Bulletin* 190 (3):302-312.

Piraino, S., D. De Vito, J. Schmich, J. Bouillon, and F. Boero, 2004 Reverse development in Cnidaria. *Canadian Journal of Zoology* 82 (11):1748-1754.

Pop, M., 2009 Genome assembly reborn: recent computational challenges. *Briefings in bioinformaticss* 10 (4):354-366.

Pratt, W.B., Y. Morishima, H.-M. Peng, and Y. Osawa, 2010 Proposal for a role of the Hsp90/Hsp70-based chaperone machinery in making triage decisions when proteins undergo oxidative and toxic damage. *Experimental biology and medicine* 235 (3):278-289.

Prouty, N., E. Roark, N. Buster, and S.W. Ross, 2011 Growth rate and age distribution of deep-sea black corals in the Gulf of Mexico. *Marine Ecology Progress Series* 423:101-115.

Putnam, N.H., M. Srivastava, U. Hellsten, B. Dirks, J. Chapman *et al.*, 2007 Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *science* 317 (5834):86-94.

Reiter, S., M. Crescenzi, B. Galliot, and W.C. Buzgariu, 2012 Hydra, a versatile model to study the homeostatic and developmental functions of cell death. *International Journal of Developmental Biology* 56 (6-7-8):593-604.

Ridley, A.J., 2006 Rho GTPases and actin dynamics in membrane protrusions and vesicle trafficking. *Trends in cell biology* 16 (10):522-529.

Robinson, M.D., D.J. McCarthy, and G.K. Smyth, 2010 edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformaticss* 26 (1):139-140.

Rogina, B., and S.L. Helfand, 2004 Sir2 mediates longevity in the fly through a pathway related to calorie restriction. *Proceedings of the National Academy of Sciences* 101 (45):15998-16003.

Ronquist, F., M. Teslenko, P. Van Der Mark, D.L. Ayres, A. Darling *et al.*, 2012 MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology* 61 (3):539-542.

Rubin, G.M., M.D. Yandell, J.R. Wortman, G.L. Gabor, C.R. Nelson *et al.*, 2000 Comparative genomics of the eukaryotes. *science* 287 (5461):2204-2215.

Sanders, S.M., and P. Cartwright, 2015 Patterns of Wnt signaling in the life cycle of Podocoryna carnea and its implications for medusae evolution in Hydrozoa (Cnidaria). *Evolution & development* 17 (6):325-336.

Sarid, J., T.D. Halazonetis, W. Murphy, and P. Leder, 1987 Evolutionarily conserved regions of the human c-myc protein can be uncoupled from transforming activity. *Proceedings of the National Academy of Sciences* 84 (1):170-173.

Schmich, J., Y. Kraus, D. De Vito, D. Graziussi, F. Boero *et al.*, 2007 Induction of reverse development in two marine Hydrozoans. *International Journal of Developmental Biology* 51 (1):45-56.

Schmid, V., A. Bally, K. Beck, M. Haller, W. Schlage *et al.*, 1991 The extracellular matrix (mesoglea) of hydrozoan jellyfish and its ability to support cell adhesion and spreading, pp. 3-10 in *Hydrobiologia*. Springer.

Shcherbata, H.R., A.S. Yatsenko, L. Patterson, V.D. Sood, U. Nudel *et al.*, 2007 Dissecting muscle and neuronal disorders in a Drosophila model of muscular dystrophy. *The EMBO journal* 26 (2):481-493.

Shenoy, A., and R. Blelloch, 2012 microRNA induced transdifferentiation. *F1000 biology reports* 4.

Shi, Q., Z. Dong, and H. Wei, 2007 The involvement of heat shock proteins in murine liver regeneration. *Cell Mol Immunol* 4 (1):53.

Simão, F.A., R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, and E.M. Zdobnov, 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformaticss* 31 (19):3210-3212.

Slamon, D.J., and M.J. Cline, 1984 Expression of cellular oncogenes during embryonic and fetal development of the mouse. *Proceedings of the National Academy of Sciences* 81 (22):7141-7145.

Snow, B.E., N. Erdmann, J. Cruickshank, H. Goldman, R.M. Gill *et al.*, 2003 Functional conservation of the telomerase protein Est1p in humans. *Current Biology* 13 (8):698-704.

Spring, J., N. Yanze, A.M. Middel, M. Stierwald, H. Gröger *et al.*, 2000 The mesoderm specification factor twist in the life cycle of jellyfish. *Developmental biology* 228 (2):363-375.

Steele, R.E., C.N. David, and U. Technau, 2011 A genomic view of 500 million years of cnidarian evolution. *TRENDS in Genetics* 27 (1):7-13.

Sullivan, J.C., and J.R. Finnerty, 2007 A surprising abundance of human disease genes in a simple "basal" animal, the starlet sea anemone (Nematostella vectensis). *Genome* 50 (7):689-692.

Taira, T., J. Maëda, T. Onishi, H. Kitaura, S. Yoshida *et al.*, 1998 AMY-1, a novel C-MYC binding protein that stimulates transcription activity of C-MYC. *Genes to Cells* 3 (8):549-565.

Takahashi, K., K. Tanabe, M. Ohnuki, M. Narita, T. Ichisaka *et al.*, 2007 Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 131 (5):861-872.

Takahashi, K., and S. Yamanaka, 2006 Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126 (4):663-676.

Tardent, P., 1963 Regeneration in the Hydrozoa. *Biological Reviews* 38 (3):293-333.

Tian, X., D. Firsanov, Z. Zhang, Y. Cheng, L. Luo *et al.*, 2019 SIRT6 is responsible for more efficient DNA double-strand break repair in long-lived species. *Cell* 177 (3):622-638. e622.

Tissenbaum, H.A., and L. Guarente, 2001 Increased dosage of a sir-2 gene extends lifespan in Caenorhabditis elegans. *Nature* 410 (6825):227-230.

Tomczyk, S., K. Fischer, S. Austad, and B. Galliot, 2015 Hydra, a powerful model for aging studies. *Invertebrate reproduction & development* 59 (sup1):11-16.

Tower, J., 2011 Heat shock proteins and Drosophila aging. *Experimental gerontology* 46 (5):355-362.

Tseng, A.-S., and M. Levin, 2008 Tail regeneration in Xenopus laevis as a model for understanding tissue repair. *Journal of dental research* 87 (9):806-816.

Turpin, F., B. Potier, J. Dulong, P.-M. Sinet, J. Alliot *et al.*, 2011 Reduced serine racemase expression contributes to age-related deficits in hippocampal cognitive function. *Neurobiology of aging* 32 (8):1495-1504.

Uringa, E.-J., J.L. Youds, K. Lisaingo, P.M. Lansdorp, and S.J. Boulton, 2011 RTEL1: an essential helicase for telomere maintenance and the regulation of homologous recombination. *Nucleic acids research* 39 (5):1647-1655.

Uzawa, J., M. Urai, T. Baba, H. Seki, K. Taniguchi *et al.*, 2009 NMR study on a novel mucin from jellyfish in natural abundance, qniumucin from Aurelia aurita. *Journal of natural products* 72 (5):818-823.

Vaupel, J.W., A. Baudisch, M. Dölling, D.A. Roach, and J. Gampe, 2004 The case for negative senescence. *Theoretical population biology* 65 (4):339-351.

Vaziri, H., and S. Benchimol, 1998 Reconstitution of telomerase activity in normal human cells leads to elongation of telomeres and extended replicative life span. *Current Biology* 8 (5):279-282.

Venteicher, A.S., and S.E. Artandi, 2009 TCAB1: driving telomerase to Cajal bodies. *Cell Cycle* 8 (9):1329-1331.

Vinarsky, V., D.L. Atkinson, T.J. Stevenson, M.T. Keating, and S.J. Odelberg, 2005 Normal newt limb regeneration requires matrix metalloproteinase function. *Developmental biology* 279 (1):86-98.

Wang, L., A. Zheng, L. Yi, C. Xu, M. Ding *et al.*, 2004 Identification of potential nuclear reprogramming and differentiation factors by a novel selection method for cloning chromatin-binding proteins. *Biochemical and biophysical research communications* 325 (1):302-307.

Watanabe, H., R. Mättner, and T.W. Holstein, 2009 Immortality and the base of multicellular life: Lessons from cnidarian stem cells, pp. 1114-1125 in *Seminars in cell & developmental biology*. Elsevier.

Weismann, A., 1883 *Die entstehung der sexualzellen bei den hydromedusen: zugleich ein Betrag zur Kenntniss des Baues und der Lebenserscheinungen dieser Gruppe*: Fischer.

Weissbach, H., F. Etienne, T. Hoshi, S.H. Heinemann, W.T. Lowther *et al.*, 2002 Peptide methionine sulfoxide reductase: structure, mechanism of action, and biological function. *Archives of Biochemistry and Biophysics* 397 (2):172-178.

Wenger, Y., and B. Galliot, 2013 RNAseq versus genome-predicted transcriptomes: a large population of novel transcripts identified in an Illumina-454 Hydra transcriptome. *BMC genomics* 14 (1):204.

Whitaker, N.J., T.M. Bryan, P. Bonnefin, A. Chang, E.A. Musgrove *et al.*, 1995 Involvement of RB-1, p53, p16INK4 and telomerase in immortalisation of human cells. *Oncogene* 11 (5):971-976.

Wood, D.E., and S.L. Salzberg, 2014 Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology* 15 (3):R46.

Yan, J., and T. Jin, 2012 Signaling network from GPCR to the actin cytoskeleton during chemotaxis. *Bioarchitecture* 2 (1):15-18.

Yang, E.V., and S.V. Byant, 1994 Developmental regulation of a matrix metalloproteinase during regeneration of axolotl appendages. *Developmental biology* 166 (2):696-703.

Yu, J., M.A. Vodyanik, K. Smuga-Otto, J. Antosiewicz-Bourget, J.L. Frane *et al.*, 2007 Induced pluripotent stem cell lines derived from human somatic cells. *science* 318 (5858):1917-1920.

Zaghlool, A., J. Halvardson, J.J. Zhao, M. Etemadikhah, A. Kalushkova *et al.*, 2016 A role for the chromatin-remodeling factor BAZ1A in neurodevelopment. *Human mutation* 37 (9):964-975.

Zdobnov, E.M., and R. Apweiler, 2001 InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformaticss* 17 (9):847-848.

Zhang, G., C. Li, Q. Li, B. Li, D.M. Larkin *et al.*, 2014 Comparative genomics reveals insights into avian genome evolution and adaptation. *science* 346 (6215):1311-1320.

Zhang, J., G. Li, L. Feng, H. Lu, and X. Wang, 2020 Krüppel-like factors in breast cancer: Function, regulation and clinical relevance. *Biomedicine & Pharmacotherapy* 123:109778.

Zhou, X.Z., and K.P. Lu, 2001 The Pin2/TRF1-interacting protein PinX1 is a potent telomerase inhibitor. *Cell* 107 (3):347-359.

Zietara, M.S., A. Arndt, A. Geets, B. Hellemans, and F.A. Volckaert, 2000 The nuclear rDNA region of Gyrodactylus arcuatus and G. branchicus (Monogenea: Gyrodactylidae). *Journal of Parasitology* 86 (6):1368-1373.

APPENDIX II.A

SPECIMEN COLLECTION AND IDENTIFCATION, PRE-ASSEMBLY PROCESING AND POST-ASSEMBLY QUALITY ASSESSMENTS

Specimen collection and identification

All *T. dohrnii* samples were collected in July 2015 in Bocas del Toro, Panama. To reduce the genetic and sampling variability that may occur among different *T. dohrnii* individuals (e.g. collection site/time, potential of different sex, etc.), all biological replicates of the different stages (polyp, medusa, cyst, reversed polyp) originated (i.e. cut, liberated or induced by rearing after starvation) from a single colony. The mitochondrial 16S region was sequenced from the colony, resulting in a 100% identical match (e-value: 0.0) with *T. dohrnii* from the Panamanian region from NCBI. The generated 16S region was deposited to GenBank under the accession number MH029858.

Transcriptome assembly

The quality of the RNA-seq reads from each library was assessed by mapping each of the datasets to the previously published *T. dohrnii* polyp and medusa transcriptomes from the Mediterranean Sea (Matsumoto et al. 2019). As discussed in Matsumoto et al. (2019) (Matsumoto et al. 2019), the Cyst 2 dataset was excluded from the analyses due to the abnormal number of reads mapped to the polyp transcriptome (9.68%) and range of estimated paired distance values (0-3,830bp) in comparison to the other *T. dohrnii* libraries (Table A1). The remaining 11 *T. dohrnii* libraries were confirmed as viable for transcriptome assembly.

Table A1- Read mapping quality assessment of RNA-seq reads of each *T. dohrnii* library from the Bocas del Toro, Panama region using the CLC alignment algorithm. [* = from Matsumoto et al. (2019) (Matsumoto et al. 2019); [R] = dataset removed from further analyses]

	Reference: Polyps from Mediterranean Sea			Reference: Medusa from Mediterranean Sea		
Library	Mapping %	Broken Read %	Est. Paired distance	Mapping %	Broken Read %	Est. Paired distance
Cyst 1 *	92.98%	10.61%	128-594 bp	94.90%	14.36%	130-594 bp
Cyst 2 *[R]	9.68%	8.73%	0-3830 bp (few pairs mapped)	91.91%	60.54%	0-240 bp
Cyst 3 *	90.97%	14.65%	172-596 bp	88.71%	15.02%	128-596 bp
Medusa 1	91.08%	9.85%	128-618 bp	94.37%	14.30%	128-620 bp
Medusa 2	91.98%	17.19%	212-487 bp	94.45%	11.31%	128-606 bp
Medusa 3	91.32%	17.67%	223-498 bp	94.05%	12.35%	128-596 bp
Polyp 1	93.14%	15.44%	205-479 bp	91.74%	10.26%	128-582 bp
Polyp 2	93.51%	19.72%	221-510 bp	91.58%	14.55%	128-606 bp
Polyp 3	90.95%	14.31%	128-598 bp	87.00%	17.93%	128-602 bp
Reversed Polyp 1	90.39%	31.65%	178-586 bp	87.32%	31.86%	128-580 bp
Reversed Polyp 2	83.74%	34.15%	126-534 bp	82.42%	36.70%	114-573 bp
Reversed Polyp 3	78.45%	44.63%	117-603 bp	75.33%	43.68%	108-636 bp

The datasets from the eleven RNA-seq libraries (Table A2) were pooled and were fed into the Trinity software (Haas et al. 2003) totaling in ~270 million paired-end (~ 540 million total) input reads (Table A3; see Appendix L for detailed method

parameters). The reads were trimmed based on quality using Trimmomatic and the redundancy was reduced via in silico normalization within the Trinity package, resulting in total of ~13.4 million (~26.8 million) reads for assembly.

Table A2. Total number of reads for each generated *T. dohrnii* library.

Library	Total Reads	Paired
Cyst 1 *	51,524,006	25,762,003
Cyst 3 *	38,495,812	19,247,906
Medusa 1	54,698,848	27,349,424
Medusa 2	48,713,252	24,356,626
Medusa 3	58,698,848	29,349,424
Polyp 1	40,201,176	20,100,588
Polyp 2	49,104,256	24,552,128
Polyp 3	47,558,790	23,779,395
Reversed Polyp 1	51,715,980	25,857,990
Reversed Polyp 2	46,315,858	23,157,929
Reversed Polyp 3	51,298,702	25,649,351

Table A3. Total number of raw, trimmed and normalized reads from the pooled *T. dohrnii* dataset.

