MAKING HEADS AND TAILS OF *MOLGULA*: NEXT GENERATION SEQUENCING ANALYSIS OF CLOSELY RELATED TAILED AND TAIL-LESS ASCIDIAN SPECIES

By

Elijah Kariem Lowe

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computer Science - Doctor of Philosophy

2015

ABSTRACT

MAKING HEADS AND TAILS OF *MOLGULA*: NEXT GENERATION SEQUENCING ANALYSIS OF CLOSELY RELATED TAILED AND TAIL-LESS ASCIDIAN SPECIES

By

Elijah Kariem Lowe

Tunicates are invertebrate chordates and are the sister group to the vertebrates. Although tunicates have little morphological resemblance to vertebrates in their adult stage, they do share several features in their larval stage: a hollow dorsal neural tube, gill slits, and a post-anal tail, containing a notochord—a group of cells organized in a rod shaped structure - the key features that classify the phyla. Within the tunicates, several ascidians have undergone tail-loss, and many of them are Molgulidae. Hybrids have been produced through the cross fertilization of two *Molgula* species (*Molgula occulta* and *Molgula oculata*), and no other solitary *Molgula* species have been known to hybridize. Here we have sequenced the transcriptomes of several developmental stages of both *M. occulta* and *M. oculata*—two closely related, free-spawning ascidian species, and their hybrid, in order to study the mechanisms behind tail loss in *M. occulta.* We were first presented with the problem of identifying the best approach for the *de novo* assembly of our transcriptomes. Here we determined that processing reads through digital normalization, a redundancy reduction step, had less of an effect on assemblies than did the assembler used. We then sequenced and assembled the genomes of *M. occulta*, *M. oculata* and a more distant species, *M. occidentalis.* This allowed us to characterize the genomes, discovering that the species are more divergent then they appear phenotypically, and also supporting better gene models. Through differential expression analysis we determined that *M. oculata* and the hybrid appear to express many shared

transcripts that are up-regulated during the formation of the ascidian tail, and that these genes are primarily upregulated in the tailed species and hybrid in relation to the tail-less species. Furthermore, of those transcripts upregulated at the tailbud stage in the hybrid but not in *M. occulta*, it appears that expression is being restored by the *M. oculata* allele. This suggests that the relative lack of differential gene expression in the neurula-to-tailbud transition in *M. occulta* is due to loss-of-function of cis-regulatory elements controlling the expression of key genes involved in tail and CNS formation.

ACKNOWLEDGMENTS

Where do I begin? I am a summation of all the people that I have come in contact with throughout my life. So, I would like to dedicate this text to some of the most influential people.

I start with my mother Patricia Lowe who always encouraged me to pursue my wildest dreams. My father William Robertson, who taught me to question the world and to always pay attention to the small things. My brother Fard Lowe, who consistently served as an inspiration to me and was someone I could always count on. To my high school counselor, Kathy Giles-Harris and high school math teacher, Christopher Reese who both saw something special in me, and pushed me to pursue it.

My brothers from Morehouse—Terron Ferguson, Ryan Shepard, John-Marcus Philips, Sean Brazier and Kelechi Kalu—who kept the laughs and good times rolling along my journey. You guys always made me feel cool, even though I'm a science nerd.

My MSU family, who are too many to name, but I'll name a few—my brother and sister, Cameron Khalfani Herman and Neem Serra, my best bud Ruby Carrillo, Dr. Barbara Thelamour, my mentor Judi Brown-Clarke, my problem solvers Connie James and Darcie Zubek, Carmel Martin-Fairey, Daniel Couvrtier, Luis Zaman, Chad Byers, Temple Smith, James Kremer, okay I said a few, so I'll stop there. I needed you all along the way, through late nights of studying, endless laughs, awesome food, much needed pep talks and just being there to help me survive Michigan.

Collaborators and friends, Billie Swalla, her lab, Lionel Christiaen, his lab, Alberto Stolfi and the wonderful Princess Claudia Racioppi, who all helped me to develop as biologist, experimentally and conceptually, and who advised me along the way. I look forward to

working with you all for years to come. And a special thanks to Billie without whom this project would not exist. We have traveled the world together (well at least a good part of the US and Europe), and done what I think is some good science together.

My advisor, Titus Brown. I cannot thank you enough for guiding me along the way. I truly lucked out when I landed in your lab. You are a great scientist, adviser, mentor, and person, and this is the only time you'll hear (or more like read) me say this.

My awesome committee—David Arnosti, Jin Chen, Ian Dworkin, and Yanni Sun—who were always tough, but fair.

And to those who pursue knowledge not for the sake of notability but just to make the world a better place.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

The origin of body plans is an age-old question. Chordates are distinguished by their body plan: most chordates form tadpole-shaped larvae that commonly have a hollow dorsal neural tube and a post-anal tail containing a central notochord flanked by bilateral muscle [86]. During embryogenesis, the notochord serves as the source of a patterning signal for neural tube and paraxial mesoderm, an addition to the axial skeleton for the larval tail [42, 111]. Tunicates are one of the three subphyla of chordates and are so grouped because of their outer covering known as a tunic, and during development form the typical tadpole larval body plan [34]. Ascidians and larvaceans are two of the groups within the Tunicates, and both develop typical chordate body plans. The ascidians exhibit their tailed body form during their larvae free-swimming stage, using the tail for locomotion before becoming sessile, undergoing metamorphosis and filter feeding for the remainder of their lives. During the free-swimming larval stage, the elongation and mobility of the tail is dependent upon the proper formation of the notochord and muscle cells [98]. In ascidians and in lower vertebrates the improper formation of the notochord leads to severely shortened larva that cannot swim or feed properly [16, 43, 111]. Out of ~3000 ascidian species, 16 are known to have independently lost their larval tail, differentiated notochord, muscle cell, and other chordate features, changing their developmental body plan. The majority of these losses have happened within the *Molgula* clade [2, 120], although the Styelidae also have two identified tail-less species [34].

Ascidians are a simple system in which to study developmental processes: their cell lineages have been traced starting at fertilization [84] to gastrulation [85, 82], they have invariant early cell lineages and a small number of cells [59], and there has been no documentation of ascidians developing without an invariant cell lineage. Ascidian development is nearly identical across distantly related solitary ascidians [60, 86]. They also have rapid embryogenesis, compact genomes, and simplified larval body plans containing few larval tissue types [12, 40, 14]. In addition to the several Molgulids that have independently lost their tail, two Molgulids, one tailed and one tail-less species, can be hybridized, offering the opportunity to study the mechanism behind evolutionarily divergent body plans [41].

Many genes in the notochord gene network have been identified by subtractive hybridization screening and microarrays [39, 30, 22, 50]. More recently, sequencing technologies such as Ion Torrent, Roche 454 and Illumina have made genome or transcriptome wide analysis more readily available for non-model species. These technologies have several advantages over microarrays: they have a wider scope, are more precise and are able to find novel genes [71]. With the advances in technology we have now sequenced the transcriptomes of both species and their hybrid. This ability to sequence the entire transcriptome at a high resolution allows us to look at pivotal time points in tail development and compare across closely related species to identify and study mechanisms that have been lost or modified during evolution.

We present a comparative study of the tailed *M. oculata* and the tail-less *M. occulta* through gene expression in order to understand the underlying factors behind tail development, tail loss and the origin of the chordate body plan. Although this study presents the first assembled *Molgula* genomes, there are a number of sequenced tunicate genomes available: in particular, we use the assembled and annotated genome of *Ciona intestinalis*, which serves as the most documented and closest complete reference for the *Molgula* and other ascidian species [14, 98, 100]. We began this project with RNA-seq data from several time points from each of the species (*M. occulta* and *M. oculata*) and their hybrid. Whole genome and transcriptome sequencing will not definitively identify the factors involved in the development of the tailed chordate body plan since other processes are occurring, but examining the correlation of gene expression patterns between the tailed, tail-less and interspecific *Molgula* hybrid allows us to filter genes associated with this process. Before we can make biological inferences from our data we have to produce a quality assembly, and because of this we first assessed the quality of an efficient low-memory assembly pipeline for our RNA-seq data. We later obtained genomic DNA and assembled the genomes of *M. occulta*, *M. oculata* and a more divergent species *M. occidentalis.* This allowed us to analyze the homology between ascidian gene networks, and build more complete transcript modules for differential expression [133].