	Total Reads	Paired
Raw Reads	538,159,214	269,079,607
After trimming	537,256,754	268,628,377
After normalization	26,804,068	13,402,034

The first and foundational transcriptome assembly (~323.124 Mbp) resulted in a total of 416,169 transcripts, 317,473 unigenes with a GC content of 38.86% and a maximum contig length of 27,093bp. The N50 of the entire assembly was 1,329bp with a median contig length of 408bp, and an average length of 776.43bp. Based on the longest unigenes (i.e. trinity 'genes/unique transcripts'), the assembly was ~184.450 Mbp with a N50 of 757bp, median contig length of 350bp, average length of 580.99bp.

Transcriptome Quality Assessments

The outputted reads from the *in-silico normalization* within trinity were mapped back to the assembled contigs to assess the completeness of the transcriptome in respect of the sequencing reads. The following stringency parameters were utilized with the CLC Genomic Workbench v8 alignment tool in the analyses: Medium- Length fraction=0.5, Similarity fraction=0.8; Stringent= Length fraction=0.8, Similarity fraction=0.8). A high number of the reads were mapped in both analyses, 95.6% and 92.4%, respectively, indicating that very little information was lost in the unassembled reads (Table A4).

Table A4. Read mapping analyses of transcriptome using the CLC Genomic Workbench alignment tool. [LF= length fraction; SF= similarity fraction; Estimated paired distance of the reads: 128-610 bp]

	Medium (LF=0.5, SF=0.8)	**Stringent (LF=0.8, SF=0.8)**
Input Reads (normalized)	26,804,068	
Mapped Reads	25,614,004	25,757,822
Read Mapping %	95.6%	92.4%

To ensure that all libraries in each stage were well represented in the assembled transcriptome, the raw reads were individually mapped back to the assembled transcripts (Table A5). All libraries within each stage were highly represented in the transcriptome, in which the mapped read percentage ranged from 97.63-99.41% using stringent parameters. Overall, a very high percentage of the sequencing reads were incorporated into our assembled transcriptome, indicating a highly complete assembly.

Table A5: Raw paired-end reads from each library mapped back to assembled transcriptome.

Library	**Mapped Reads**	**Mapping %**
Cyst 1	51,218,630	99.41%
Cyst 2	38,196,626	99.22%
Polyp 1	39,941,300	99.35%
Polyp 2	48,645,838	99.07%
Polyp 3	46,973,254	98.77%
Medusa 1	54,243,416	99.17%
Medusa 2	48,343,380	99.25%
Medusa 3	58,129,597	99.31%
Reversed Polyp 1	51,066,990	98.75%
Reversed Polyp 2	45,218,859	97.63%
Reversed Polyp 3	50,176,926	97.81%

The BUSCO (Simão et al. 2015) tool was utilized to determine the completeness of the transcriptome in terms of gene content using the Metazoa database. The Metazoa analysis reported 97.9% completeness (95.6% complete, 2.2% partial), indicating that our assembly is highly complete in terms of gene content (Table A6).

Table A6: Gene content completeness analyses of the transcriptome using the BUSCO tool.

	Metazoa database Total BUSCOs: 978
Complete BUSCOs	935 (95.6%)
Partial BUSCOs	22 (2.2%)
Complete+Partial BUSCOs	957 (97.9%)
Missing BUSCOs	21 (2.2%)

APPENDIX II.B

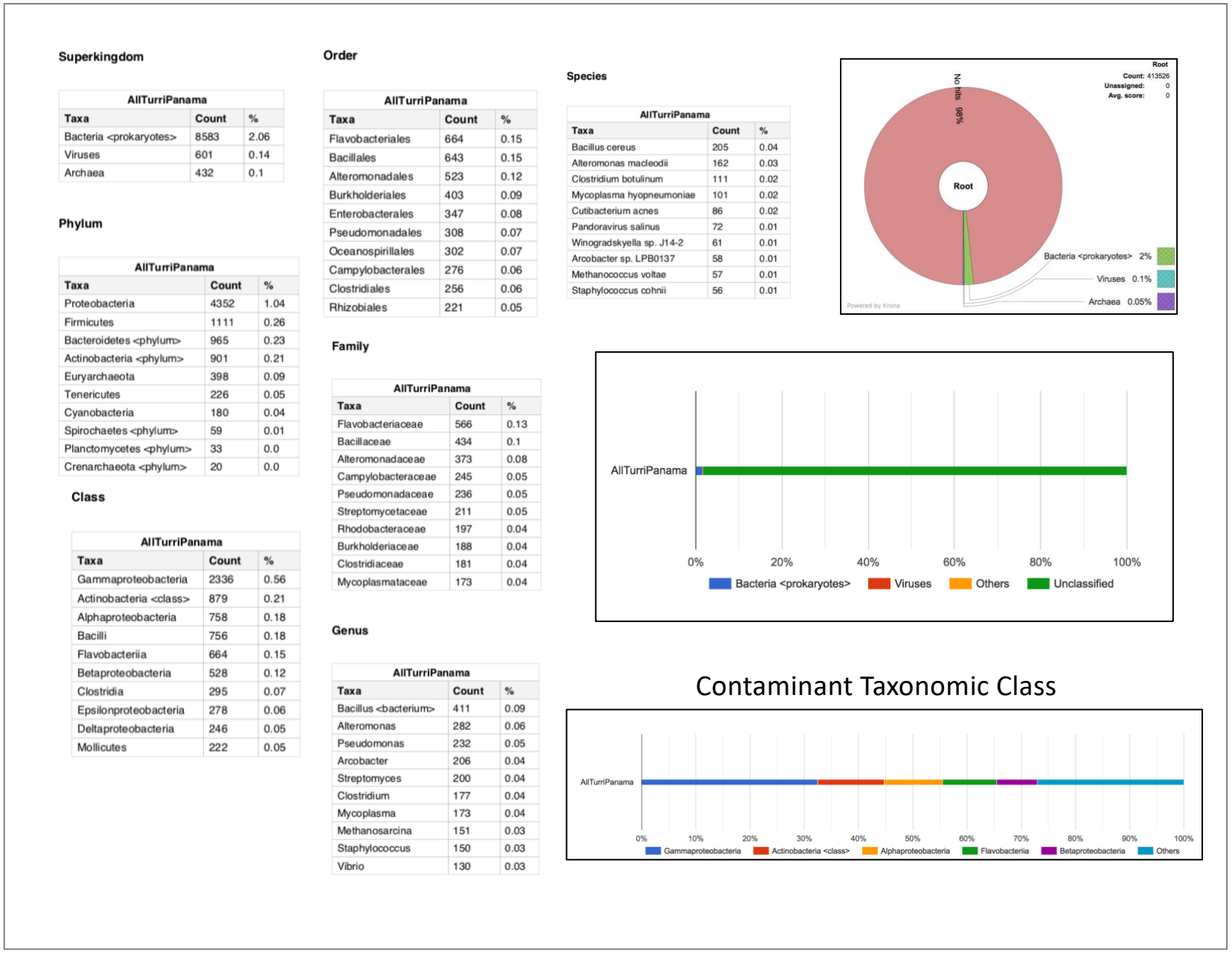
TRANSCRIPTOME TRIMMING AND FILTERING OF BIOLOGICAL CONTAMINANTS

Bacteria, Archaea and Viral contaminant removal

Though sequences Poly A-tails were excluded during cDNA library preparation to exclude the majority of bacterial sequences, the Kraken metagenomic classification tool (Wood and Salzberg 2014) was utilized to further filter contigs based on operational taxonomic units (OTUs) that belong to Bacteria, Archaea and virus species (database: All Bacterial, Archaeal and Viral Genomes in RefSeq). Out of the 416,169 assembled contigs, 9,616 sequences (2.31%) were classified to be from bacterial, archaeal and viral sources (Figure B1). Despite excluding prokaryotic sequences from our cDNA library, the metagenomic classification can be useful to provide insight on the bacterial, archaeal and viral species that are part of the microbiome within *T. dohrnii* and/or of the environment in Bocas del Toro, Panama (Atlantic). The most common contaminant taxonomic class among Bacteria, Archaea and Viral groups was Gammaproteobacteria, representing 56% of the contaminant classified contigs (Figure B1). Other common classes include Actinobacteria (21%), Alphaproteobacteria (18%) and Bacilli (18%). The new transcriptome statistics with classified sequences removed is reported in Table B1 (filtered contaminants #1).

Only contigs that were larger than 400bp were kept for further analyses. This approach is similar to that applied to other Cnidarian transcriptomes (Kitchen et al. 2015; Sanders and Cartwright 2015a, 2015b). Six duplicated sequences (i.e. exactly the same) were removed from the transcript dataset. Short contigs are less likely to provide biological meaning to our analyses as they tend to have poor coverage and quality (e.g. artifacts), often under-represented, often have no assigned protein or function (i.e. functional annotation) and can come from contaminant, non-target organisms. Though they may provide some information in RNA-seq analyses, short sequences are harder to validate with complications in statistical power, confidence and biological interpretation as mRNA transcripts are longer in natural systems. In the *T. dohrnii* cyst transcriptome that was annotated in Matsumoto et al. (2019) (Matsumoto et al. 2019), only 14.40% of the contigs shorter than 400 bp had blast hits and 9.61% were annotated with GO terms.

Our newly trimmed transcriptome (~258.305 Mbp) resulted in 206,159 transcripts and 129,607 unigenes with a GC content of s38.38% (Table B1). The new N50 of the trimmed assembly is 1,725 with a median contig length of 828 bp and an average length of 1,252.91 bp. Based on the longest unigenes (~128.976 Mbp), the N50 was 1,185 bp with a median length of 676 bp and average length of 995.11 bp.



Superkingdom

AllTurriPanama		
Taxa	Count	%
Bacteria <prokaryotes>	8583	2.06
Viruses	601	0.14
Archaea	432	0.1

Phylum

AllTurriPanama		
Taxa	Count	%
Proteobacteria	4352	1.04
Firmicutes	1111	0.26
Bacteroidetes <phylum>	965	0.23
Actinobacteria <phylum>	901	0.21
Euryarchaeota	398	0.09
Tenericutes	226	0.05
Cyanobacteria	180	0.04
Spirochaetes <phylum>	59	0.01
Planctomycetes <phylum>	33	0.0
Crenarchaeota <phylum>	20	0.0

Class

AllTurriPanama		
Taxa	Count	%
Gammaproteobacteria	2336	0.56
Actinobacteria <class>	879	0.21
Alphaproteobacteria	758	0.18
Bacilli	756	0.18
Flavobacteriia	664	0.15
Betaproteobacteria	528	0.12
Clostridia	295	0.07
Epsilonproteobacteria	278	0.06
Deltaproteobacteria	246	0.05
Mollicutes	222	0.05

Order

AllTurriPanama		
Taxa	Count	%
Flavobacteriales	664	0.15
Bacillales	643	0.15
Alteromonadales	523	0.12
Burkholderiales	403	0.09
Enterobacterales	347	0.08
Pseudomonadales	308	0.07
Oceanospirillales	302	0.07
Campylobacterales	276	0.06
Clostridiales	256	0.06
Rhizobiales	221	0.05

Family

AllTurriPanama		
Taxa	Count	%
Flavobacteriaceae	566	0.13
Bacillaceae	434	0.1
Alteromonadaceae	373	0.08
Campylobacteraceae	245	0.05
Pseudomonadaceae	236	0.05
Streptomycetaceae	211	0.05
Rhodobacteraceae	197	0.04
Burkholderiaceae	188	0.04
Clostridiaceae	181	0.04
Mycoplasmataceae	173	0.04

Genus

AllTurriPanama		
Taxa	Count	%
Bacillus <bacterium>	411	0.09
Alteromonas	282	0.06
Pseudomonas	232	0.05
Arcobacter	206	0.04
Streptomyces	200	0.04
Clostridium	177	0.04
Mycoplasma	173	0.04
Methanosarcina	151	0.03
Staphylococcus	150	0.03
Vibrio	130	0.03

Species

AllTurriPanama		
Taxa	Count	%
Bacillus cereus	205	0.04
Alteromonas macleodii	162	0.03
Clostridium botulinum	111	0.02
Mycoplasma hyopneumoniae	101	0.02
Cutibacterium acnes	86	0.02
Pandoravirus salinus	72	0.01
Winogradskyella sp. J14-2	61	0.01
Arcobacter sp. LPB0137	58	0.01
Methanococcus voltae	57	0.01
Staphylococcus cohnii	56	0.01

Figure B1: Classified contigs from Bacteria, Archaea and Viral sources (top 10 taxa from each group). [Database: All RefSeq Bacteria, Archaea and Virus genomes].

Table B1. Assembly statistics for the transcriptomes for the pooled *T. dohrnii* dataset. [Pre-trim/filter=original assembly; Filtered contaminants= Bacteria, Archaea and Viral species filtered using Kraken metagenomic classifier; Trimmed= only contigs larger than 400bp kept; Filtered contaminants #2 = biological contaminants found in annotations filtered from dataset (see methods)]

		Pre-trim/filter	Filtered contaminants #1	Trimmed (>400bp)	Filtered contaminants #2
	# of unique transcripts	317,473	310,142	129,603	127,645
	# of transcripts	416,169	406,553	206,159	204,031
	% GC	38.86%	38.72%	38.38%	38.29%
	Minimum length	201	201	401	401
	Maximum length	27,093	27,093	27,093	27,093
Based on all transcripts	**Contig N10**	4,268	4,259	4,584	4,595
	Contig N20	3,076	3,069	3,415	3,426
	Contig N30	2,347	2,342	2,696	2,706
	Contig N40	1,785	1,780	2,163	2,173
	Contig N50	1,329	1,324	1,725	1,734
	Median contig length	408	406	828	832
	Mean contig length	776.43	773.35	1,252.92	1258.07
	Total assembled bases	~323.124 Mb	~314.406 Mb	~258.301 Mb	~256.685
Based on longest unigene	**Contig N10**	3,348	3,338	3,909	3,928
	Contig N20	2,147	2,139	2,706	2,724
	Contig N30	1,472	1,462	2,013	2,028
	Contig N40	1,044	1,034	1,627	1,538
	Contig N50	757	749	1,185	1,194
	Median contig length	350	349	675	676
	Mean contig length	580.99	577.9	995.12	998.99
	Total assembled bases	~184.450 Mb	~179.231 Mb	~128.971 Mb	~127.517

BLASTx Eukaryota biological contamination removal

BLASTx using the NCBI's Non-Redundant (NR) database was performed on the filtered, deduplicated and trimmed contigs (206,159 transcripts) using a e-value cutoff of e^{-3}. The species distribution of all BLAST hits (maximum number of hits per contig: 20) indicated that the top three most represented taxa are *Stylophora pistillata* (120,307 hits), *Hydra vulgaris* (104,252 hits) and *Exaiptasia pallida* (101,621 hits) (Figure B2), and the overall top six taxa were all cnidarians (indicated in green). There are, however, three taxa that could represent biological contaminants, *Acanthamoeba*, *Thecamonas*, and *Acytostelium* species (indicated in red).

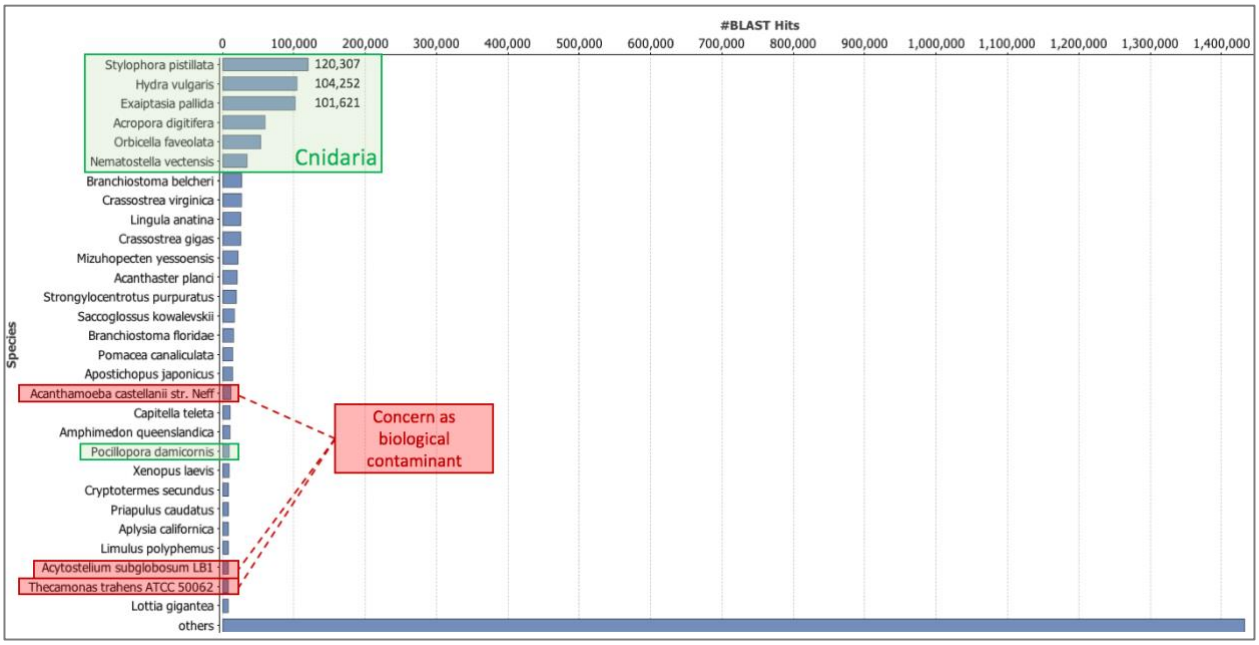


Figure B2: Species distribution of all BLASTx hits (pre-filter biological contaminant #2). [total contigs with hits: 108,060; top 20 blast hits saved; Green: Cnidarian taxa; Red: Taxa concerned as biological contaminant]

A closer look at the top-hit species distribution shows that there is an unusually high representation of fungal and protozoan species (e.g. slime molds, amoebas) (indicated in red), following or ranked between cnidarian taxa (indicated in green) (Figure B3). A total of 8,564 contigs had top-hits that belonged to the following five genera of concern: *Thecamonas*, *Acanthamoeba*, *Planoprotostelium*, *Abelmoschus* and *Acytostelium*. The sequence similarity distribution of the predicted contaminant sequences indicates a number of sequences that are highly similar (>95% sequence similarity indicated in red box, Figure B4) to the named taxa (565 sequences with top hits greater than 95%), indicating towards true biological contaminants. These sequences were removed from our transcriptome before subsequent analyses.

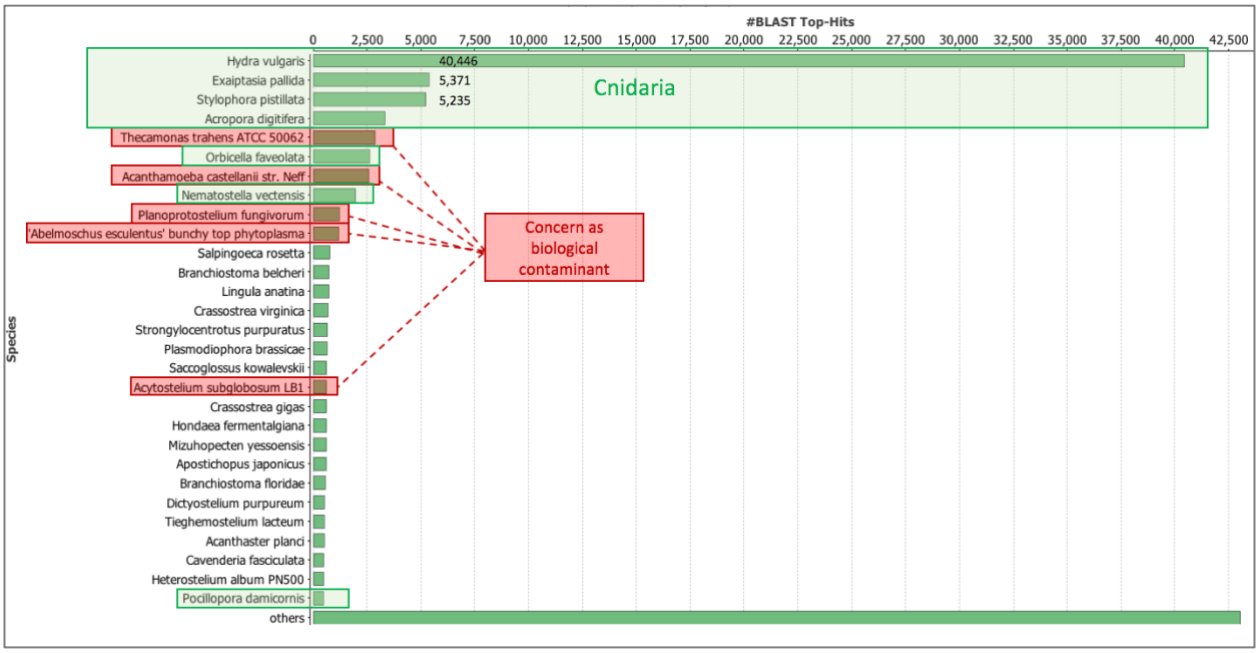


Figure B3: Species distribution of top BLASTx hits (pre-filter biological contaminant #2). [total contigs: 108,060; Green: Cnidarian taxa; Red: Taxa concerned as biological contaminant]

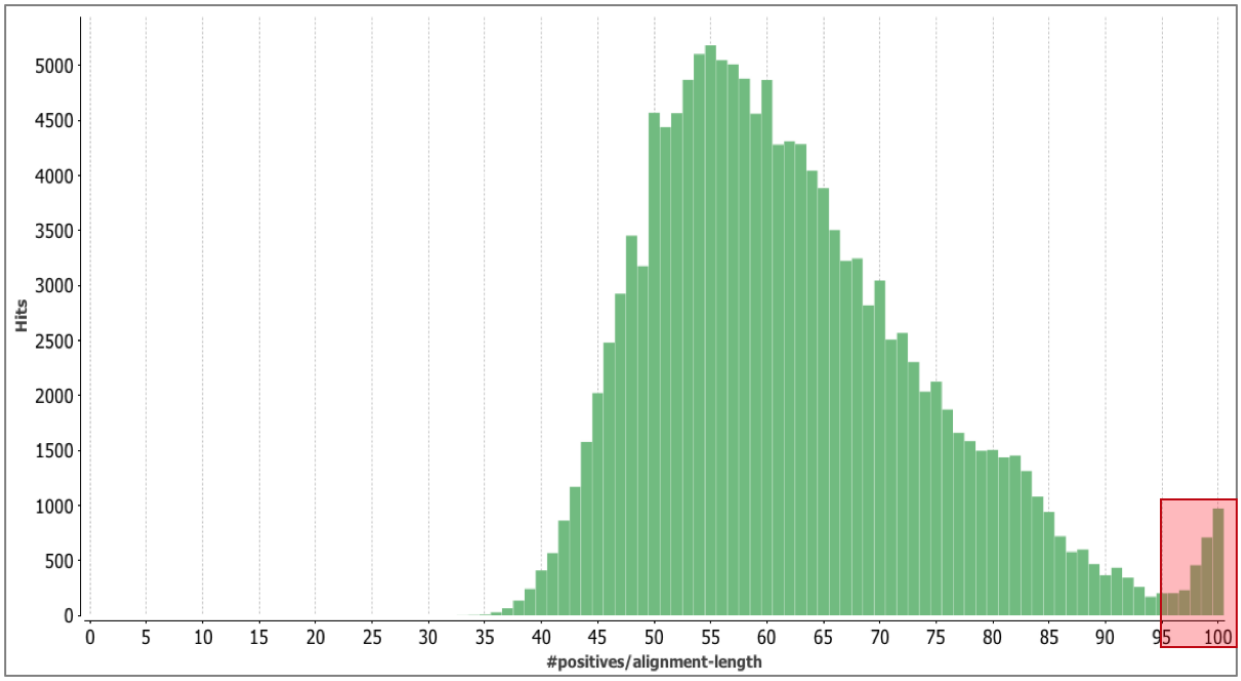


Figure B4: Sequence similarity distribution of contaminant BLASTx hits. [Genera: *Thecamonas, Acanthamoeba, Planoprotostelium, Abelmoschus* and *Acytostelium*]

Additionally, the species distribution of the hits among predicted contaminant sequences indicate that a large number of the hits belong to non-metazoan taxa, particularly fungal and protozoan, with exception of a number of metazoans (indicated in orange) (Figure B5). The contigs that incorporate metazoan hits are likely from *T. dohrnii* and not a contaminant. Among the metazoan taxa were cnidarians (*S. pistillata, E. pallida*), lancelets (*B. belcheri*) and mollusks (*L. anatina, Crassostrea* sp. and *P. canaliculata*). Due to the nature of the laboratory rearing of the cyst and reversed polyp stages, and the marine benthic environment of polyp colonies, fungal and epibiontic organisms (living associated with various taxa) are probable to be present as biological contaminants within our transcriptome. Figure B6 portrays *T. dohrnii* polyps that were growing on a crab with other epibiontic species in the shallow and tropical waters of Bocas del Toro, Panama (Atlantic). Additionally, Matsumoto et al., (2019) (Matsumoto et al. 2019) reported contamination of the foraminiferan *Reticulomyxa filosa* (Protist) in the individually constructed polyp transcriptome of *T. dohrnii* from the Mediterranean Sea, Italy. Epibiontic protists have been found on hydrozoan colonies (Bavestrello et al. 2008). To best avoid contamination from such organisms, precautions were taken during specimen collection, where individual polyp hydranths from the top of the colony with the least amount of visible fouling organisms were cut off and preserved for subsequent processing. In our newly assembled transcriptome from Bocas del Toro, Panama, there was a total of 300 sequences with top-hits that belong to the genus *Reticulomyxa* and only 2 had >95% sequence similarity with both sequencing incorporating Metazoa among the top 20 hits, confirming that there is very little concern for *R. filosa* contamination in the polyp sequencing reads.

To best ensure that the sequences belonged to actual biological contaminants, the contigs were re-blasted against the NR Metazoa database (taxid: 33208) during subsequent IPS, EggNOG and KEGG annotation analyses. 1,563 out of 7,999 contigs that had no hits against metazoan proteins were predicted to be contaminant sequences from fungal or protozoan sources, and thus removed from the transcriptome. In total, 2,128 contigs were removed from the transcriptome and a total of 204,031 contigs remain for subsequent annotation processing.

The newly filtered transcriptome (~265.685 Mbp) resulted in 204,031 transcripts and 127,645 unigenes with a GC content of 38.29% (Table 4). The new N50 of the trimmed assembly is 1,734 bp with a median contig length of 832 bp and an average length of 1,258.07 bp. Based on the longest unigenes (~127.517 Mbp), the N50 was 1,194 with a median length of 676 bp and average length of 998.99 bp.

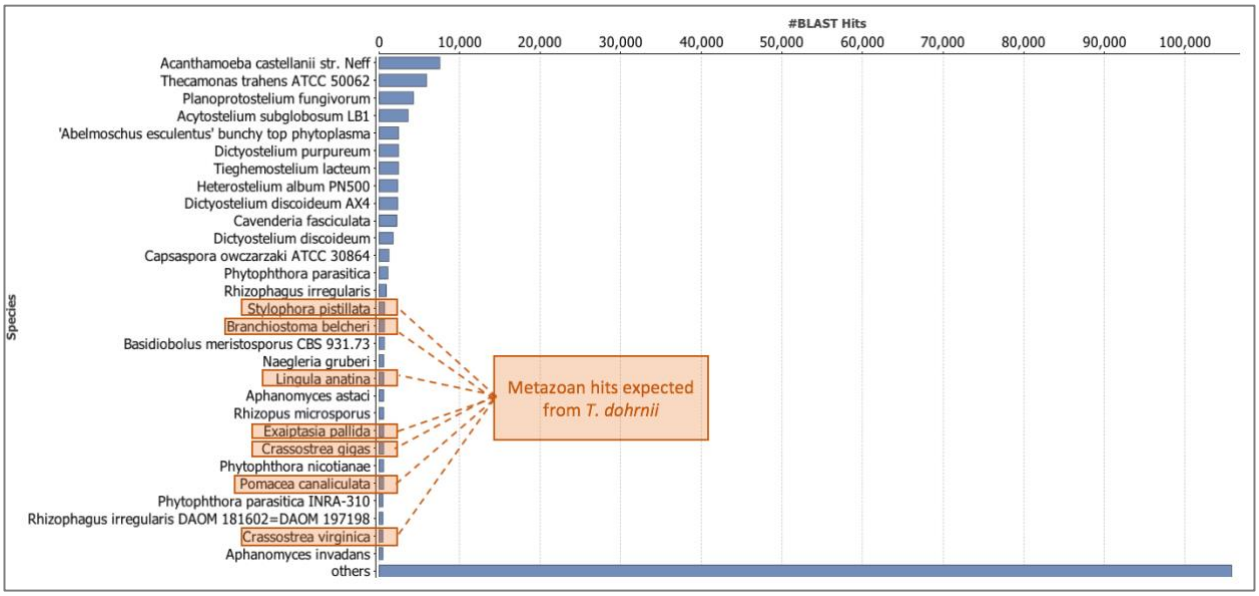


Figure B5: Species distribution of contaminant BLASTx hits. [Genera: *Thecamonas, Acanthamoeba, Planoprotostelium, Abelmoschus* and *Acytostelium;* Orange=Metazoan species]

Figure B6: *T. dohrnii* polyps growing on the surface of a crab along with other epibiontic species in Bocas del Toro, Panama (Atlantic). [Top left: Entire crab; top right: closeup of crab anterior with polyps; bottom left: closeup of the oral region of the crab with polyps; bottom right: closeup of appendages of the crab with polyps].

APPENDIX II.C

TRANSCRIPTOME FUNCTIONAL ANNOTATION

BLAST2GO pipeline (BLASTx, GO mapping, annotation)

After contaminant removal, 115,932 out of 204,031 (56.82%) contigs had BLAST hits (Table C1). Among the hits, 96,438 contigs (83.18% out of sequences with BLAST hits) were mapped with GO terms, and 72,167 contigs (74.83% out of sequences with GO terms) were B2G annotated.

Table C1: Total number of annotated contigs based on each annotation method. [Total number of contigs in transcriptome: 204,031; Blue: Cumulative total number of contigs with any type of annotation; Green: Cumulative number of contigs B2G annotated with GO term, *- not included as annotation/gene description]

Total contigs in transcriptome: 204,031		
B2G	Contigs with Blast hits	115,932
	Contigs mapped with GO terms	96,438
	Contigs B2G annotated	72,167
	Total contigs with annotation	115,932
	Total contigs with GO term	72,167
IPS	Contigs with IPS hits	149,259
	IPS hits with GO term	61,642
	Total contigs with annotation	161,717
	Total contigs with GO term	78,162
KEGG	Contigs with KEGG hits w/GO term	11,005
	Total contigs with annotation	161,717
	Total contigs with GO term	78,162
COG	Contigs with COG hits w/GO term	67,514
	Total contigs with annotation	161,717
	Total contigs with GO term	85,782
Rfam	Contigs with Rfam hits	115
	Total contigs with annotation	161,832
	Total contigs with GO term	85,897
EST	Contigs with EST hits*	418
	EST hits with Blast hits	178
	Contigs mapped with GO terms	63
	Total contigs with annotation	162,010
	Total contigs with GO term	85,960

InterProScan annotations

Out of 204,031 total contigs, 149,259 sequences (73.16%) had at least one IPS hit and 61,642 sequences were annotated with GO terms (Table C1). The results from IPS were merged with the B2G annotations to confirm previous and find new GO terms. 67,362 new GO terms were added, totaling in 293,026 GO terms found in our transcriptome. 5,995 uncharacterized sequences were newly were annotated with GO terms, and ultimately, a total of 78,162 contigs were annotated with a GO term.