Chapter 2

Literature Review

2.1 Ascidian tail development

The notochord is one of the distinguishing characteristics of chordates: it is responsible for the extension of the larval tail which is typical for chordate body plan. In their adult forms, ascidians and their vertebrate cousins bear little resemblance to one another, however during development they have similar body plans that include the notochord [40]. Ascidians are known for their bilateral and invariant cell cleavage, and their development is well described up to the gastrulation stage [84, 85, 82]. Like vertebrates such as Xenopus, ascidians depend on maternally localized determinants to regulate cell moments and division, while the development of the early body plans are similar, the location and identity of these determinants are different in the ascidian and in vertebrates [60]. Solitary ascidian notochords typically originate from two cell lineages, with the primary notochord deriving from the "A" blastomere and the secondary notochord deriving from the "B" blastomere [84]. Both "A" and "B" blastomeres can be identified at the 4-cell embryonic stage. At this stage the blastomeres are labeled according to the Conklin convention: "a" and "A" label the anterior animal and vegetal blastomeres, while "b" and "B" label the posterior animal and vegetal blastomeres [11] . Although the notochords cells have been traced back to the 4-cell stage, notochord induction does not occur until the 32-cell stage. By the 64-cell stage there are 10 notochord cell precursors, the 8 primary precursor notochord cells—A lineage—which are no longer multipotent, and the 2 secondary notochord cells which are not restricted until the 110-cell stage [85, 139, 140, 58]. Two additional stages of cell division occur, one at gastrulation and one at neurulation, ending with the 40 notochord cells that are typical of most solitary ascidian tadpole larvae [11]. At the onset of neurulation the notochord begins to form, starting with the closing of the neural tube and posterior movement of the notochord and muscle cells, followed by the mediolateral convergence of the notochord cells to the midline and then the polarization and intercalation of the cells through a process known as convergence and extension [117]. At this point the larval tail is constructed of a notochord flanked by 3 rows of muscles on each side, and both notochord and muscle cell derive from the same blastomeres [85]. While the arrangement of the notochord cells is a stochastic process, the anterior 32-cells—primary notochord cells—are always formed by the A7.3 and A7.7 blastomere and the posterior most 8—secondary—notochord cells are always formed by the B8.6 blastomere; however, the ordering of the 32 most anterior is not determinate, as cells from both the A7.3 and A7.7 intercalate in a random order (Figure 2.2)[84, 85, 78, 117, 51]. This process, along with muscle cell development, drives the formation of the larval tail [78, 39, 117].

The ascidian tail is used for dispersal during the free-swimming larval stage, where the ascidian locates a substrate to attach. After becoming sessile, the ascidian tail is absorbed into the trunk region as the larvae metamorphose into the adult, the form in which they filter feed for the duration of their lives. Although a tailed larvae is typical of most ascidians, several species within the Stolidobranchia order have individually undergone tail-loss, and many of these species fall in the family Molgulidae [2, 42, 34, 69]. Species without tails tend to have lower speciation rates and smaller geographical ranges [68]. The tail-less—anural—species develop in a similar manner and are indistinguishable from their

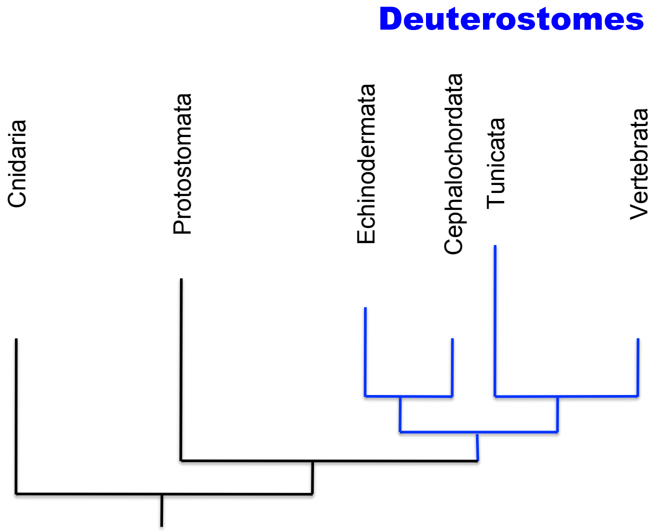


Figure 2.1: **Phylographic sketch.** Phylogenetic placement of Cnidaria and the Bilateria (Protostomata and Deuterostoma). Deuterostomes, which in Greek means second mouth, are distinguished in the Bilateria by forming their anus first. Tunicata and Cephalachordata are the sister groups to Vertebrata. This subphyla shares several characteristics; a notochord, a hollow dorsal nerve cord, and a post-anal tail at some point in their life cycles. (Blue) branches are Deuterostomes.

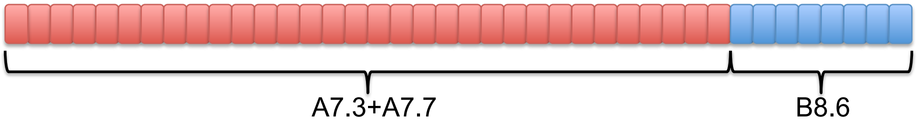


Figure 2.2: **Notochord cells.**The primary notochord cells (red), also known as the A-lineage, are specified at the 64-cell stage. There are a total of 32 primary notochord cells that come from the A7.3 and A7.7 blastomere, and the intercalation of the cells happen in a stochastic manner. The secondary notochord cells (blue) derive from the B8.6 blastomere and are specified at the 110-cell stage, one cell division after the primary notochord cells.

tailed—urodele—counterparts up to late gastrulation [2, 120, 39]. When studying other direct and indirect developers, such as the sea urchins *Heliocidaris erythrogramma* and *Heliocidaris tuberculata*, the body plans diverge at an earlier developmental stage. *H. erythrogramma*, a direct developer, forgoes the typical pluteus larval body plan and in doing so changes its developmental plan [27]. In *H. erythrogramma* and other direct developers, blastomeres are equivalently sized, cleavage orientation remains parallel and cell division is synchronous. In contrast, after the 4 cell division an unequal three-tiered cleavage is established in *H. tuberculata* and other indirect developers along with the loss of synchronous cell division. Cell fates also differs between direct and indirect sea urchin species [138]. The change in body plan in *Molgula* is a much more recent evolutionary occurrence in comparison to *Heliocidaris*, and the comparison gives insight into early onsite of changing body plans, and may demonstrate alternative methods to the changing of early cell fate. Additionally studying the *Molgula* adds to the understanding of the ancestral chordate body plan.

Anural ascidians lack several urodele features including an intercalated and extended notochord, differentiated muscle cells and the otolith sensory organ. The absence of differentiated muscles cells and intercalated notochord are the likely cause of tail-lessness in these species [78, 120]. The development of several tail-less species has been studied in some detail. *Molgula tectiformis* notochord cells do not divide again after the 10 precursor cells are formed and *M. occulta* stops dividing after 20 cells [42]. The same occurs in *M. bleizi*, however after the 20 notochord cells are formed, the embryo attempts to make a tail but never completes the process [123]. It has also been shown that chordate embryos without fully developed notochord and/or muscle cells do not fully elongate or fail completely to develop a tail [42, 125, 111]. Vertebrates, cephalochordates and most ascidians have tailed larvae and interspecific *Molgula occulta/oculata* hybrids can restore the urodele features, evidence that the ancestral chordate had a tailed larvae and that the mechanism for tail development was present in anural ascidians but was lost over evolutionary time [2, 39].