KEGG annotations

The Kyoto Encyclopedia of Genes and Genomes (KEGG) database was utilized to map enzyme codes (EC) to our transcriptome. A total of 19,474 contigs were

annotated with an EC (Figure C1). Among the EC classes, hydrolases were the most abundantly present, with 11,005 annotated contigs, then transferases with 4,881 annotated contigs, and oxidoreductases with 2,379 annotated contigs.

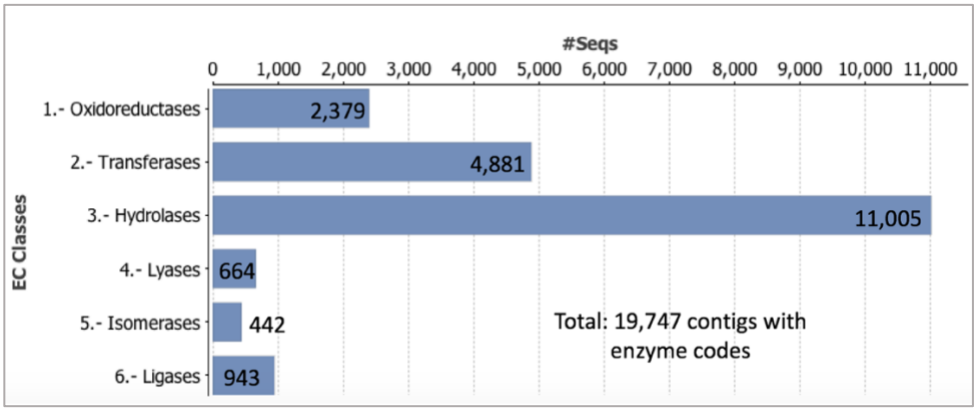


Figure C1: KEGG Enzyme code annotation distribution.

EggNOG annotations

The Conserved Orthologous Groups (COG) mapping tool against the EggNOG database within B2G was utilized to further annotate our transcriptomes with GO terms. A e-value cutoff of e^{-3} and bit score cutoff of 60 was used to only merge high quality annotations from EggNOG with our B2G and IPS annotations to confirm previous and find new GO terms. 67,514 sequences out of 204,031 total contigs (33.09%) were mapped with 1,031,536 GO terms found in the EggNOG database, and merged onto the existing B2G and IPS annotations to confirm and find new annotations. 11,551 contigs with no previous GO annotation were mapped, accumulating to 85,782 contigs mapped with at least one GO term.

RFAM annotations

Non-coding RNA (ncRNA) are mRNA sequences that do not get translated into a protein sequence, but are captured during RNA-seq as they have undergone transcription. ncRNA, also referred to previously as 'junk DNA', are common in animal genomes, but functions are not always known (Cheng et al., 2005; Birney et al., 2007; Bakel et al., 2010). Recent genomic and transcriptomic analyses have uncovered that they contribute to a number of human diseases, such as cancer and neurodegeneration (Esteller, 2011; Adams, 2017; Distefano, 2018; Lekka, 2018). The ncRNA database in Rfam was used to further characterized transcripts that had no prior annotation (i.e. no B2G, IPS, KEGG, EggNog, or hydrozoan EST annotations). A total of 42,251 contigs were analyzed. There were two uncharacterized contigs larger than 10,000 bp which was above the maximum input sequence length and could not be analyzed in Rfam. 169 new GO terms among 115 newly annotated contigs were merged to existing annotations. Among the annotated contigs, (Figure C2). Ultimately, 161,832 out of 204,031 contigs (79.32%) had at least one type of annotation (i.e. BLASTx, IPS, EggNOG, KEGG, Rfam) and 85,897 (41.10%) were annotated with GO terms.

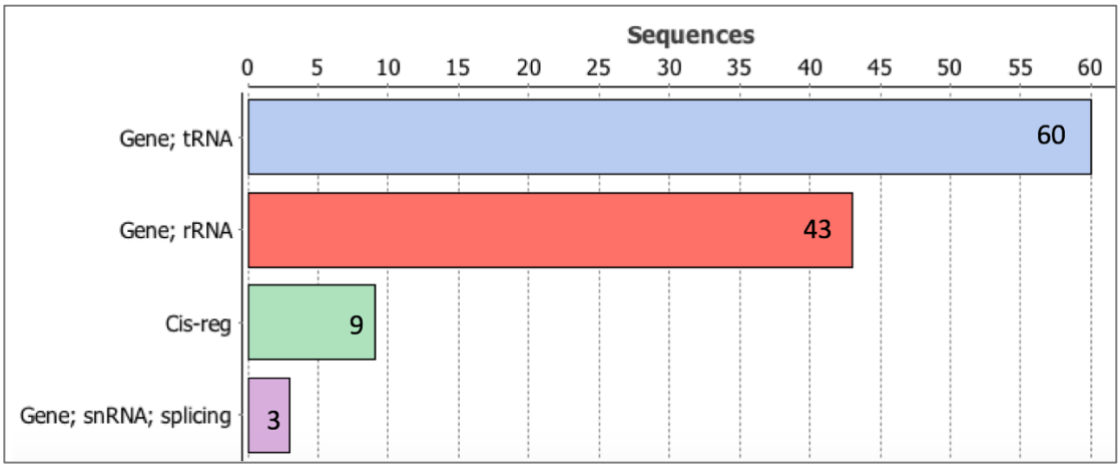


Figure C2: Rfam annotation sequence distribution where 169 new GO terms were added to 115 non-coding RNA sequences.

Hydrozoan EST/BLASTx annotations

RNA-seq data that are shorter in sequence length or constructed from lower quality reads can result in difficulties in finding the correct protein alignment. The uncharacterized sequences can first be aligned to the expressed sequence tag (EST) database, where it can then once again be compared to a protein databased to further increase the completeness of out transcriptome annotations. BLASTn using the Hydrozoa EST database (taxid: 6074, hydrozoans) was performed on transcripts that had no form of annotation (i.e. B2G, InterProScan, RFam, KEGG, EggNOG). Out of the 42,199 un-annotated transcripts, 418 transcripts had significant matches to hydrozoan EST sequences. These sequences were then compared to the Hydrozoa protein subset of the NR database (TaxID: 6074, hydrozoans) via BLASTx, and GO mapping/annotation was subsequently performed. 178 new contigs had Blast hits and 63 sequences were functionally annotated with GO terms. In total, 85,960 out of 204,031 contigs (42.13%) were annotated with GO terms and 162,010 contigs (79.40%) with any type of annotation from the utilized databases, specifically BLASTx, IPS, EggNOG, KEGG, Rfam and the Hydrozoa EST. There were ultimately 7,470 contigs with no annotation in the transcriptome that have potential to be novel transcripts from genes that have yet to be characterized. There were 7,470 total contigs over 1000 bp that did not have any annotation, particularly likely to be novel.

APPENDIX II.D

PRE-PROCESSING READS AND ALIGNMENT OF RNA-SEQ LIBRARIES

Quality trimming and alignment of RNA-seq libraries

Illumina-based platforms has been reported to be prone to the decrease of sequencing quality towards the 3' end, subsequently negatively impacting read-mapping analyses (Fuller et al. 2009). However, for differential gene expression pipelines, aggressive trimming of sequencing reads is unnecessary and often detrimental for downstream analyses (Williams et al. 2016). Prior to generating count tables for each individual library, reads were lightly trimmed based on quality from the 3' end with a phred score cutoff of 10, as recommended in (Williams et al. 2016). Though the alignment generated both gene and transcript/isoform-based expression data, analyses were performed based only on gene level-based analyses (i.e. all transcripts/isoforms are categorized with the same gene and function; based only 'unique' trinity genes). Differentiating among different outcomes of alternative splicing (i.e. based on isoform/transcript-level analyses) is both interesting and important, but with no available genome for *T. dohrnii*, distinguishing between true isoforms and fragmentation of contigs is difficult and will result in erroneous expression data. This nature of transcriptome assemblies in contrast to genome assemblies is portrayed when comparing the transcriptome and genome of the same species (Gold et al. 2019; Brekhman et al. 2015), where the number of total contigs in genomes are much higher than the number of total transcripts assembled. Using the assumption that most isoforms for a single gene will have similar GO terms and annotations, the longest isoform for each gene was used to represent GO annotations for each gene.

All libraries were normalized to eliminate systematic effects (i.e. differences in library size) and make accurate comparisons among stages and their replicates, and genes with low counts were filtered across libraries using a count per million (CPM) filter value of 1 (corresponding to approximate counts of 10-15 per gene) as recommended for DGE analyses (Chen et al. 2016). Subsequently, the maSigPro Bioconductor package (Nueda et al. 2014) for time-series DGE analyses was utilized to perform sequential DGE analyses in the following order of lifecycle stages: 1) Polyp (hydranth from colony), 2) Medusa, 3) Cyst, 4) Reversed Polyp. Differentially expressed genes were identified and categorize their gene-expression profiles (i.e. different models/patterns of gene activity) based on hierarchal clustering.

A multidimensional scaling plot (MDS) plot representing the differences in expression data between replicates and among stages was produced to confirm sample quality (Figure D1). A well-controlled experiment will portray the largest sources of expression variation to be between different lifecycle stages of replicates rather than among replicates

The MSD plot for our data portrays that the biological replicates for each stage cluster together in close proximity, particularly the Medusa stage (Figure D1, red) and the colonial Polyp stage (green), while the Cyst (purple) and the Reversed Polyp stage (blue) show the replicates being slightly more dissimilar. This could potentially be explained

by the nature of the both of the stages collected. The cyst stage is the intermediate stage during the polyp-to-medusa rejuvenation, and thus, is more difficult to be consistent despite preservation during a specific time-point in which morphological traits appear or disappear (i.e. stage defined as attached to a surface with a complete perisarc). The reversed polyp is also a stage in which consistency is slightly more difficult to obtain for similar reasons, where each individual may be at a slightly different stage of the rejuvenation (i.e. just rejuvenated vs. moved onto elongating stolons to produce more polyps) despite the best attempt to preserve each replicate with exactly the same features (i.e. stage defined as the production of a single functional polyp). Additionally, the cyst and particularly the reversed polyp stages likely incorporate more biological contaminants from culturing specimen in petri dishes, such as various fungi species reported in the pre-filtered BLAST results making samples less similar than the colonial polyp and medusa stage.

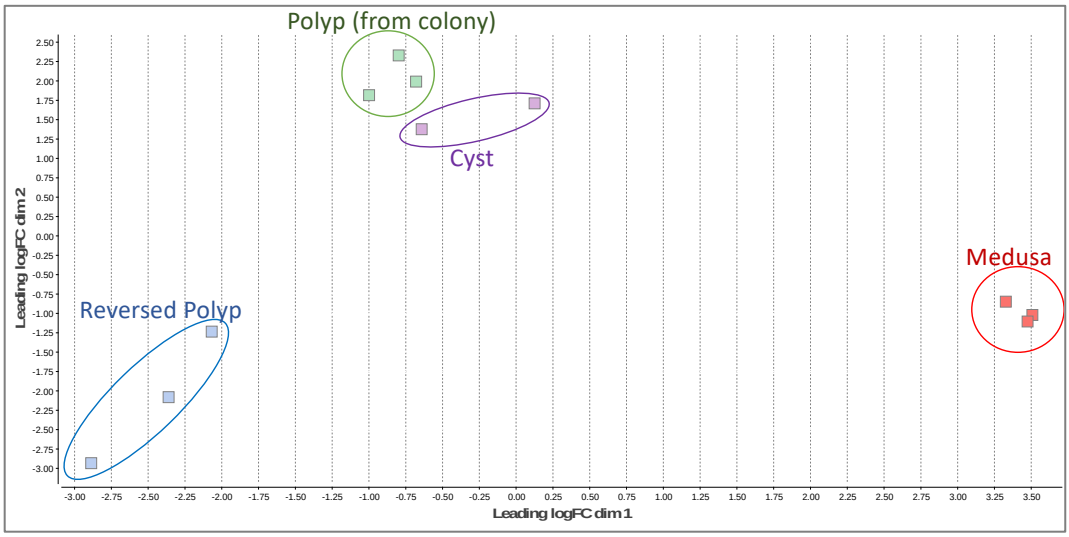


Figure D1: MDS plot of expression data of all eleven libraries used in the DGE analysis [Polyp (colony)-green, Medusa-red, Cyst-purple, Reversed Polyp-blue; LogFC- measure in change of expression, where + means upregulated and – means downregulated].

APPENDIX II.E

SEQUENTIAL LIFE HISTORY DIFFERENTIAL GENE EXPRESSION ANALYSIS

Among the significant DE genes, 1,257 genes started with a statistically significant DE genes at the first time-point (i.e. stage 1: Polyp), 1,334 genes portrayed a significant linear enrichment or repression during the reverse development sequence, and 1,445 genes portrayed a significant curved response (i.e. change in linear behavior), which could indicate transitory behavior of genes. Genes in the same hierarchical cluster show similar patterns of gene expression during the reverse development sequence of *T. dohrnii*, starting with the colonial polyp (hydranth) stage and ending with the reversed polyp.

A functional gene enrichment analysis via Fischer's Exact Test of the combined 224 genes in Cluster 5 was performed to identify specific biological processes that were the enriched and suppressed in the Cyst. There was a total of 209 over- and 3 under-expressed categories within the Cyst and 74 over- and 3 under-expressed when simplified to the most specific biological category (i.e. most specific child term within a GO lineage). The reduced dataset was sorted from the most significantly differentially expressed GO term and visualized in an enrichment chart. The most specific enriched biological process in the Cluster 5 was 'Nematode larval development (GO:0002119)' and 'Positive regulation of growth rate (GO:0030307)' (Figure E1). Though *T. dohrnii* is evolutionarily distant from nematodes, comparative embryology (i.e. similarities in embryos among animals) indicates that there may be developmental networks that are similar among animals (Garfield and Wray 2009; Kuo 2019; Richards 2009). Additionally, the Cyst has commonalities with planulae, as both stages precede the juvenile polyp stage (Efroni et al. 2016; Harland 2018; Martinet et al. 2016). Furthermore, categories related the response to DNA damage and protein monoubiquitination were found to be highly enriched. Processes associated to cytoskeleton and chromosome organization, both specific child-GO terms of broader mitotic cell division processes, were also found to be suppressed (Appendix II.F).