In order to study specific mechanisms of tail loss, we can study closely related anural and urodele species. One such pair of species, *M. oculata* and *M. occulta*, both of the Roscovita clade, have been shown to produce hybrids in lab conditions. Of the known *Molgula* species *M. occulta* and *M. oculata* are the only two that can hybridize. Although *M. occulta* and *M. oculata* have been found to dwell in the same habitat, hybrids have not been found in nature and have only been produced in lab conditions. Fertilizing *M. oculata* eggs with *M. occulta* sperm in most cases produce embryos with fully formed tails. The reciprocal cross (*M. occulta* eggs X *M. oculata* sperm) produces a hybrid embryo with 20 notochord cells like *M. occulta*, however the notochord cells converge and extend like *M. oculata* [120]. The ascidian tail has been shown to form in the presence of notochord and the absence of muscles cells [78]. Tail development is similar in short-tailed hybrids, the notochord is not flanked by muscle as in tailed species and the tail is only as long as the notochord [123]. Hybrid embryos that develop urodele features are batch specific, and tails develop only in batches of *M. occulta* eggs that express the *p58* protein which is associated with cytoskeleton [119, 39]. Additionally, in hybrid embryos in which urodele features are restored, the number of cells that express acetylcholinesterase (AChE) in a vestigial muscle cell lineage increased in comparison to hybrids lacking urodele features and *M. occulta* [41]. This, along with evidence that the ancestral notochord—the axochord—is muscle based [57], suggests the need for both notochord and muscles cell lineages for the formation of the ascidian tail.

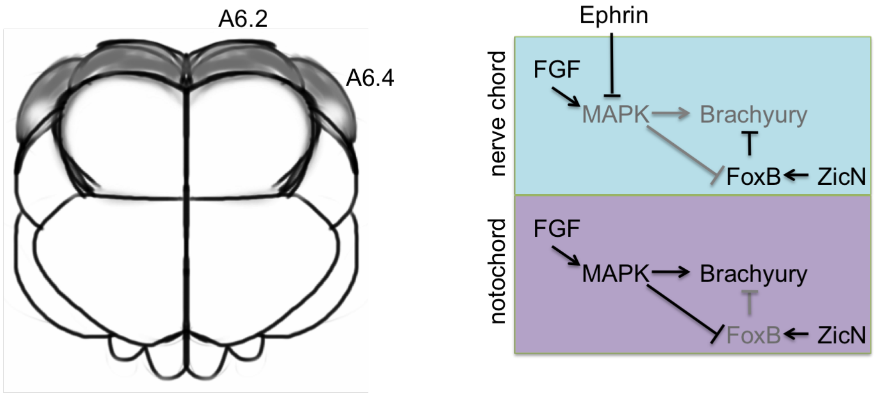


Figure 2.3: **Notochord/nerve chord precursor at 32 cell stage.** At the 32 cell stage the A6.2 and A6.4 (grey) cells of the ascidian embryo are the notochord/nerve chord precursors. Notochord induction begins at the 32-cell stage by *FGF9/16/20*. MAPK then actives *brachyury* while repressing *FoxB* in the notochord precursor cells. MAPK is repressed by *Ephrin* in the nerve chord precursors by neighboring cells. *ZicN* actives *FoxB* which represses the expression of *brachyury*, this occurs in the nerve chord cells because of the repression of MAPK by Ephrin. Genes names in black are active and gene names in grey are inactive. This figure was adopted from Hashimoto *et. al* [26]

2.2 Notochord development as seen through *Brachyury*

Brachyury, a T-box transcription factor, has been identified as essential for notochord development [140]. *Bra* was first discovered in mouse, where heterozygotes develop with shorten tails, and homozygotes fail to form a tail and die in utero [28]. The notochord is induced by *bra* in both vertebrates and ascidians, with consistent timing and expression of *bra* in the notochord and mesoderm of mouse, xenopus, zebrafish and chicken [48], with *bra* expressed exclusively in the notochord cells in ascidians [139]. Notochord induction is regulated by the *FGF/MAPK/Ets* signaling cascade (Figure 2.3) [77]. In particular, the A6.2 and A6.4 notochord/nerve cord precursors are induced by *FGF9/16/20* at the 32-cell stage, just after the 7th cell cleavage [97]. It was observed from isolation experiments that notochord/nerve cord precursors that lose *FGF9/16/20* competence at the 32-cell stage assume the default nerve cord cell fate, and the converse occurs for presumptive nerve cord blastomeres that are introduce to *FGF*, they forgo their default nerve cord fate and become notochord [140, 77]. *FGF9/16/20* activates *MAPK* which induces *bra* and represses *FoxB* [26]. *FoxB* represses the activation of *bra* predominantly through the binding of Fox BS1. Without the repression of FoxB TF, the notochord cell fate is not induced. *FoxB* is activated by *ZicN* and is present in both nerve cord and notochord precursors, however FoxB is repressed by *MAPK* in the notochord cell lineage at the 64-cell stage [26]. *Ephrin* ligand is expressed in the epidermis and signals to the future nerve chord cells, inhibiting *FGF/MAPK* pathway. Notochord cells, not being in contact with the epidermis, are free to activate *FGF/MAPK*, and activate *bra*.

Although *bra* is necessary, its presence does not guarantee a tail. *M. occulta* and *M. tectiformis*, two tailless *Molgula*, both express *bra*. In both cases *bra* expression stops earlier than that of *M. oculata*, while notochord development is slightly different between the two species. *Bra* is expressed in the 10 precursor notochord cells in *M. occulta*, which undergo another round of cell division, while this final division does not occur in *M. tectiformis.* In both *M. occulta* and *M. tectiformis*, larva-specific muscle actin genes have become pseudogenes, however the mutation in the muscle actin genes are not the same between the two species [123, 42]. *Manx*, which encodes for a zinc-finger protein, is another gene involved in tail development. *Manx* was identified through subtractive screening, and observed to be absent or lowly expressed in *M. occulta* compared to *M. oculata* [123]. The expression of *manx* was restored in the interspecific hybrids with urodele features, and demonstrated to be necessary, by abolishing the zygotic expression through antisense phosphorothioate oligodeoxynucleotides, which abolished the urodele features [121, 122].

After cell specification, the notochord cells must converge, intercalate and extend. The Planar Cell Polarity (PCP) pathway is involved in cell movement during this process and mutations in *prickle*—a known PCP gene—have been shown to cause a shortened ascidian tail by affecting both the mediolateral intercalation and the elongation of the ascidian tail [43]. The *pk* mutant *aimless* produces a truncated tail, however the polarity of the nuclei is established, showing that prickle does not establish polarity within the cell but polarity between cells, acting in a local manner and perhaps as a global organizer [43, 51]. However, even in the absence of the PCP pathway considerable convergence and elongation of the notochord was observed in Ciona, driven by a presumed boundary effect [132].

Many of the upstream genes and transcription factors that interact with *bra* have been studied in detail, through knock-outs and cell isolation experiments. A larger scale subtractive screening was done to identify genes downstream of *bra*, in which 39 genes were initially found [31]. A number of these genes have been characterized, identifying functions such as extracellular matrix components (*cadherin 8, entactin, fibronectin, laminin alpha1, alpha4, and beta1*, and *thrombospondin*), cell shape and polarity (*pk, trop, ERM, ACL*), and axon guidance (*netrin, semaphorin 3A*), amongst a host of other biological processes [30, 32, 54]. Additionally, downstream genes regulated by *bra* have been examined by using ChIP-seq to identify many known genes in the network, as well as to discover new genes [52, 45].