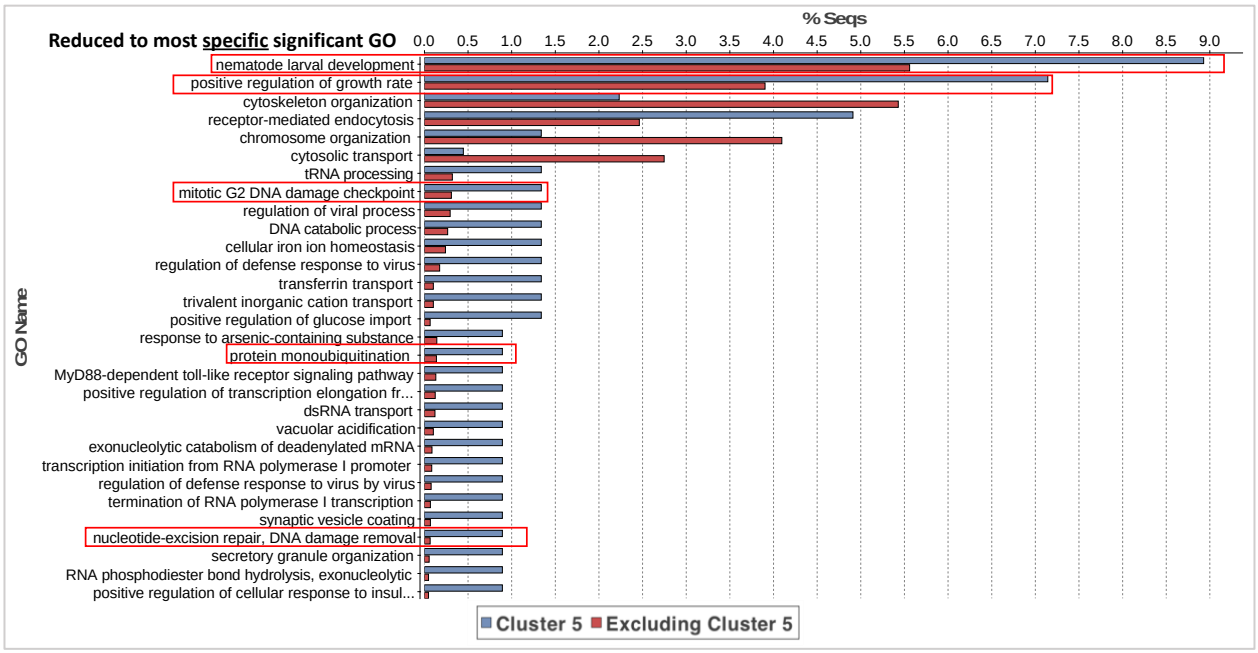


Figure E1: Functional gene enrichment analysis of Cluster 5, where significantly enriched and suppressed categories were reduced to the most specific terms (sorted by: highest in Cluster 5 or excluding Cluster 5).

APPENDIX II.F

List of DEGs from the sequential DGE analysis discussed in Chapter II is included as a separate file.

APPENDIX II.G

PAIR-WISE DIFFERENTIAL GENE EXPRESSION ANALYSES

Colonial Polyp vs. Reversed Polyp

A total of 2,905 DE gene were identified when the Reversed Polyp was compared to the Polyp stage, where 1,622 genes were up-regulated in the Reversed Polyp stage and 1,283 were downregulated (i.e. upregulated in the Polyp stage) (Figure 5A in main text). There were substantially less DE genes in the Reversed Polyp vs. Polyp analysis in comparison to the Medusa vs. Polyp. It was expected that the Medusa and Polyp would have more differences in genetic networks and biological processes as the stages occupy completely different niches and reproductive strategies (i.e. planktonic, solitary, sexual reproduction vs. benthic, colonial, asexual reproduction), while the Reversed Polyp and Polyp are the same lifecycle stages. It was noted that there is possibility of leftover biological contaminants from fungal species in the Reversed Polyp stage that were unable to be filtered during the previous BLASTx step and may contribute to genes being up-regulated in the Reversed Polyp. Some of the enriched genes, such as reproduction and symbiotic process related-genes, may be a result of contaminants introduced from the rearing process, but genuine asexual reproduction processes or other symbiotic processes cannot be ruled out.

Medusa vs. Colonial Polyp

The Reversed Polyp libraries were not used in the comparison to aim for a more accurate analysis, as there was potential leftover fungal and contaminants that may have resulted from the culturing process. A total of 4,486 DE gene were identified when the Medusa was compared to the Polyp stage, where 2,432 genes were up-regulated in the Medusa stage and 2,053 were downregulated (i.e. upregulated in the Polyp stage). Due to the large number of DE genes, the top 50 DE genes were used as our introductory investigation of the differences in the genetic and biological networks of the Medusa and the Polyp stage.

APPENDIX II.H

List of DEGs from the Pair-wise DGE analyses discussed in Chapter II is included as a separate file.

APPENDIX II.I

List of enriched genes from the Pair-wise DGE analysis of the Medusa vs. Polyp analysis discussed in Chapter II is included as a separate file.

APPENDIX III.A

List of SIRT sequences identified among super-transcripts discussed in Chapter III is included as a separate file.

APPENDIX III.B

List of telomere and telomerase-related sequences identified among super-transcripts discussed in Chapter III is included as a separate file.

APPENDIX III.C

List of HSP70 and HSP90 sequences identified among super-transcripts discussed in Chapter III is included as a separate file.

APPENDIX III.D

YAMANAKA (OCT4, SOX2, KLF4, C-MYC) AND THOMPSON FACTORS (LIN28, NANOG) ANALYSES

Table D1: Function and significance of Yamanaka factors. Oct4, Sox2, Klf4 and c-Myc are the core transcription factors that induced pluripotent stem cells in mammalian somatic cells.

Transcription factor	Function
Oct4 (Octamer-binding transcription factor 4) Other names: Oct-3, POU5	Mammalian transcription factor encoded by the POU5F1 gene that is associated with the maintenance and renewal of undifferentiated embryonic stem cells that remains active during embryonic development (Pan et al. 2002; Takahashi et al. 2007; Takahashi and Yamanaka 2006; Shi and Jin 2010). Among the four Yamanaka factors, Oct4 is indispensable and fundamental to induce a pluripotent state, where the transcript alone was capable of inducing pluripotency in neural cells in an experimental setting (Kim et al. 2009a; Kim et al. 2009b). Highly expressed and serves as a repressor of cellular specialization, and gets silenced as the cell undergoes differentiation (Pesce and Schöler 2001; Zaehres et al. 2005). Additionally, has been reported to promote tumorigenesis by inhibiting apoptosis (Tai et al. 2005; Chiou et al. 2008; Saigusa et al. 2009).
Sox2 (Sex determining region Y-box 2)	Similar to and often found associating with Oct4, the Sox2 transcription factor is also associated with the maintenance of pluripotency in undifferentiated embryonic and neural stem cells, and plays a fundamental role in embryonic development in mammals (Avilion et al. 2003; Kiernan et al. 2005). Sox2 and Oct4 interact and bind DNA cooperatively to regulate genes involved in stem cells (Rizzino 2013). Plays a large role in cell fate determination and stem cells in the central nervous system. Additionally, has been involved in the formation of cancers and tumors (Liu et al. 2013; Boumahdi et al. 2014).
c-Myc Other names: Myc	Proto-oncogene (mutation contributes to cancer) that plays a large role in regulating gene expression associated with cell transformation and the maintenance of the cell cycle, though modifying apoptosis and metabolism (Kerr et al. 1994; Dang 1999; Dang et al. 1999). This regulator is known to universally control 15% of all genes in humans (Gearhart et al. 2007), and is highly expressed in many cancerous cells and activates growth factor-related genes (Dubik and Shiu 1992; Chen et al. 2001).
Klf4 (Krueppel-like factor 4)	Transcription factor that is associated with cell proliferation, apoptosis, differentiation and somatic cell reprogramming; play a large role in regenerative responses to DNA damage and maintaining cellular stability; Involved in embryonic development (Segre et al. 1999; El-Karim et al. 2013). Known to promote the formation of tumors but also has been shown to be an oncogenic suppressor (Rowland et al. 2005; Rowland and Peeper 2006; Guan et al. 2010). Reported to play a large role in development of bones and the overall skeletal system (Kim et al. 2014).

Table D2: tBlastx analysis of the Yamanaka and Thomson genes (abbreviated as Thom.) in the *T. dohrnii* transcriptome. All 12 query sequences, which included all variants of the six transcription factors reported in GenBank, had hits with regions of high scoring pairs (HSPs). The majority of variants had similar or exactly the same sequence hits, bit-score and e-value, as for c-Myc, Klf4, and Nanog. However, Oct4 had two sequences among the variants in our T. dohrnii transcriptome that had the lowest e-value and highest bit score, one being 1,614 bp in length (variant 1 and 3) and the other 337 bp in length (variant 2 and 4). In addition, the sequence for Oct4 variant 2 and 4 was an isoform of the Lin28 hit sequence in T. dohrnii, which was also a short 323 bp sequence. Only sequences that were larger than 400 bp were utilized for the subsequent annotation search.

	Query	GenBank ID	Query Length	# of HSPs	Lowest E-value	Highest Bit Score	Sequence in *T. dohrnii*	Sequence length
Yamanaka Factors	c-Myc variant 1	NM_002467.6	4515	11	7.66E-21	103.71	TRINITY_DN99862_c1_g1_i1	2172
	c-Myc variant 2	NM_001354870.1	3721	12	1.22E-20			
	Klf4 variant 1	NM_001314052.1	3014	843	6.57E-48	193.97	TRINITY_DN110215_c0_g1_i7	4156
	Klf4 variant 2	NM_004235.5	2903	846	6.32E-48			
	Oct4 variant 1	NM_002701.5	1430	27	9.95E-32	127.99	TRINITY_DN89582_c0_g1_i1	1614
	Oct4 variant 3	NM_001173531.2	1589	100	3.93E-31	127.08		
	Oct4 variant 2	NM_203289.5	2075	493	5.85E-35	138.99	TRINITY_DN103581_c10_g1_i2	337
	Oct4 variant 4	NM_001285986.1	2300	483	6.54E-35			
	Sox2	NM_003106.3	2520	51	4.66E-42	174.27	TRINITY_DN102764_c1_g1_i4	6567
Thom.	Lin28	NM_024674.5	4024	850	4.34E-35	151.82	TRINITY_DN103581_c10_g1_i1	323
	Nanog variant 1	NM_024865.3	2103	329	5.38E-34	139.91	TRINITY_DN107889_c3_g1_i3	225
	Nanog variant 2	NM_001297698.1	2055	324	5.25E-34			

Table D3: RNA-seq analysis of the Yamanaka and Thomson factors in *T. dohrnii*.
[Input reads: all reads from *T. dohrnii*: 538,159,214 (269,079,607 paired-end) reads; LF= length fraction, SF=similarity fraction]

	Query	GenBank ID	Query Length	Consensus length	# reads mapped	Single reads	Paired reads	Avg. Coverage
SF: 0.3, LF: 0.3	c-Myc variant 1	NM_002467.6	4515	3719	806123	744081	62042	3,694.27
	Klf4 variant 1	NM_001314052.1	3014	3011	11940732	8619924	3320808	117,184.37
	Oct4 variant 1	NM_002701.5	2300	2298	626809	599767	27042	4,538.73
	Sox2	NM_003106.3	2520	2521	408620	387684	20936	2,726.81
	Lin28	NM_024674.5	4024	4020	1924579	1849325	75254	9,342.08
	Nanog variant 1	NM_024865.3	2103	2102	2300200	2271906	28294	21,434.09
SF: 0.5, LF: 0.5	c-Myc variant 1	NM_002467.6	4515	405	578	564	14	7.68
	Klf4 variant 1	NM_001314052.1	3014	556	464,486	459,564	4,922	6,813.09
	Oct4 variant 1	NM_002701.5	2300	348	1,940	1,850	90	97.65
	Sox2	NM_003106.3	2520	559	137	137	0	2.68
	Lin28	NM_024674.5	4024	842	2194	2084	110	51.31
	Nanog variant 1	NM_024865.3	2103	433	1,508	1,480	28	63.96
SF: 0.7, LF: 0.7	c-Myc variant 1	NM_002467.6	4515	0	0	0	0	0
	Klf4 variant 1	NM_001314052.1	3014	0	0	0	0	0
	Oct4 variant 1	NM_002701.5	2300	325	1497	1351	146	85.3
	Sox2	NM_003106.3	2520	100	2	2	0	0.08
	Lin28	NM_024674.5	4024	632	1224	1144	80	37.03
	Nanog variant 1	NM_024865.3	2103	321	619	537	82	38
SF: 0.9, LF: 0.9	c-Myc variant 1	NM_002467.6	4515	0	0	0	0	0
	Klf4 variant 1	NM_001314052.1	3014	0	0	0	0	0
	Oct4 variant 1	NM_002701.5	2300	308	52	52	0	3.27
	Sox2	NM_003106.3	2520	0	0	0	0	0
	Lin28	NM_024674.5	4024	326	8	8	0	0.28
	Nanog variant 1	NM_024865.3	2103	307	92	90	2	6.4

APPENDIX III.E

List of Yamanaka Factor gene family sequences identified among super-transcripts discussed in Chapter III is included as a separate file.

APPENDIX IV.A

Appendix IV.A: N50 statistics and BUSCO scores of all seven assembly trails from genomic HiFi reads.

	Genomic HiFi long-reads assembly						
	Canu #1	**Canu #2**	**Flye #1**	**Flye #2**	**SPAdes #1**	**SPAdes #2**	**IPA Assembler**
#Contigs	10,734	9,380	2,996	3,090	208,726	48,408	1,723
Total Size (bp)	312,401,202	328,234,377	211,363,623	231,540,628	714,570,551	133,432,306	348,119,487
N50	50,686	64,429	126,531	170,348	42,010	63,343	347,125
Mean	29,104	34,993	70,549	74,811	3,423	2,756	202,043
Median	14,465	16,211	39,545	35,896	258	218	98,823
GC%	40.47	39.88	39.27	39.27	40.90	38.60	42.21
Largest	4,990,683	5,108,046	5,116,097	5,116,225	1,415,270	3,375,708	5,127,597
Shortest	1,028	1,060	609	204	128	128	15,656
BUSCO	**C+F: 11.9% (115)**	**C+F: 14.8% (141)**	**C+F: 11.3 (108)**	**C+F: 12.8% (122)**	**C+F: 17.7% (168)**	**C+F: 12.8% (122)**	**C+F: 12.4% (118)**
	C: 10.1% (97)	C: 12.1% (115)	C: 9.6% (92)	C: 11.1% (106)	C: 13.6% (129)	C: 11.1% (106)	C: 11.0% (105)
	S: 5.2% (50)	S: 6.1% (58)	S: 4.7% (45)	S: 5.8% (55)	S: 7.7% (73)	S: 5.8% (55)	S: 5.0 (48)
	D: 4.9% (47)	D: 6.0% (57)	D: 4.9% (47)	D: 5.3% (51)	D: 5.9% (56)	D: 5.3% (51)	D: 6.0% (57)
	F: 1.8% (17)	F: 2.7% (26)	F: 1.7% (16)	F: 1.7% (16)	F: 4.1% (39)	F: 1.7% (16)	F: 1.4% (13)
	M: 88.1% (840)	M: 85.2% (813)	M: 88.7% (846)	M: 87.2% (832)	M: 82.3% (786)	M: 832 (87.2)	M: 87.6% (836)

APPENDIX IV.B

Appendix IV.B: BLASTn using the NT database (all taxa) on the reads classified in Kraken as microbial reads. Reads that did not have hits against the NT database were placed into the Assembly Dataset (total 880,830 reads), presumed to belong to our target specimen.

All Corrected Reads- NT BLAST (all taxa)			
Fasta File	**Sequences**	**w/ BLAST**	**No BLAST**
All_Cor0	200,000	161,615	38,385
All_Cor200000	200,000	160,566	39,434
All_Cor400000	200,000	162,638	37,362
All_Cor600000	200,000	161,808	38,192
All_Cor800000	200,000	161,740	38,260
All_Cor1000000	200,000	161,391	38,609
All_Cor1200000	200,000	161,948	38,052
All_Cor1400000	200,000	161285	38,715
All_Cor1600000	200,000	166,725	33,275
All_Cor1800000	200,000	162,134	37,866
All_Cor2000000	200,000	160,102	39,898
All_Cor2200000	200,000	159,073	40,927
All_Cor2400000	200,000	160,160	39,840
All_Cor2600000	200,000	159,912	40,088
All_Cor2800000	200,000	161,869	38,131
All_Cor3000000	200,000	161,362	38,638
All_Cor3200000	200,000	161,457	38,543
All_Cor3400000	200,000	159,870	40,130
All_Cor3600000	200,000	160,827	39,173
All_Cor3800000	200,000	159,540	40,460
All_Cor4000000	200,000	160,087	39,913
All_Cor4200000	200,000	158,441	41,559
All_Cor4400000	117,631	92,251	25,380
Total:	**4,517,631**	**3,636,801**	**880,830**

APPENDIX IV.C

Appendix IV.C: BLASTn using the Metazoa (taxid: 33208) in NT database on the reads with hits against the NT database. Reads that had hits against the Metazoa database were placed into the Assembly Dataset (total 895,973 reads), presumed to belong to our target specimen.