Larvaceans are pelagic tunicates that also develop in a typical chordate manner, featuring a notochord. However, larvaceans notochords contain only 20 cells [106, 15]. The larvacean *Oikopleura dioica* retains its tail during its adult life stage and at this point *bra* is not expressed in the adult notochord, however, *bra* is expressed in the same manner in the developing larval notochord as ascidians [1, 83]. When comparing gene networks for the extent of variation, *Oikopleura* did not exhibit the same mechanism for tail development as *Ciona*: of the 50 *bra* target genes previously identified in *Ciona*, only 26 of them had orthologs in the *Oikopleura* genome, meaning that almost 50% of candidate bra target genes are not present [53]. Of the genes that did show homology, expression ranged from notochord-specific to tail-general, including expression in possible notochord to tissues that were clearly not the notochord. From this we can infer that additional genes have gained function in notochord formation in the *C. intestinalis* gene regulatory network, and the ancestral tunicate may have had a small core set of genes for notochord formation.

2.3 Assembling and analyzing data

One of the major advances in biological sciences in the past 20 years has been the implementation of sequencing technologies. These technologies allow us to examine biological systems genomically, with increasing ease. The first widespread sequencing technology was Sanger sequencing in the 1986, but Sanger sequencing was not broadly used until 10 years later, when automated sequencers became available. Another technology, microarrays, which became popular starting in the mid '90s, allowed us to look at a wide spectrum of genes and understand relative expression within a sample. For example, Kobayashi et al. [50] isolated and analyzed gene expression in notochord (A7.3+A7.7) and nerve cord (A7.4+A7.8) precursors using microarrays. This study was able to identify 106 genes expressed in the notochord precursor and 68 expressed in the nerve cord precursor at the 64-cell stage. Of these the genes, 36 notochord genes and 25 nerve chord genes were confirmed via whole mount in situ hybridization in the respective cells. This demonstrates the power of this technique, however, prior knowledge of the gene sequences involved was needed. *C. intestinalis* was sequenced using Sanger sequencing, and is the best assembled and annotated ascidian genome [14]. In addition to long (Sanger) reads, scaffolding was done using scaffold-joining guided by paired-end expressed sequence tags and bacterial artificial chromosome (BAC) sequences, and BAC chromosomal in situ hybridization data [102]. Sanger sequencing is able to sequence whole genomes without the need of prior knowledge to identify novel genes but was costly and time consuming[76, 63].

Sanger was the first generation of sequencing technologies, and currently both second and third generation are in use, with Roche 454, Ion Torrent, Illumina and PacBio being the most widespread. These technologies produce data far more easily and at a much lower cost than Sanger sequencing [76]. There are many trade-offs for each of the technologies, including cost per MB, sequencing time, prep cost, error rate and sequencing bias; in particular, 454 and PacBio have longer reads than Illumina and Ion Torrent, 800 bp and 1+kbp, respectively. However, both Illumina and Ion Torrent's short reads are cheaper to generate, produce more reads and are better for digital counting; in addition, PacBio has a very high error rate [20]. Illumina and Ion Torrent have the lowest error rates and while Ion Torrent is more sensitive to single nucleotide polymorphisms, it also calls many more false positives. Illumina has become the most used NGS technology because it is the most versatile and performs the best in general [93]. The associated drop in sequencing price has yielded many of the assembled genomes within the Tunicata phyla. Outside of this project there are ten tunicate genomes assembled; *C. intestinalis* (species "Type A" and "Type B"), *C. savignyi*, *Oikopleura dioica*, *Botryllus schlosseri*, *Halocynthia aurantium*, *H. roretzi*, *Phallusia fumigata*, *P. mammilata*, and *Didemnum vexillum*, but no *Molgula* genomes. Molgula genomes and transcriptomes, specifically the tailed *M. oculata* and the tail-less *M. occulta* would make great systems to study the development of the chordate tail and chordate body plan.

Chapter 3

Evaluating a lightweight transcriptome assembly pipeline on two closely related ascidian species

3.1 Introduction

Next generation sequencing (NGS) has allowed us to study organisms with a broader lens, looking at entire genomes and transcriptomes instead of single genes. This capability is particularly important for non-model organisms where little prior knowledge may be available, and where NGS readily enables whole-transcriptome analyses [137], allowing us to study organisms that are ecologically or evolutionarily interesting.

There are now several sequencing technologies, Illumina being one of the most versatile [20], that can produce millions of short reads ranging from 75 to 150 bp in length at a low cost [142]. As sequencing costs continue to drop, transcriptomes from multiple developmental stages of non-model organisms can easily be sequenced. Various types of *de novo* assembly algorithms and reference based assembly approaches have been developed to handle this massive influx of transcriptomic data [92, 134, 110]. It has been shown in some cases that

[0] This chapter is in review at PeerJ: Lowe, Swalla, and Brown, 2014 (`http://dx.doi.org/10.7287/peerj.preprints.505v1`).

mapping mRNA-seq reads to a reference genome yields better transcriptomes than *de novo* assemblies, even if the genome is 5-15% divergent [133]. However, with many non-model organisms, no closely related reference genome is available.

De novo is the only solution for transcriptome assembly of organisms with no evolutionarily close reference genome. Transcriptome assemblers such as Trinity [21] and Velvet/Oases [141, 103] use De Bruijn-graph based *de novo* approaches which build graphs connecting the reads based on k-mer overlap. These graphs are then traversed via an Eulerian path algorithm to assemble transcripts. Because De Bruijn graphs are based on exact matches between DNA words, increasing numbers of sequencing errors result in an exponential number of new paths, adding to the complexity of the graph and, in turn, increasing the assembly time and memory requirements [92]. This is both time consuming and limits the ability to assembly a transcriptome by the amount of available resources.

Here we have sequenced the transcriptomes of several developmental stages of *Molgula occulta* and *Molgula oculata*—two closely related, free-spawning ascidian species, with no available reference genome. *Ciona intestinalis* and *Ciona savignyi* are the closest related ascidian species with well-assembled genomes, but are not close enough to use as a nucleotide reference for transcriptome construction. In this paper, we describe an efficient, easy to follow protocol for the transcriptome assembly of two Molgulid developmental transcriptomes. A crucial part of this protocol is the use of a preprocessing step that normalizes read abundances prior to assembly, called "digital normalization." We study the effect of digital normalization on assemblies performed with both Trinity and Velvet/Oases. We compare our approach to the results of running Trinity and Velvet/Oases without digitally normalized reads and show that our approach recovers 99% the same gene content but has significantly reduced requirements for time and memory. This reduction in time and memory lets us assemble transcriptomes efficiently using cloud resources, making our results exceptionally easy to reproduce [23], and more broadly enabling transcriptome assembly by researchers without access to large computer resources.

3.2 Methods

3.2.1 Sequencing preparation

M. occulta and *M. oculata* were collected by dredging off the shores of Roscoff, France near La Station Biologique. Swalla et al have previously described the maintenance [120] and culturing [122] of the animals. The transcriptomes of *M. occulta* and *M. oculata* were sequenced at Michigan State University (MSU) in the Research Technology Support Facility on Illumina HiSeq 2000. Five lanes of sequences were generated for *M. occulta*, two lanes of the gastrula stage (F+3), one of neurula (F+4), one of early tailbud (F+5), and one from the tailbud (F+6) stage (Table 1). Three lanes of sequences were generated for *M. oculata*, one each for the gastrula, neurula and tailbud stage. 10μg of RNA were sequenced for each stage with the exception of *M. occulta* F+4, where 1.05μg of RNA was sequenced. On average each embryonic stage yielded 48 million reads of 75 base pairs (bp) in length with paired-end insert lengths of 250 bp. All reads can be found in the NCBI short read archive (SRA) under accession number SRP040134.

3.2.2 Assembly protocol

Below is an overview of the steps used for the *de novo* assembly and annotation of our transcriptomes.

1. Quality trimming and filtering of raw reads.

2. Apply digital normalization to decrease data size.

3. Assemble transcriptome.

4. Assess transcriptome quality.

5. BLAST (gene recovery/identification).

Scripts used to run these steps can be found in the following GitHub repository: `https://github.com/ged-lab/2014-mrnaseq-cloud`

3.2.3 Pre-assembly read trimming and normalization

Low quality bases were trimmed and low quality reads were removed using quality-trim-pe.py found in the scripts directory of the repository. A hard trim is done at a Phred quality score of 33 and reads less than 30 base pairs in length are discarded. This process creates a paired and singleton fastq file for each library because of the removal of low quality reads. The filtering of reads allows for better assembly and better mapping, although it may also reduce sensitivity to low-expressed transcripts [64, 67]. The reads were initially 75 bp long, and the average base pair (bp) length was 63 bp after quality trimming and filtering. After quality trimming reads were either directly assembled, or first preprocessed with digital normalization and then assembled.

Digital normalization (diginorm) is a technique that down samples reads from highly abundant transcripts while retaining approximately the full sequence information content of the reads [5]. Here, for each species, reads from all stages were normalized together to build a common reference transcriptome; reads were normalized to a k-mer coverage of 20 with the k-mer size set to 20 as well. The initial data set from *M. occulta* contained 237 million reads from 5 lanes, and *M. oculata* contained 150 million total reads; after digital normalization, the *M. occulta* dataset was reduced to 91.6 million reads and *M. oculata* was reduced to 50 million reads, a 60% and 77% reduction respectively (Table 3.1).

Table 1: Read counts

Sample	Number of reads	Reads kept	Percentage kept	Accession Number
M. occulta F+3	42,174,510	-	-	SRR1197985
M. occulta F+3.2	50,018,302	-	-	SRR1197986
M. occulta F+4	44,948,983	-	-	SRR1199464
M. occulta F+5	53,692,296	-	-	SRR1199259
M. occulta F+6	45,782,981	-	-	SRR1199268
***M. occulta* Total**	**236,617,072**	**91,316,419**	**38.6%**	
M. oculata F+3	47,045,433	-	-	SRR1197522
M. oculata F+4	52,890,938	-	-	SRR1197965
M. oculata F+6	50,156,895	-	-	SRR1197972
***M. oculata* Total**	**150,093,266**	**49,957,980**	**33.3%**	

Table 3.1: **Digitally normalized reads.** The number of reads sequenced before and after digital normalization is shown for each lane of sequencing. The percentage of total reads kept after digital normalization is shown in bold. *M. occulta* had approximately ~237 million reads and was reduced to 91 million reads, a 60% reduction. *M. oculata* had 150 million reads and was reduced by 77% to ~50 million reads.

3.2.4 Transcriptome assembly

We used the Trinity (r20140413p1) and Velvet/Oases (v1.2.08/v0.2.08) assembler packages, both of which have been broadly used on other data sets [133, 21, 103]. Velvet was initially developed to assemble genomes, and the Oases add-on package was developed for transcriptome assembly, since transcriptomes have variable coverage and many isoforms. Since Oases cannot be run without Velvet, we refer below to transcriptomes assembled with Velvet and Oases as Oases assemblies. Unlike Trinity, Oases requires the choice of a k-mer overlap for assembly; we chose several k values ranging from k = 21 to k = 35, for odd values of k, with scaffolding turned off. After assembly, the Oases transcriptomes with the highest number of blast hits to *C. intestinalis* were selected for further analysis. The Trinity assembler was run with default parameters.

All assemblies were performed on the Michigan State University (MSU) High Performance computing cluster (HPCC). All diginorm assemblies were repeated on Amazon EC2 machines as a proof of concept. After assembly, transcripts shorter than 200 bp in length were removed, and CD-HIT was used to eliminate small transcripts with 99% identity to longer transcripts using the following command: "cd-hit-est -i <transcript file>-c 0.99 -o <output file>" [62].

To choose the best k-mer parameter for the Oases assemblies, *C. intestinalis* proteins were searched with TBLASTN (e-value cutoff of 1e-6) against each Oases assembly and the transcriptome with the most hits was selected for further analysis.

3.2.5 Gene identification

We used standalone BLAST to find reciprocal best hits (RBH) between the eight assembled transcriptomes and the *C. intestinalis* proteome retrieved from NCBI under search term "(ciona intestinalis) AND Ciona intestinalis [porgn:__txid7719]". At the time of retrieval there were 16,123 sequences and they were downloaded and stored in the GitHub repository under the file name "ciona_transcriptome.fa" in case the sequences change on NCBI. An e-value cutoff of 1e-6 was used as a minimum threshold for transcript identity. The find-reciprocal-2.py script was used to identify the RBH.

3.2.6 Read mapping

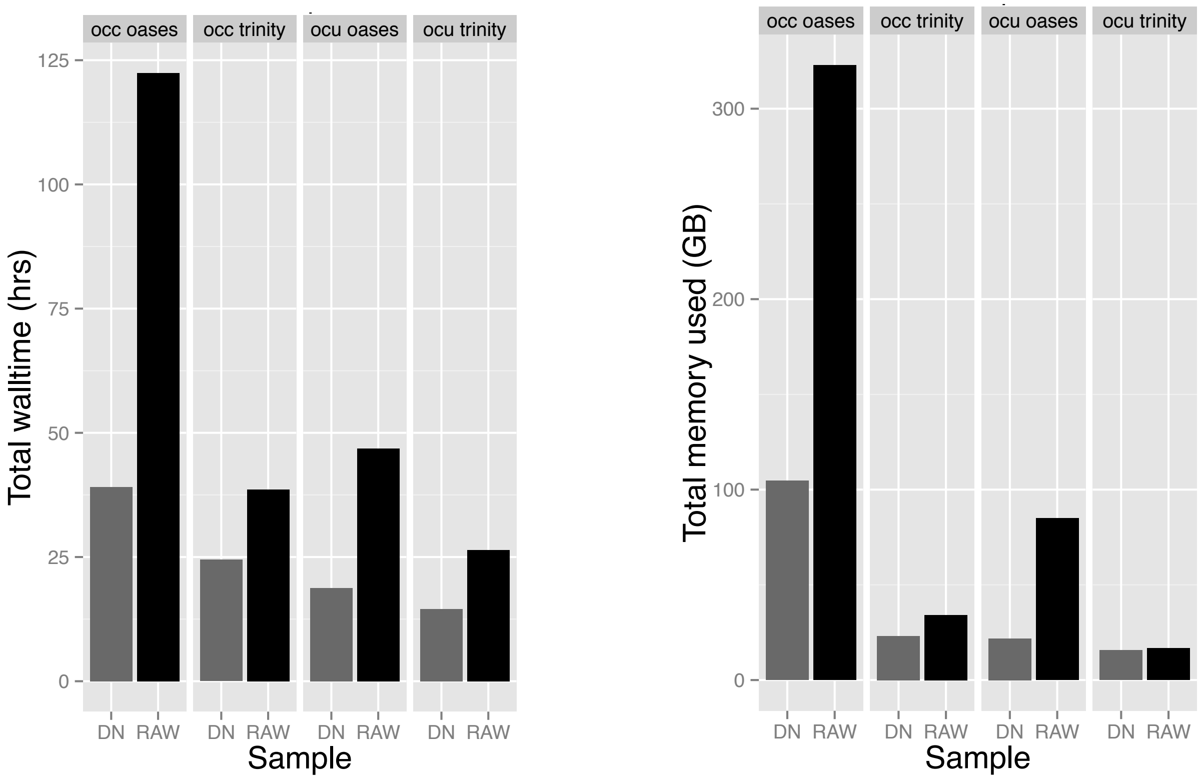
To determine the inclusion of reads in the various transcriptome assemblies trimmed reads were mapped to their respective species using bowtie2 v2.2.1 [56]. For both unnormalized read and diginorm assemblies the full set of trimmed reads were used for mapping. Default parameters were used, and both paired ends and singletons were mapped. Samtools v0.1.19 [61] was used for format conversion from SAM to BAM format, and also to calculate the percentage of mapped reads. The BAM files were also used to calculate the coverage of transcripts.