Corrected Reads (All cells)- Metazoa BLAST			
Fasta File	**Sequences**	**w/ BLAST**	**No BLAST**
All_Cor0	161,615	42,261	119,354
All_Cor200000	160,566	42,863	117,703
All_Cor400000	162,638	42,752	119,886
All_Cor600000	161,808	42,465	119,343
All_Cor800000	161,740	43,403	118,337
All_Cor1000000	161,391	43,389	118,002
All_Cor1200000	161,948	43,721	118,227
All_Cor1400000	161,285	42,739	118,546
All_Cor1600000	166,725	45,497	121,228
All_Cor1800000	162,134	40,893	121,241
All_Cor2000000	160,102	36,238	123,864
All_Cor2200000	159,073	36,788	122,285
All_Cor2400000	160,160	37,054	123,106
All_Cor2600000	159,912	38,086	121,826
All_Cor2800000	161,869	37,763	124,106
All_Cor3000000	161,362	37,803	123,559
All_Cor3200000	161,457	38,027	123,430
All_Cor3400000	159,870	37,198	122,672
All_Cor3600000	160,827	37,238	123,589
All_Cor3800000	159,540	36,644	122,896
All_Cor4000000	160,087	36,231	123,856
All_Cor4200000	158,441	36,567	121,874
All_Cor4400000	92,251	20,353	71,898
Total:	**3,636,801**	**895,973**	**2,740,828**

APPENDIX IV.D

Appendix IV.D: BLASTn using NT database (all taxa) and extract reads that had a e-value greater than 0 (so excludes 0 e-value reads), and less than 90% sequence identity against the top microbial BLAST hit. Reads that met these criteria were placed into the Assembly Dataset (total 1,622,274 reads). This step ensured there was no removal of reads that showed conserved/similar regions between my target organism and microbial taxa.

Corrected Reads (All cells)- Metazoa BLAST			
Fasta File	**Sequences**	**E-value 0; Similarity ≥90%**	**Selected for Assembly**
All_Cor0	119,354	51,875	67,479
All_Cor200000	117,703	51,054	66,649
All_Cor400000	119,886	52,378	67,508
All_Cor600000	119,343	51,236	68,107
All_Cor800000	118,337	52,629	65,708
All_Cor1000000	118,002	51,435	66,567
All_Cor1200000	118,227	52,535	65,692
All_Cor1400000	118,546	51,966	66,580
All_Cor1600000	121,228	52,195	69,033
All_Cor1800000	121,241	49,025	72,216
All_Cor2000000	123,864	48,352	75,512
All_Cor2200000	122,285	47,993	74,292
All_Cor2400000	123,106	47,500	75,606
All_Cor2600000	121,826	47,383	74,443
All_Cor2800000	124,106	48,620	75,486
All_Cor3000000	123,559	47,247	76,312
All_Cor3200000	123,430	47,630	75,800
All_Cor3400000	122,672	47,003	75,669
All_Cor3600000	123,589	48,400	75,189
All_Cor3800000	122,896	47,451	75,445
All_Cor4000000	123,856	49,004	74,852
All_Cor4200000	121,874	47,202	74,672
All_Cor4400000	71,898	28,441	43,457
Total:	**2,740,828**	**1,118,554**	**1,622,274**

APPENDICES REFERENCES

Abad, M., L. Mosteiro, C. Pantoja, M. Canamero, T. Rayon *et al.*, 2013 Reprogramming in vivo produces teratomas and iPS cells with totipotency features. *Nature* 502 (7471):340-345.

Adachi, K., H. Miyake, T. Kuramochi, K. Mizusawa, and S.-i. Okumura, 2017 Genome size distribution in phylum Cnidaria. *Fisheries science* 83 (1):107-112.

Adams, J.C., and A. Brancaccio, 2015 The evolution of the dystroglycan complex, a major mediator of muscle integrity. *Biology open* 4 (9):1163-1179.

Adler, A.S., S. Sinha, T.L. Kawahara, J.Y. Zhang, E. Segal *et al.*, 2007 Motif module map reveals enforcement of aging by continual NF-κB activity. *Genes & development* 21 (24):3244-3257.

Al-Shahrour, F., R. Díaz-Uriarte, and J. Dopazo, 2004 FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics* 20 (4):578-580.

Alvarado, A.S., and S. Yamanaka, 2014 Rethinking differentiation: stem cells, regeneration, and plasticity. *Cell* 157 (1):110-119.

Amano, H., and E. Sahin, 2019 Telomeres and sirtuins: at the end we meet again. *Molecular & cellular oncology* 6 (5):e1632613.

Anandarajan, M., S. Paulraj, and R. Tubman, 2009 ABCA3 Deficiency: an unusual cause of respiratory distress in the newborn. *The Ulster medical journal* 78 (1):51.

Avilion, A.A., S.K. Nicolis, L.H. Pevny, L. Perez, N. Vivian *et al.*, 2003 Multipotent cell lineages in early mouse development depend on SOX2 function. *Genes & development* 17 (1):126-140.

Babicki, S., D. Arndt, A. Marcu, Y. Liang, J.R. Grant *et al.*, 2016 Heatmapper: web-enabled heat mapping for all. *Nucleic acids research* 44 (W1):W147-W153.

Bai, S., R. Thummel, A.R. Godwin, H. Nagase, Y. Itoh *et al.*, 2005 Matrix metalloproteinase expression and function during fin regeneration in zebrafish: analysis of MT1-MMP, MMP2 and TIMP2. *Matrix biology* 24 (4):247-260.

Bankevich, A., S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin *et al.*, 2012 SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology* 19 (5):455-477.

Barber, L.J., J.L. Youds, J.D. Ward, M.J. McIlwraith, N.J. O'Neil *et al.*, 2008 RTEL1 maintains genomic stability by suppressing homologous recombination. *Cell* 135 (2):261-271.

Bavestrello, G., C. Cerrano, C. Di Camillo, S. Puce, T. Romagnoli *et al.*, 2008 The ecology of protists epibiontic on marine hydroids. *Journal of the marine Biological Association of the United Kingdom* 88 (8):1611-1617.

Bavestrello, G., C. Sommer, and M. Sarà, 1992 Bi-directional conversion in Turritopsis nutricula (Hydrozoa). *Scientia Marina* 56 (2-3):137-140.

Beausejour, C.M., and J. Campisi, 2006 Ageing: balancing regeneration and cancer. *Nature* 443 (7110):404.

Bellizzi, D., S. Dato, P. Cavalcante, G. Covello, F. Di Cianni *et al.*, 2007 Characterization of a bidirectional promoter shared between two human genes related to aging: SIRT3 and PSMD13. *Genomics* 89 (1):143-150.

Bellizzi, D., G. Rose, P. Cavalcante, G. Covello, S. Dato *et al.*, 2005 A novel VNTR enhancer within the SIRT3 gene, a human homologue of SIR2, is associated with survival at oldest ages. *Genomics* 85 (2):258-263.

Bessereau, J.-L., 2006 Transposons in C. elegans. *WormBook*:1.

Blasco, M.A., 2007 Telomere length, stem cells and aging. *Nature chemical biology* 3 (10):640.

Bodnar, A.G., M. Ouellette, M. Frolkis, S.E. Holt, C.-P. Chiu *et al.*, 1998 Extension of life-span by introduction of telomerase into normal human cells. *science* 279 (5349):349-352.

Borradaile, N.M., A. Watson, and J.G. Pickering, 2011 Regeneration and Aging: Regulation by Sirtuins and the NAD+ Salvage Pathway, pp. 289-298 in *Regenerative Nephrology*. Elsevier.

Bosch, T.C., A. Klimovich, T. Domazet-Lošo, S. Gründer, T.W. Holstein *et al.*, 2017 Back to the basics: cnidarians start to fire. *Trends in neurosciences* 40 (2):92-105.

Boumahdi, S., G. Driessens, G. Lapouge, S. Rorive, D. Nassar *et al.*, 2014 SOX2 controls tumour initiation and cancer stem-cell functions in squamous-cell carcinoma. *Nature* 511 (7508):246.

Bowles, J., G. Schepers, and P. Koopman, 2000 Phylogeny of the SOX family of developmental transcription factors based on sequence and structural indicators. *Developmental biology* 227 (2):239-255.

Brekhman, V., A. Malik, B. Haas, N. Sher, and T. Lotan, 2015 Transcriptome profiling of the dynamic life cycle of the scypohozoan jellyfish Aurelia aurita. *BMC genomics* 16 (1):74.

Calderwood, S., 2007 Molecular chaperones and the ubiquitin proteasome system in aging. *The Ubiquitin Proteasome System in the Central Nervous System. Nova*:537-552.

Calvo, R., and H.A. Drabkin, 2000 Embryonic genes in cancer. *Annals of oncology* 11:207-218.

Chapman, J.A., E.F. Kirkness, O. Simakov, S.E. Hampson, T. Mitros *et al.*, 2010 The dynamic genome of Hydra. *Nature* 464 (7288):592.

Chappell, J., and S. Dalton, 2013 Roles for MYC in the establishment and maintenance of pluripotency. *Cold Spring Harbor perspectives in medicine* 3 (12):a014381.

Chen, C.-R., Y. Kang, and J. Massagué, 2001 Defective repression of c-myc in breast cancer cells: a loss at the core of the transforming growth factor β growth arrest program. *Proceedings of the National Academy of Sciences* 98 (3):992-999.

Chen, S., and G. Parmigiani, 2007 Meta-analysis of BRCA1 and BRCA2 penetrance. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology* 25 (11):1329.

Chen, Y., A.T. Lun, and G.K. Smyth, 2016 From reads to genes to pathways: differential expression analysis of RNA-Seq experiments using Rsubread and the edgeR quasi-likelihood pipeline. *F1000Research* 5.

Cheng, H., G.T. Concepcion, X. Feng, H. Zhang, and H. Li, 2021 Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature methods* 18 (2):170-175.

Chiou, S.-H., C.-C. Yu, C.-Y. Huang, S.-C. Lin, C.-J. Liu *et al.*, 2008 Positive correlations of Oct-4 and Nanog in oral cancer stem-like cells and high-grade oral squamous cell carcinoma. *Clinical Cancer Research* 14 (13):4085-4095.

Clough, R.L., R. Sud, N. Davis-Silberman, R. Hertzano, K.B. Avraham *et al.*, 2004 Brn-3c (POU4F3) regulates BDNF and NT-3 promoter activity. *Biochemical and biophysical research communications* 324 (1):372-381.

Collas, P., and A.-M. Håkelien, 2003 Reprogramming somatic cells for therapeutic applications. *e-biomed: the journal of regenerative medicine* 4 (2):7-13.

Conesa, A., S. Götz, J.M. García-Gómez, J. Terol, M. Talón *et al.*, 2005 Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21 (18):3674-3676.

Dang, C.V., 1999 c-Myc target genes involved in cell growth, apoptosis, and metabolism. *Molecular and cellular biology* 19 (1):1-11.

Dang, C.V., L.M. Resar, E. Emison, S. Kim, Q. Li *et al.*, 1999 Function of the c-Myc oncogenic transcription factor. *Experimental cell research* 253 (1):63-77.

Davidson, N.M., A.D. Hawkins, and A. Oshlack, 2017 SuperTranscripts: a data driven reference for analysis and visualisation of transcriptomes. *Genome biology* 18 (1):148.

De Vito, D., S. Piraino, J. Schmich, J. Bouillon, and F. Boero, 2006 Evidence of reverse development in Leptomedusae (Cnidaria, Hydrozoa): the case of Laodicea undulata (Forbes and Goodsir 1851). *Marine Biology* 149 (2):339.

Derevyanko, A., K. Whittemore, R.P. Schneider, V. Jiménez, F. Bosch *et al.*, 2017 Gene therapy with the TRF 1 telomere gene rescues decreased TRF 1 levels with aging and prolongs mouse health span. *Aging cell* 16 (6):1353-1368.

Dobin, A., C.A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski *et al.*, 2013 STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* 29 (1):15-21.

Dobrovolskaia, M.A., A.E. Medvedev, K.E. Thomas, N. Cuesta, V. Toshchakov *et al.*, 2003 Induction of in vitro reprogramming by Toll-like receptor (TLR) 2 and TLR4 agonists in murine macrophages: effects of TLR "homotolerance" versus "heterotolerance" on NF-κB signaling pathway components. *The Journal of Immunology* 170 (1):508-519.

Drolet, D.W., K.M. Scully, D.M. Simmons, M. Wegner, K. Chu *et al.*, 1991 TEF, a transcription factor expressed specifically in the anterior pituitary during embryogenesis, defines a new class of leucine zipper proteins. *Genes & development* 5 (10):1739-1753.

Dubik, D., and R. Shiu, 1992 Mechanism of estrogen activation of c-myc oncogene expression. *Oncogene* 7 (8):1587-1594.

Dubrez, L., S. Causse, N.B. Bonan, B. Dumetier, and C. Garrido, 2019 Heat-shock proteins: chaperoning DNA repair. *Oncogene*:1-14.

DuBuc, T.Q., N. Traylor-Knowles, and M.Q. Martindale, 2014 Initiating a regenerative response; cellular and molecular features of wound healing in the cnidarian Nematostella vectensis. *BMC biology* 12 (1):24.

Edgar, R.C., 2004 MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 32 (5):1792-1797.

Efroni, I., A. Mello, T. Nawy, P.-L. Ip, R. Rahni *et al.*, 2016 Root regeneration triggers an embryo-like sequence guided by hormonal interactions. *Cell* 165 (7):1721-1733.

El-Karim, E.A., E.G. Hagos, A.M. Ghaleb, B. Yu, and V.W. Yang, 2013 Krüppel-like factor 4 regulates genetic stability in mouse embryonic fibroblasts. *Molecular cancer* 12 (1):89.

Flores, I., and M.A. Blasco, 2010 The role of telomeres and telomerase in stem cell aging. *FEBS letters* 584 (17):3826-3830.

Fuchs, B., W. Wang, S. Graspeuntner, Y. Li, S. Insua *et al.*, 2014 Regulation of polyp-to-jellyfish transition in Aurelia aurita. *Current Biology* 24 (3):263-273.

Fuller, C.W., L.R. Middendorf, S.A. Benner, G.M. Church, T. Harris *et al.*, 2009 The challenges of sequencing by synthesis. *Nature biotechnology* 27 (11):1013.

Garfield, D.A., and G.A. Wray, 2009 Comparative embryology without a microscope: using genomic approaches to understand the evolution of development. *Journal of Biology* 8 (7):65.

Gauchat, D., F. Mazet, C. Berney, M. Schummer, S. Kreger *et al.*, 2000 Evolution of Antp-class genes and differential expression of Hydra Hox/paraHox genes in anterior patterning. *Proceedings of the National Academy of Sciences* 97 (9):4493-4498.

Gearhart, J., E.E. Pashos, and M.K. Prasad, 2007 Pluripotency redux—advances in stem-cell research. *New England Journal of Medicine* 357 (15):1469-1472.