3.3 Results

3.3.1 Digital normalization reduces the resources needed for assembly

The *M. oculata* unnormalized read data set assembled with Oases used 44 CPU hours and 85 GB of RAM. The Oases assembly done with the digitally normalized reads took ~22 CPU hours and 21 GB of RAM (Figure 3.1a); this includes the time and memory required to run the digital normalization pipeline. *M. occulta* diginorm Oases assembly required over 100 GB of RAM, and the raw read Oases used 300 GB of RAM. The raw read Oases assemblies for both species took twice as long and needed at least three times as much memory when compared to the diginorm reads.

The difference in assembly time and memory between diginorm and raw reads was not as large when using the Trinity assembler. Diginorm completed its assemblies several hours faster than assembling raw reads, ~15 hours compared to ~26 hours for *M. oculata* and ~24 hours compared to ~39 hours for *M. occulta*. *M. oculata* unnormalized reads did not require much more memory than the normalized reads—16.8 GB and 15.65 GB, respectively. Diginorm had a larger effect on *M. occulta*, assembling *M. occulta* normalized reads with 23.17



(a) *Time to complete assemblies*

(b) *Memory to complete assemblies*

Figure 3.1: **Wall time and memory requirements for assemblies.** Wall time (a) in hours to complete the diginorm (DN) and raw read (RAW) assemblies for both species and assemblers. Oases assembled multiple k's, 21 $\leq$ k $\leq$ 35 opposed to Trinity that uses only a single k. This is one reason the assembly times differed. (b) Shows the memory used to assemble each of the transcriptomes. M. oculata (ocu) transcriptomes assemble in less time than M. occulta (occ) because they have fewer lanes of reads to assemble. In all cases diginorm required less time and memory to complete the assembly.

GB of RAM versus 34.14 GB of RAM for the unnormalized reads (Figure 3.1b).

3.3.2 Assembly statistics varied by preprocessing approach and assembler

Oases run with the diginormed reads yielded fewer total transcripts than Oases run with the unnormalized reads. The *M. oculata* diginorm assembly produced 300 fewer transcripts, and the *M. occulta* diginorm assembly produced 227 fewer transcripts (Table 3.2). Digital

Species	Method	N50	Mean transcripts length	Total number of transcripts	Total number of base pairs
M. occulta	DN Oases	14,606	888	89,465	79,447,700
M. occulta	Oases	14,492	912	89,692	81,824,388
M. occulta	DN Trinity	14,738	978	96,287	94,200,549
M. occulta	Trinity	12,300	914	87,090	79,672,435
M. oculata	DN Oases	7,274	1,478	39,438	58,291,461
M. oculata	Oases	7,158	1,380	39,738	54,869,493
M. oculata	DN Trinity	10,141	1,450	57,105	82,856,337
M. oculata	Trinity	8,018	1,275	49,265	62,817,433

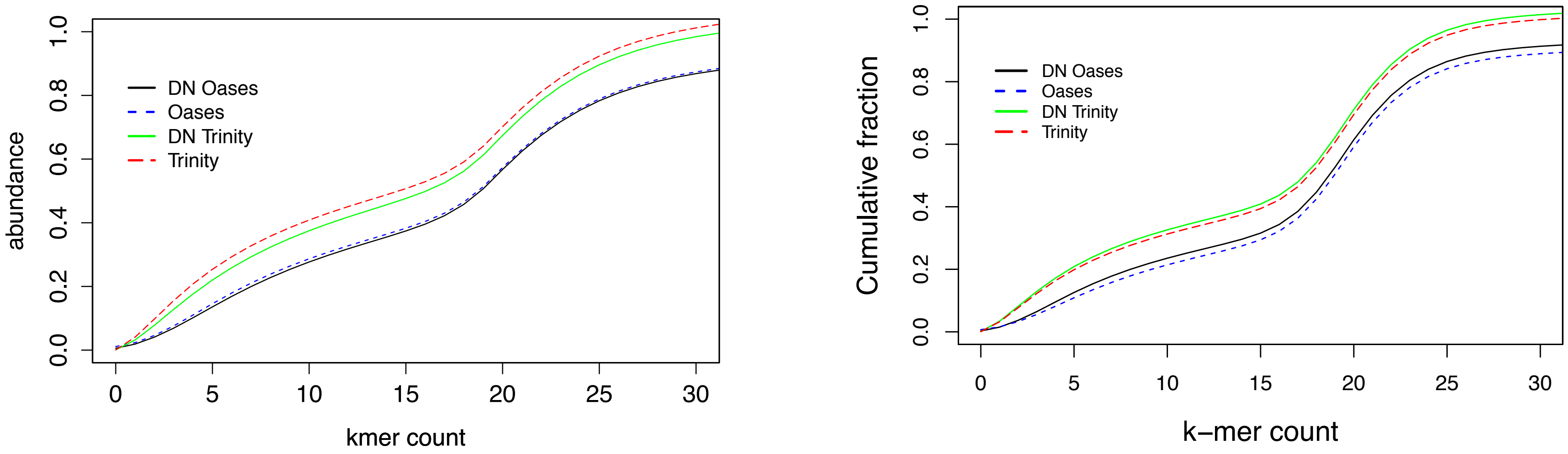
Table 3.2: **Transcriptome metrics.** Several metrics used to assess the assembled transcriptomes. The N50, mean transcript length, total number of transcripts and total number of base pairs are listed for each transcriptomes.

normalization had the opposite affect when using Trinity for assembly, increasing the total number of assembled transcripts by 7,840 for *M. oculata* and 9,197 for *M. occulta*.

Trinity produces 6.8k (7.6%) more transcripts than Oases for *M. occulta* using the digitally normalized reads, and a 2.6k (2.9%) decrease in the number of transcripts using the unnormalized reads. Trinity assembled more transcripts for both *M. oculata* assemblies, a 17.6k (44.8%) increase for diginorm and a 9.5k (24%) increase for the raw reads.

3.3.3 Trinity assemblies include more low-abundance k-mers than Oases assemblies

We next examined the k-mer spectrum of the assembled transcripts using k-mer abundances from the digitally normalized reads. The k-mer spectrum is an account of the information content of the reads and can be used to evaluate the ability of the assemblers to recover low-abundance transcripts [92]. We first used digital normalization to reduce the reads to a median k-mer coverage of 20, so that the k-mer frequency spectrum peaked at a coverage of 20, and then plotted a cumulative abundance plot of those k-mers shared between the normalized reads and the assemblies. The results, displayed in Figure 2, show that Trinity recovers more low-abundance k-mers. Also note that between assemblies done with the same assemblers, the k-mer distributions were very similar, suggesting that the k-mer spectrum is reflective of the underlying graph traversal algorithm used by the assembler. In addition the Trinity assemblies included more unique k-mers (Figure 3.3)



(a) *M. occulta*

(b) *M. oculata*

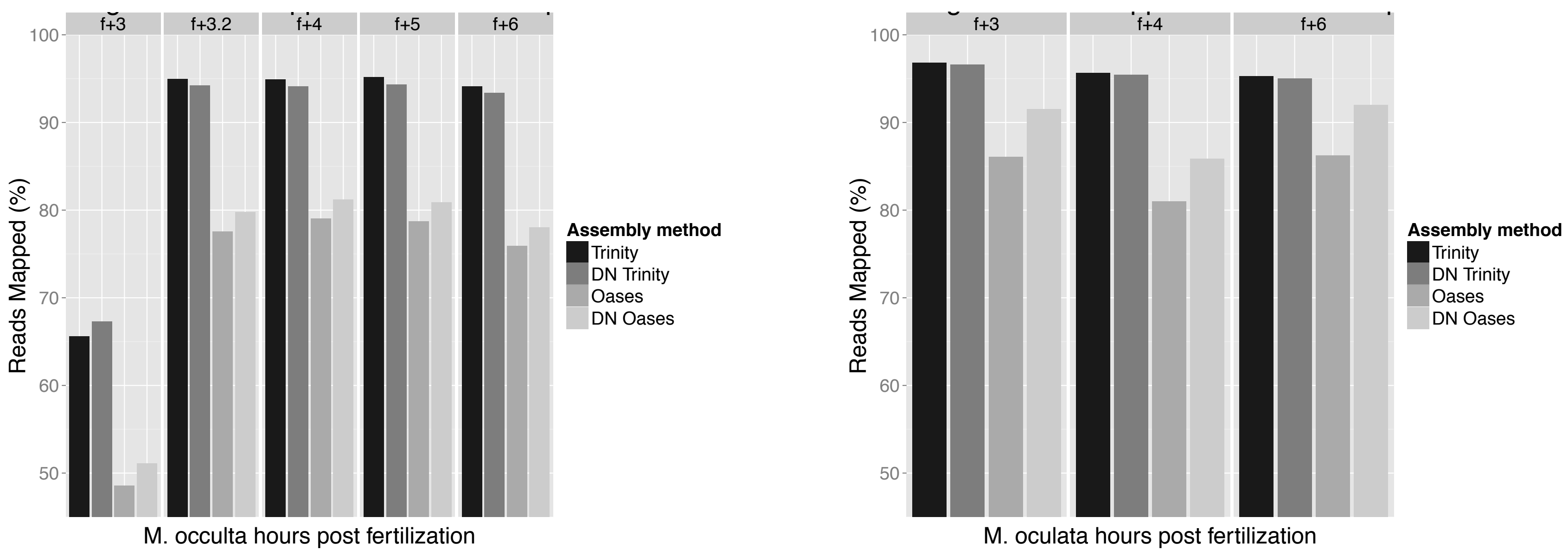
Figure 3.2: **K-mer distribution.** The k-mer distribution is shown for each assembler and assembly condition, diginorm (DN) and unnormalized reads. The k-mer distribution is the coverage of a given k-mer verses how many k-mers of that coverage are incorporated in the respective assemblies. Both Oases and Trinity assemblies are shown for (3.2a) the *M. occulta* k-mer distribution and (3.2b) the *M. oculata* k-mer distribution. Trinity had a higher k-mer distribution for both species, reflective of the inclusion of more low abundance reads into the Trinity assemblies.

Species	Method	n = 1	n = 2	n ≥ 3
M. occulta	DN Oases	60.7	18.4	20.9
M. occulta	Oases	60.3	17.4	22.3
M. occulta	DN Trinity	68.5	17.5	14
M. occulta	Trinity	73.5	16	10.5
M. oculata	DN Oases	65	17.7	17.3
M. oculata	Oases	67.1	16.4	16.5
M. oculata	DN Trinity	66.1	17.3	16.6
M. oculata	Trinity	74.2	15	10.8

Table 3.3: **Multiplicity.** The k-mer multiplicity shows uniqueness of each assembly. All k-mers with a multiplicity of one are unique. Trinity has a higher percentage of unique k-mers when comparing assemblers. The unnormalized Trinity had the highest number of unique k-mers overall.

3.3.4 Read mapping shows high inclusion of reads in the assembled transcriptomes

We mapped the quality-filtered reads to the assembled transcriptomes to evaluate their inclusiveness. The F+3 stage of reads from *M. occulta* had the lowest percentage of mapped reads, with the Oases unnormalized assembly mapping only 49% of the reads, and the Trinity unnormalized assembly mapping 67% (Figure 3.3a). This was an isolated case: all other Oases assemblies contained at least 75% of the reads for each time point and the Trinity assemblies contained at least 93% of the reads for each time point. Trinity raw read assemblies tended to contain slightly more reads than the diginorm assemblies, while the opposite was true for Oases; however, in no case did the mapping of raw-reads assembly differ from the diginorm assemblies in more than 3% of their read content.



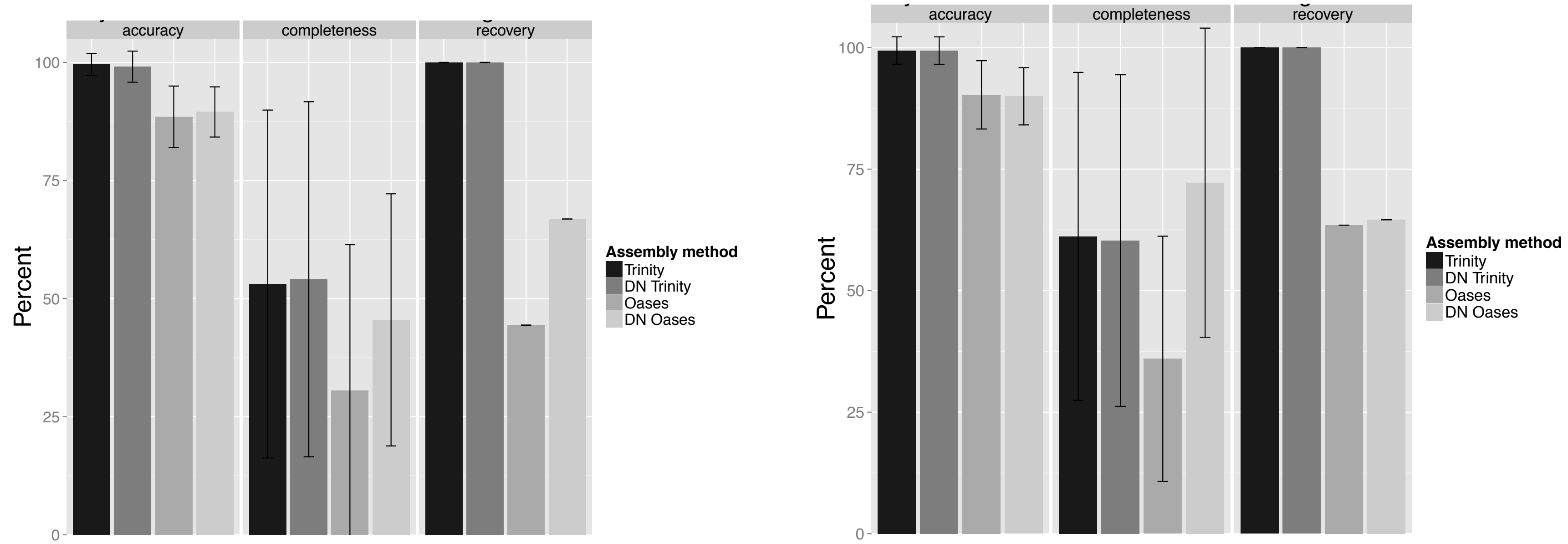
(a) *M. occulta*

(b) *M. oculata*

Figure 3.3: **Read mapping.** Unnormalized reads were mapped back to each of the assemblies to determine the inclusion of reads in the assembly. *M. occulta* first round of gastrulation reads (f+3), showed the lowest mapping quality for all assemblies, with the lowest being raw Oases at 48.57%. *M. occulta* f+3 is the only case where mapping is less than 74% and the only case where DN Trinity mapped more reads than raw Trinity. *M. oculata* unnormalized Oases performed the worst, with the Trinity assembly having the best mappings. Trinity assemblies have more mapped reads than Oases for all conditions, with at least 93% read mapping for both species. Raw Trinity typically mapped slightly more reads than DN, and the opposite occurs for Oases, with DN having more reads mapped to its assembly. Note that the Y axis starts at 45%.