Gold, D.A., R.D. Gates, and D.K. Jacobs, 2014 The early expansion and evolutionary dynamics of POU class genes. *Molecular biology and evolution* 31 (12):3136-3147.

Gold, D.A., T. Katsuki, Y. Li, X. Yan, M. Regulski *et al.*, 2019 The genome of the jellyfish Aurelia and the evolution of animal complexity. *Nature ecology & evolution* 3 (1):96.

Götz, S., J.M. García-Gómez, J. Terol, T.D. Williams, S.H. Nagaraj *et al.*, 2008 High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic acids research* 36 (10):3420-3435.

Grabowska, W., E. Sikora, and A. Bielak-Zmijewska, 2017 Sirtuins, a promising target in slowing down the ageing process. *Biogerontology* 18 (4):447-476.

Greiss, S., and A. Gartner, 2009 Sirtuin/Sir2 phylogeny, evolutionary considerations and structural conservation. *Molecules and cells* 28 (5):407.

Guan, H., L. Xie, F. Leithäuser, L. Flossbach, P. Möller *et al.*, 2010 KLF4 is a tumor suppressor in B-cell non-Hodgkin lymphoma and in classic Hodgkin lymphoma. *Blood* 116 (9):1469-1478.

Guindon, S., and O. Gascuel, 2003 A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology* 52 (5):696-704.

Gurevich, A., V. Saveliev, N. Vyahhi, and G. Tesler, 2013 QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29 (8):1072-1075.

Haas, B.J., A.L. Delcher, S.M. Mount, J.R. Wortman, R.K. Smith Jr *et al.*, 2003 Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic acids research* 31 (19):5654-5666.

Haas, B.J., A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood *et al.*, 2013 De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature protocols* 8 (8):1494.

Harland, R.M., 2018 A new view of embryo development and regeneration. *science* 360 (6392):967-968.

Harrington, L., T. McPhail, V. Mar, W. Zhou, R. Oulton *et al.*, 1997 A mammalian telomerase-associated protein. *science* 275 (5302):973-977.

He, J., L. Zheng, W. Zhang, and Y. Lin, 2015 Life cycle reversal in Aurelia sp. 1 (Cnidaria, Scyphozoa). *PLoS One* 10 (12):e0145314.

Hertzano, R., M. Montcouquiol, S. Rashi-Elkeles, R. Elkon, R. Yücel *et al.*, 2004 Transcription profiling of inner ears from Pou4f3 ddl/ddl identifies Gfi1 as a target of the Pou4f3 deafness gene. *Human molecular genetics* 13 (18):2143-2153.

Higgins, D.G., A.J. Bleasby, and R. Fuchs, 1992 CLUSTAL V: improved software for multiple sequence alignment. *Bioinformatics* 8 (2):189-191.

Hoesel, B., and J.A. Schmid, 2013 The complexity of NF-κB signaling in inflammation and cancer. *Molecular cancer* 12 (1):86.

Holtfreter, J., 1946 Structure, motility and locomotion in isolated embryonic amphibian cells. *Journal of morphology* 79 (1):27-62.

Hou, Y., H. Wei, Y. Luo, and G. Liu, 2010 Modulating expression of brain heat shock proteins by estrogen in ovariectomized mice model of aging. *Experimental gerontology* 45 (5):323-330.

Hsieh, P.N., G. Zhou, Y. Yuan, R. Zhang, D.A. Prosdocimo *et al.*, 2017 A conserved KLF-autophagy pathway modulates nematode lifespan and mammalian age-associated vascular dysfunction. *Nature communications* 8 (1):1-12.

Hsu, A.-L., C.T. Murphy, and C. Kenyon, 2003 Regulation of aging and age-related disease by DAF-16 and heat-shock factor. *science* 300 (5622):1142-1145.

Jia, G., L. Su, S. Singhal, and X. Liu, 2012 Emerging roles of SIRT6 on telomere maintenance, DNA repair, metabolism and mammalian aging. *Molecular and cellular biochemistry* 364 (1-2):345-350.

Jopling, C., S. Boue, and J.C.I. Belmonte, 2011 Dedifferentiation, transdifferentiation and reprogramming: three routes to regeneration. *Nature reviews Molecular cell biology* 12 (2):79.

Kaeberlein, M., and R.W. Powers III, 2007 Sir2 and calorie restriction in yeast: a skeptical perspective. *Ageing research reviews* 6 (2):128-140.

Kaity, B., R. Sarkar, B. Chakrabarti, and M.K. Mitra, 2018 Reprogramming, oscillations and transdifferentiation in epigenetic landscapes. *Scientific reports* 8 (1):1-12.

Kent, W.J., 2002 BLAT—the BLAST-like alignment tool. *Genome research* 12 (4):656-664.

Kerr, J.F., C.M. Winterford, and B.V. Harmon, 1994 Apoptosis. Its significance in cancer and cancer therapy. *Cancer* 73 (8):2013-2026.

Kiernan, A.E., A.L. Pelling, K.K. Leung, A.S. Tang, D.M. Bell *et al.*, 2005 Sox2 is required for sensory organ development in the mammalian inner ear. *Nature* 434 (7036):1031.

Kim, D., J.M. Paggi, C. Park, C. Bennett, and S.L. Salzberg, 2019 Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature biotechnology* 37 (8):907-915.

Kim, J.B., B. Greber, M.J. Araúzo-Bravo, J. Meyer, K.I. Park *et al.*, 2009a Direct reprogramming of human neural stem cells by OCT4. *Nature* 461 (7264):649.

Kim, J.B., V. Sebastiano, G. Wu, M.J. Araúzo-Bravo, P. Sasse *et al.*, 2009b Oct4-induced pluripotency in adult neural stem cells. *Cell* 136 (3):411-419.

Kim, J.H., K. Kim, B.U. Youn, J. Lee, I. Kim *et al.*, 2014 Kruppel-like factor 4 attenuates osteoblast formation, function, and cross talk with osteoclasts. *J Cell Biol* 204 (6):1063-1074.

Kincaid, B., and E. Bossy-Wetzel, 2013 Forever young: SIRT3 a shield against mitochondrial meltdown, aging, and neurodegeneration. *Frontiers in aging neuroscience* 5:48.

Kitchen, S.A., C.M. Crowder, A.Z. Poole, V.M. Weis, and E. Meyer, 2015 De novo assembly and characterization of four anthozoan (phylum Cnidaria) transcriptomes. *G3: Genes, genomes, genetics* 5 (11):2441-2452.

Koç, A., and V.N. Gladyshev, 2007 Methionine sulfoxide reduction and the aging process.

Kolmogorov, M., J. Yuan, Y. Lin, and P.A. Pevzner, 2019 Assembly of long, error-prone reads using repeat graphs. *Nature biotechnology* 37 (5):540-546.

Kondoh, H., and R. Lovell-Badge, 2015 *Sox2: biology and role in development and disease*: Academic Press.

Koonin, E.V., L. Aravind, and A.S. Kondrashov, 2000 The impact of comparative genomics on our understanding of evolution. *Cell* 101 (6):573-576.

Koren, S., B.P. Walenz, K. Berlin, J.R. Miller, N.H. Bergman *et al.*, 2017 Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome research* 27 (5):722-736.

Kortschak, R.D., G. Samuel, R. Saint, and D.J. Miller, 2003 EST analysis of the cnidarian Acropora millepora reveals extensive gene loss and rapid sequence divergence in the model invertebrates. *Current Biology* 13 (24):2190-2195.

Kubota, S., 2005 Distinction of two morphotypes of Turritopsis nutricula medusae (Cnidaria, Hydrozoa, Anthomedusae) in Japan, with reference to their different abilities to revert to the hydroid stage and their distinct geographical distributions.

Kubota, S., 2011 Repeating rejuvenation in Turritopsis, an immortal hydrozoan (Cnidaria, Hydrozoa).

Kuhlbrodt, K., B. Herbarth, E. Sock, J. Enderich, I. Hermans-Borgmeyer *et al.*, 1998 Cooperative function of POU proteins and SOX proteins in glial cells. *Journal of Biological Chemistry* 273 (26):16050-16057.

Kuo, D.-H., 2019 Comparative Embryology as a Way to Understand Evolution, pp. 57-72 in *Old Questions and Young Approaches to Animal Evolution*. Springer.

Lefebvre, V., B. Dumitriu, A. Penzo-Méndez, Y. Han, and B. Pallavi, 2007 Control of cell fate and differentiation by Sry-related high-mobility-group box (Sox) transcription factors. *The international journal of biochemistry & cell biology* 39 (12):2195-2214.

Li, B., and C.N. Dewey, 2011 RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* 12 (1):323.

Li, H., 2018 Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* 34 (18):3094-3100.

Li, J.-y., D.-h. Guo, P.-c. Wu, and L.-s. He, 2018 Ontogeny reversal and phylogenetic analysis of Turritopsis sp. 5 (Cnidaria, Hydrozoa, Oceaniidae), a possible new species endemic to Xiamen, China. *PeerJ* 6:e4225.

Limbert, C., G. Päth, R. Ebert, V. Rothhammer, M. Kassem *et al.*, 2011 PDX1-and NGN3-mediated in vitro reprogramming of human bone marrow-derived mesenchymal stromal cells into pancreatic endocrine lineages. *Cytotherapy* 13 (7):802-813.

Liu, H., Y. Yang, Y. Ge, J. Liu, and Y. Zhao, 2019 TERC promotes cellular inflammatory response independent of telomerase. *Nucleic acids research* 47 (15):8084-8095.

Liu, K., B. Lin, M. Zhao, X. Yang, M. Chen *et al.*, 2013 The multiple roles for Sox2 in stem cell maintenance and tumorigenesis. *Cellular signalling* 25 (5):1264-1271.

Lorian, V., 1989 In vitro simulation of in vivo conditions: physical state of the culture medium. *Journal of clinical microbiology* 27 (11):2403.

Löw, P., 2011 The role of ubiquitin–proteasome system in ageing. *General and comparative endocrinology* 172 (1):39-43.

Lundberg, A.S., W.C. Hahn, P. Gupta, and R.A. Weinberg, 2000 Genes involved in senescence and immortalization. *Current opinion in cell biology* 12 (6):705-709.

Lynch, M., and J.S. Conery, 2003 The origins of genome complexity. *science* 302 (5649):1401-1404.

Lynch, M., and B. Walsh, 2007 *The origins of genome architecture*: Sinauer Associates Sunderland, MA.

Ma, Y., X. Zhang, H. Ma, Y. Ren, Y. Sun *et al.*, 2014 Bioinformatic analysis of the four transcription factors used to induce pluripotent stem cells. *Cytotechnology* 66 (6):967-978.

Martinet, C., P. Monnier, Y. Louault, M. Benard, A. Gabory *et al.*, 2016 H19 controls reactivation of the imprinted gene network during muscle regeneration. *Development* 143 (6):962-971.

Martínez, D.E., and D. Bridge, 2012 Hydra, the everlasting embryo, confronts aging. *International Journal of Developmental Biology* 56 (6-7-8):479-487.

Matsumoto, Y., and M.P. Miglietta, 2021 Cellular reprogramming and immortality: Expression profiling reveals putative genes involved in Turritopsis dohrnii's life cycle reversal. *Genome biology and evolution* 13 (7):evab136.

Matsumoto, Y., and M.P. Miglietta, In review Cellular reprogramming and immortality: Expression profiling reveals putative genes involved in Turritopsis dohrnii's life cycle reversal.

Matsumoto, Y., S. Piraino, and M.P. Miglietta, 2019 Transcriptome characterization of reverse development in Turritopsis dohrnii (Hydrozoa, Cnidaria). *G3: Genes, genomes, genetics* 9 (12):4127-4138.

Matsunaga, A., M. Tsugawa, and J. Fortes, 2008 Cloudblast: Combining mapreduce and virtualization on distributed resources for bioinformatics applications, pp. 222-229 in *eScience, 2008. eScience'08. IEEE Fourth International Conference on*. IEEE.

Merrell, A.J., and B.Z. Stanger, 2016 Adult cell plasticity in vivo: de-differentiation and transdifferentiation are back in style. *Nature reviews Molecular cell biology* 17 (7):413.

Miglietta, M.P., and H.A. Lessios, 2009 A silent invasion. *Biological Invasions* 11 (4):825-834.

Miglietta, M.P., D. Maggioni, and Y. Matsumoto, 2018 Phylogenetics and species delimitation of two hydrozoa (phylum Cnidaria): Turritopsis (McCrady, 1857) and Pennaria (Goldfuss, 1820). *Marine Biodiversity*:1-16.

Millane, R.C., J. Kanska, D.J. Duffy, C. Seoighe, S. Cunningham *et al.*, 2011 Induced stem cell neoplasia in a cnidarian by ectopic expression of a POU domain transcription factor. *Development* 138 (12):2429-2439.

Miller, D.J., D.C. Hayward, J.S. Reece-Hoyes, I. Scholten, J. Catmull *et al.*, 2000 Pax gene diversity in the basal cnidarian Acropora millepora (Cnidaria, Anthozoa): implications for the evolution of the Pax gene family. *Proceedings of the National Academy of Sciences* 97 (9):4475-4480.

Miller, D.M., S.D. Thomas, A. Islam, D. Muench, and K. Sedoris, 2012 c-Myc and cancer metabolism, pp. 5546-5553. AACR.

Monk, M., and C. Holding, 2001 Human embryonic genes re-expressed in cancer cells. *Oncogene* 20 (56):8085.

Moskovitz, J., S. Bar-Noy, W.M. Williams, J. Requena, B.S. Berlett *et al.*, 2001 Methionine sulfoxide reductase (MsrA) is a regulator of antioxidant defense and lifespan in mammals. *Proceedings of the National Academy of Sciences* 98 (23):12920-12925.

Motoyama, H., S. Ogawa, A. Kubo, S. Miwa, J. Nakayama *et al.*, 2009 In vitro reprogramming of adult hepatocytes into insulin-producing cells without viral vectors. *Biochemical and biophysical research communications* 385 (1):123-128.

Murshid, A., T. Eguchi, and S.K. Calderwood, 2013 Stress proteins in aging and life span. *International Journal of Hyperthermia* 29 (5):442-447.

Murthy, M., and J.L. Ram, 2015 Invertebrates as model organisms for research on aging biology. Taylor & Francis.

Neves, S.R., P.T. Ram, and R. Iyengar, 2002 G protein pathways. *science* 296 (5573):1636-1639.

Nosenko, T., F. Schreiber, M. Adamska, M. Adamski, M. Eitel *et al.*, 2013 Deep metazoan phylogeny: when different genes tell different stories. *Molecular phylogenetics and evolution* 67 (1):223-233.

Nowak, T., D. Januszkiewicz, M. Zawada, M. Pernak, K. Lewandowski *et al.*, 2006 Amplification of hTERT and hTERC genes in leukemic cells with high expression and activity of telomerase. *Oncology reports* 16 (2):301-305.

Nueda, M.J., S. Tarazona, and A. Conesa, 2014 Next maSigPro: updating maSigPro bioconductor package for RNA-seq time series. *Bioinformatics* 30 (18):2598-2602.

Nurk, S., B.P. Walenz, A. Rhie, M.R. Vollger, G.A. Logsdon *et al.*, 2020 HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome research* 30 (9):1291-1305.

Okada, T., 1991 *Transdifferentiation: flexibility in cell differentiation*: Oxford University Press, USA.

Oviedo, N.J., and W.S. Beane, 2009 Regeneration: The origin of cancer or a possible cure?, pp. 557-564 in *Seminars in cell & developmental biology*. Elsevier.

Pan, G.J., Z.Y. Chang, H.R. Schöler, and P. Duanqing, 2002 Stem cell pluripotency and transcription factor Oct4. *Cell research* 12 (5):321.

Patruno, M., M.C. Thorndyke, M.C. Carnevali, F. Bonasoro, and P.W. Beesley, 2001 Growth factors, heat-shock proteins and regeneration in echinoderms. *Journal of Experimental Biology* 204 (5):843-848.

Pesce, M., and H.R. Schöler, 2001 Oct-4: gatekeeper in the beginnings of mammalian development. *Stem cells* 19 (4):271-278.

Petrou, G., and T. Crouzier, 2018 Mucins as multifunctional building blocks of biomaterials. *Biomaterials science* 6 (9):2282-2297.

Piraino, S., F. Boero, B. Aeschbach, and V. Schmid, 1996 Reversing the life cycle: medusae transforming into polyps and cell transdifferentiation in Turritopsis nutricula (Cnidaria, Hydrozoa). *The Biological Bulletin* 190 (3):302-312.

Piraino, S., D. De Vito, J. Schmich, J. Bouillon, and F. Boero, 2004 Reverse development in Cnidaria. *Canadian Journal of Zoology* 82 (11):1748-1754.

Pop, M., 2009 Genome assembly reborn: recent computational challenges. *Briefings in bioinformatics* 10 (4):354-366.

Pratt, W.B., Y. Morishima, H.-M. Peng, and Y. Osawa, 2010 Proposal for a role of the Hsp90/Hsp70-based chaperone machinery in making triage decisions when proteins undergo oxidative and toxic damage. *Experimental biology and medicine* 235 (3):278-289.

Prouty, N., E. Roark, N. Buster, and S.W. Ross, 2011 Growth rate and age distribution of deep-sea black corals in the Gulf of Mexico. *Marine Ecology Progress Series* 423:101-115.

Putnam, N.H., M. Srivastava, U. Hellsten, B. Dirks, J. Chapman *et al.*, 2007 Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *science* 317 (5834):86-94.

Qin, M., S. Wu, A. Li, F. Zhao, H. Feng *et al.*, 2019 LRScaf: improving draft genomes using long noisy reads. *BMC genomics* 20 (1):1-12.

Reiter, S., M. Crescenzi, B. Galliot, and W.C. Buzgariu, 2012 Hydra, a versatile model to study the homeostatic and developmental functions of cell death. *International Journal of Developmental Biology* 56 (6-7-8):593-604.

Richards, R.J., 2009 *The meaning of evolution: The morphological construction and ideological reconstruction of Darwin's theory*: University of Chicago Press.

Ridley, A.J., 2006 Rho GTPases and actin dynamics in membrane protrusions and vesicle trafficking. *Trends in cell biology* 16 (10):522-529.

Rizzino, A., 2013 Concise review: The Sox2-Oct4 connection: Critical players in a much larger interdependent network integrated at multiple levels. *Stem cells* 31 (6):1033-1039.

Robinson, M.D., D.J. McCarthy, and G.K. Smyth, 2010 edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26 (1):139-140.

Rogina, B., and S.L. Helfand, 2004 Sir2 mediates longevity in the fly through a pathway related to calorie restriction. *Proceedings of the National Academy of Sciences* 101 (45):15998-16003.

Ronquist, F., M. Teslenko, P. Van Der Mark, D.L. Ayres, A. Darling *et al.*, 2012 MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology* 61 (3):539-542.

Rowland, B.D., R. Bernards, and D.S. Peeper, 2005 The KLF4 tumour suppressor is a transcriptional repressor of p53 that acts as a context-dependent oncogene. *Nature cell biology* 7 (11):1074.

Rowland, B.D., and D.S. Peeper, 2006 KLF4, p21 and context-dependent opposing forces in cancer. *Nature Reviews Cancer* 6 (1):11.

Rubin, G.M., M.D. Yandell, J.R. Wortman, G.L. Gabor, C.R. Nelson *et al.*, 2000 Comparative genomics of the eukaryotes. *science* 287 (5461):2204-2215.

Saigusa, S., K. Tanaka, Y. Toiyama, T. Yokoe, Y. Okugawa *et al.*, 2009 Correlation of CD133, OCT4, and SOX2 in rectal cancer and their association with distant recurrence after chemoradiotherapy. *Annals of surgical oncology* 16 (12):3488-3498.

Sanders, S.M., and P. Cartwright, 2015a Interspecific differential expression analysis of RNA-Seq data yields insight into life cycle variation in hydractiniid hydrozoans. *Genome biology and evolution* 7 (8):2417-2431.

Sanders, S.M., and P. Cartwright, 2015b Patterns of Wnt signaling in the life cycle of Podocoryna carnea and its implications for medusae evolution in Hydrozoa (Cnidaria). *Evolution & development* 17 (6):325-336.

Sarid, J., T.D. Halazonetis, W. Murphy, and P. Leder, 1987 Evolutionarily conserved regions of the human c-myc protein can be uncoupled from transforming activity. *Proceedings of the National Academy of Sciences* 84 (1):170-173.

Schmich, J., Y. Kraus, D. De Vito, D. Graziussi, F. Boero *et al.*, 2007 Induction of reverse development in two marine Hydrozoans. *International Journal of Developmental Biology* 51 (1):45-56.

Schmid, V., A. Bally, K. Beck, M. Haller, W. Schlage *et al.*, 1991 The extracellular matrix (mesoglea) of hydrozoan jellyfish and its ability to support cell adhesion and spreading, pp. 3-10 in *Hydrobiologia*. Springer.

Segre, J.A., C. Bauer, and E. Fuchs, 1999 Klf4 is a transcription factor required for establishing the barrier function of the skin. *Nature genetics* 22 (4):356.

Shcherbata, H.R., A.S. Yatsenko, L. Patterson, V.D. Sood, U. Nudel *et al.*, 2007 Dissecting muscle and neuronal disorders in a Drosophila model of muscular dystrophy. *The EMBO journal* 26 (2):481-493.

Shenoy, A., and R. Blelloch, 2012 microRNA induced transdifferentiation. *F1000 biology reports* 4.

Shi, G., and Y. Jin, 2010 Role of Oct4 in maintaining and regaining stem cell pluripotency. *Stem cell research & therapy* 1 (5):39.

Shi, Q., Z. Dong, and H. Wei, 2007 The involvement of heat shock proteins in murine liver regeneration. *Cell Mol Immunol* 4 (1):53.

Simão, F.A., R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, and E.M. Zdobnov, 2015 BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31 (19):3210-3212.

Slamon, D.J., and M.J. Cline, 1984 Expression of cellular oncogenes during embryonic and fetal development of the mouse. *Proceedings of the National Academy of Sciences* 81 (22):7141-7145.

Snow, B.E., N. Erdmann, J. Cruickshank, H. Goldman, R.M. Gill *et al.*, 2003 Functional conservation of the telomerase protein Est1p in humans. *Current Biology* 13 (8):698-704.

Spring, J., N. Yanze, A.M. Middel, M. Stierwald, H. Gröger *et al.*, 2000 The mesoderm specification factor twist in the life cycle of jellyfish. *Developmental biology* 228 (2):363-375.

Steele, R.E., C.N. David, and U. Technau, 2011 A genomic view of 500 million years of cnidarian evolution. *Trends in Genetics* 27 (1):7-13.

Sullivan, J.C., and J.R. Finnerty, 2007 A surprising abundance of human disease genes in a simple "basal" animal, the starlet sea anemone (Nematostella vectensis). *Genome* 50 (7):689-692.

Tai, M.-H., C.-C. Chang, L.K. Olson, and J.E. Trosko, 2005 Oct4 expression in adult human stem cells: evidence in support of the stem cell theory of carcinogenesis. *Carcinogenesis* 26 (2):495-502.

Taira, T., J. Maëda, T. Onishi, H. Kitaura, S. Yoshida *et al.*, 1998 AMY-1, a novel C-MYC binding protein that stimulates transcription activity of C-MYC. *Genes to Cells* 3 (8):549-565.

Takahashi, K., K. Tanabe, M. Ohnuki, M. Narita, T. Ichisaka *et al.*, 2007 Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 131 (5):861-872.

Takahashi, K., and S. Yamanaka, 2006 Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126 (4):663-676.

Tardent, P., 1963 Regeneration in the Hydrozoa. *Biological Reviews* 38 (3):293-333.

Tian, X., D. Firsanov, Z. Zhang, Y. Cheng, L. Luo *et al.*, 2019 SIRT6 is responsible for more efficient DNA double-strand break repair in long-lived species. *Cell* 177 (3):622-638. e622.

Tissenbaum, H.A., and L. Guarente, 2001 Increased dosage of a sir-2 gene extends lifespan in Caenorhabditis elegans. *Nature* 410 (6825):227-230.

Tomczyk, S., K. Fischer, S. Austad, and B. Galliot, 2015 Hydra, a powerful model for aging studies. *Invertebrate reproduction & development* 59 (sup1):11-16.

Tower, J., 2011 Heat shock proteins and Drosophila aging. *Experimental gerontology* 46 (5):355-362.

Tseng, A.-S., and M. Levin, 2008 Tail regeneration in Xenopus laevis as a model for understanding tissue repair. *Journal of dental research* 87 (9):806-816.

Turpin, F., B. Potier, J. Dulong, P.-M. Sinet, J. Alliot *et al.*, 2011 Reduced serine racemase expression contributes to age-related deficits in hippocampal cognitive function. *Neurobiology of aging* 32 (8):1495-1504.

Uringa, E.-J., J.L. Youds, K. Lisaingo, P.M. Lansdorp, and S.J. Boulton, 2011 RTEL1: an essential helicase for telomere maintenance and the regulation of homologous recombination. *Nucleic acids research* 39 (5):1647-1655.

Uzawa, J., M. Urai, T. Baba, H. Seki, K. Taniguchi *et al.*, 2009 NMR study on a novel mucin from jellyfish in natural abundance, qniumucin from Aurelia aurita. *Journal of natural products* 72 (5):818-823.

Vaser, R., I. Sović, N. Nagarajan, and M. Šikić, 2017 Fast and accurate de novo genome assembly from long uncorrected reads. *Genome research* 27 (5):737-746.

Vaupel, J.W., A. Baudisch, M. Dölling, D.A. Roach, and J. Gampe, 2004 The case for negative senescence. *Theoretical population biology* 65 (4):339-351.

Vaziri, H., and S. Benchimol, 1998 Reconstitution of telomerase activity in normal human cells leads to elongation of telomeres and extended replicative life span. *Current Biology* 8 (5):279-282.

Venteicher, A.S., and S.E. Artandi, 2009 TCAB1: driving telomerase to Cajal bodies. *Cell Cycle* 8 (9):1329-1331.

Vinarsky, V., D.L. Atkinson, T.J. Stevenson, M.T. Keating, and S.J. Odelberg, 2005 Normal newt limb regeneration requires matrix metalloproteinase function. *Developmental biology* 279 (1):86-98.

Wang, L., A. Zheng, L. Yi, C. Xu, M. Ding *et al.*, 2004 Identification of potential nuclear reprogramming and differentiation factors by a novel selection method for cloning chromatin-binding proteins. *Biochemical and biophysical research communications* 325 (1):302-307.

Watanabe, H., R. Mättner, and T.W. Holstein, 2009 Immortality and the base of multicellular life: Lessons from cnidarian stem cells, pp. 1114-1125 in *Seminars in cell & developmental biology*. Elsevier.

Weismann, A., 1883 *Die entstehung der sexualzellen bei den hydromedusen: zugleich ein Betrag zur Kenntniss des Baues und der Lebenserscheinungen dieser Gruppe*: Fischer.

Weissbach, H., F. Etienne, T. Hoshi, S.H. Heinemann, W.T. Lowther *et al.*, 2002 Peptide methionine sulfoxide reductase: structure, mechanism of action, and biological function. *Archives of Biochemistry and Biophysics* 397 (2):172-178.

Wenger, Y., and B. Galliot, 2013 RNAseq versus genome-predicted transcriptomes: a large population of novel transcripts identified in an Illumina-454 Hydra transcriptome. *BMC genomics* 14 (1):204.

Whitaker, N.J., T.M. Bryan, P. Bonnefin, A. Chang, E.A. Musgrove *et al.*, 1995 Involvement of RB-1, p53, p16INK4 and telomerase in immortalisation of human cells. *Oncogene* 11 (5):971-976.

Williams, C.R., A. Baccarella, J.Z. Parrish, and C.C. Kim, 2016 Trimming of sequence reads alters RNA-Seq gene expression estimates. *BMC bioinformatics* 17 (1):103.

Wood, D.E., J. Lu, and B. Langmead, 2019 Improved metagenomic analysis with Kraken 2. *Genome biology* 20 (1):1-13.

Wood, D.E., and S.L. Salzberg, 2014 Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology* 15 (3):R46.

Xu, M., L. Guo, S. Gu, O. Wang, R. Zhang *et al.*, 2020 TGS-GapCloser: a fast and accurate gap closer for large genomes with low coverage of error-prone long reads. *GigaScience* 9 (9):giaa094.

Xue, W., J.-T. Li, Y.-P. Zhu, G.-Y. Hou, X.-F. Kong *et al.*, 2013 L_RNA_scaffolder: scaffolding genomes with transcripts. *BMC genomics* 14 (1):604.

Yan, J., and T. Jin, 2012 Signaling network from GPCR to the actin cytoskeleton during chemotaxis. *Bioarchitecture* 2 (1):15-18.

Yang, E.V., and S.V. Byant, 1994 Developmental regulation of a matrix metalloproteinase during regeneration of axolotl appendages. *Developmental biology* 166 (2):696-703.

Yu, J., M.A. Vodyanik, K. Smuga-Otto, J. Antosiewicz-Bourget, J.L. Frane *et al.*, 2007 Induced pluripotent stem cell lines derived from human somatic cells. *science* 318 (5858):1917-1920.

Zacharias, H., B. Anokhin, K. Khalturin, and T.C. Bosch, 2004 Genome sizes and chromosomes in the basal metazoan Hydra. *Zoology* 107 (3):219-227.

Zaehres, H., M.W. Lensch, L. Daheron, S.A. Stewart, J. Itskovitz-Eldor *et al.*, 2005 High-efficiency RNA interference in human embryonic stem cells. *Stem cells* 23 (3):299-305.

Zaghlool, A., J. Halvardson, J.J. Zhao, M. Etemadikhah, A. Kalushkova *et al.*, 2016 A role for the chromatin-remodeling factor BAZ1A in neurodevelopment. *Human mutation* 37 (9):964-975.

Zdobnov, E.M., and R. Apweiler, 2001 InterProScan–an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17 (9):847-848.

Zhang, G., C. Li, Q. Li, B. Li, D.M. Larkin *et al.*, 2014 Comparative genomics reveals insights into avian genome evolution and adaptation. *science* 346 (6215):1311-1320.

Zhang, J., G. Li, L. Feng, H. Lu, and X. Wang, 2020 Krüppel-like factors in breast cancer: Function, regulation and clinical relevance. *Biomedicine & Pharmacotherapy* 123:109778.

Zhou, X.Z., and K.P. Lu, 2001 The Pin2/TRF1-interacting protein PinX1 is a potent telomerase inhibitor. *Cell* 107 (3):347-359.

Zhu, B.-H., J. Xiao, W. Xue, G.-C. Xu, M.-Y. Sun *et al.*, 2018 P_RNA_scaffolder: a fast and accurate genome scaffolder using paired-end RNA-sequencing reads. *BMC genomics* 19 (1):175.

Zietara, M.S., A. Arndt, A. Geets, B. Hellemans, and F.A. Volckaert, 2000 The nuclear rDNA region of Gyrodactylus arcuatus and G. branchicus (Monogenea: Gyrodactylidae). *Journal of Parasitology* 86 (6):1368-1373.