3.3.5 All assemblies recovered transcripts with high accuracy but varied completeness

mRNAseq assembly accuracy can be calculated based on known transcripts generated from longer reads or reference genomes [133, 72]. We use Molgulid nucleotide sequences from NCBI to measure accuracy, and we define accuracy as the average BLAST identity score for the best match for each gene recovered [61]. There are 178 sequences from within the Molgula clade in the NCBI database. With the exception of *M. occulta* unnormalized Oases assembly, all assemblies have hits to at least 113 out of these Molgula sequences (Figure 4). The Trinity assemblies for both species have hits to all 178 sequences. Oases assemblies have hits for more sequences using digital normalized reads, two additional hits for *M. oculata* and 40 additional hits for *M. occulta. M. oculata* assemblies hits have high average accuracy in the 90 and 99 percentile for Oases and Trinity, respectively. Completeness is the percentage of a gene, transcript or protein that is recovered. Within the *M. oculata* assemblies, the unnormalized Oases assembly has the lowest average completeness at 36%, the Trinity assemblies round out at 60% and the digital normalized Oases assembly has the highest average completeness at 72%. (Note that many of the *Molgula* sequences are genomic, which includes intronic regions, so we would expect this to lower the completeness scores.)



(a) *M. occulta*

(b) *M. oculata*

Figure 3.4: **Accuracy, completeness and recovery rate against known Molgula sequences.** The NCBI has 178 Molgula sequence in its database. Transcripts were searched against these sequences using BLASTN with a cut-off of 1e-12. Trinity assemblies performed the best, recovering all known sequences. The *M. occulta* unnormalized assembly performed the worst, only recovering 79 (44%) of the transcripts. *M. occulta* tended to recover fewer of the known transcripts as well.

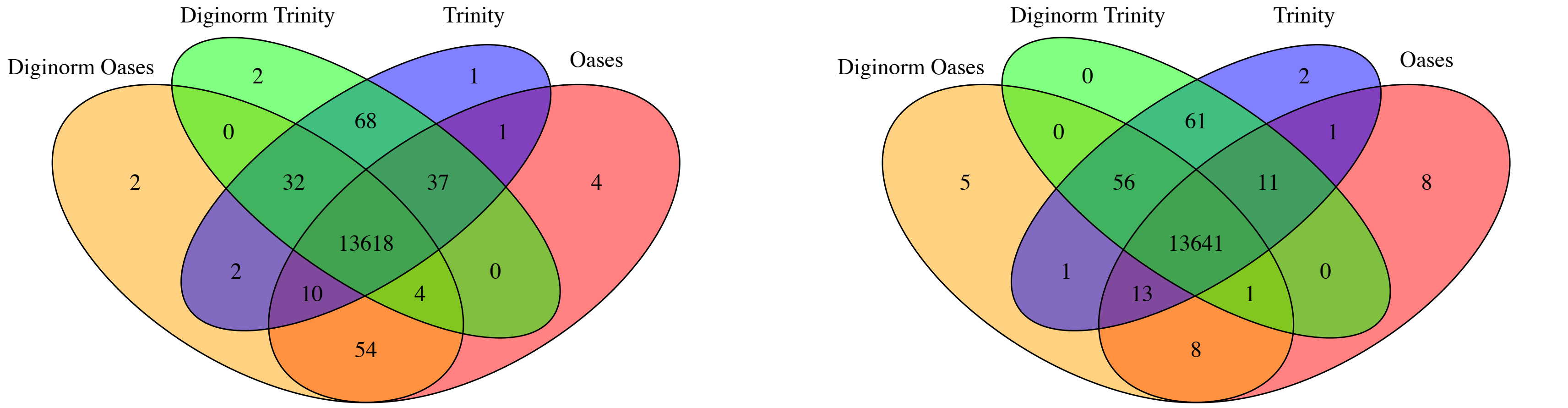
Of these 178 nucleotide sequences, 8 of them are *M. occulta* sequences and 15 of them are *M. oculata* sequences. All *M. occulta* assemblies recovered all 8 of the NCBI *M. occulta* sequences with a 94% or greater accuracy. *M. oculata* assemblies recovered *M. oculata* transcripts at a 93% accuracy as well. *M. occulta* assemblies produced the lowest completeness of the two species, 41% and 43% for unnormalized Oases and diginorm Oases respectively, and 75% for both Trinity assemblies. *M. oculata* assemblies produced more complete transcripts 66, 75, 86, and 83 percent for unnormalized Oases, Diginorm Oases, unnormalized Trinity and Diginorm Trinity respectively.

3.3.6 Both unnormalized and normalized assemblies recovered many of the same transcripts

We evaluated the two diginorm and unnormalized assemblies against one another to test whether either method missed significant portions of the transcriptome assembled by the other. We used BLAT to compare unnormalized and diginorm assemblies in both directions. In *M. occulta*, both methods recovered at least 93% of the transcripts, with Trinity diginorm recovering ~99% of Trinity's unnormalized assembly. *M. oculata* assemblies showed high overlap as well, all recovering greater than 98% of each other with the exception of diginorm Oases recovering 94% of unnormalized Oases assembly.

3.3.7 Homology search against the *Ciona* proteome shows similar recovery of ascidian genes across assemblies

We used *Ciona intestinalis* to evaluate the completeness of our transcriptomes. *C. intestinalis* has an assembled genome that is well annotated and is the closest available genome to the Molgulids. *C. intestinalis* has a genome of 160 Mb and contains ~16,000 genes [99]. A total of 13, 835 (86%) of the *C. intestinalis* proteins found in NCBI had hits in the *M. occulta* transcriptomes (Figure 5), with 2,288 genes (14%) having no hits due presumably to either lack of expression, high divergence, or loss *M. occulta.* When comparing transcripts excluded by either diginorm or unnormalized reads for all assemblies, the unnormalized read assemblies produced an additional 0.04% hits to *C. intestinalis* and there was additional 0.03% for the diginorm assemblies. There was little difference between the assemblies when compared to *C. intestinalis*, with 99% of the *C. intestinalis* genes being found in all *M. occulta* assemblies (Figure 4a). Eighty-six percent of the *C. intestinalis* proteins had matches in the *M. occulta* and *M. oculata* assemblies with less than 1% difference in presence between the several assemblies (Figure 4b).



(a) *M. occulta*

(b) *M. oculata*

Figure 3.5: **Gene recovery, raw reads versus normalized.** Gene similarity with *C. intestinalis* via BLAST for *M. occulta* (left) and *M. oculata* (right). Each oval represent the total number of putative homologous sequences recovered. In both species the Trinity assembler assembled more homologous sequences. There was almost complete overlap in homology for both assemblers and both assembly conditions.

We next examined the difference between the unnormalized and digitally normalized assemblies. Transcripts in the unnormalized assembly with BLAST hits to *C. intestinalis* but without hits in diginorm assemblies were extracted, and searched using BLASTN against the diginorm assemblies; we found fragmented versions of these transcripts, suggesting that they were partially assembled. We then mapped the diginorm reads to the extracted unnormalized transcripts and found that some portions of the transcripts were not covered by the normalized reads. This demonstrates that these transcripts were lost due to a loss of information from the diginorm process. However, the overall loss was minimal and complemented by an increase in the recovery of other conserved transcripts; this is clearly a direction for further study.

3.3.8 CEGMA analysis shows high recovery of genes

CEGMA uses a list of highly conserved eukaryotic proteins to evaluate genome and transcriptome completeness [89]. We used CEGMA to analyze the number of protein families that are present in each assembly. The default CEGMA parameters were used for analysis. CEGMA reports recovery as "complete" or "partial", where a match is marked as "complete" if 70% or more of the amino acid sequence is recovered. More than 90% of the CEGMA genes were recovered completely in each of the transcriptome assemblies, while greater than 98% of the CEGMA genes were recovered at least partially.

3.4 Discussion

3.4.1 Transcriptome assembly accurately recovers known transcripts and many genes

All of the transcriptome assemblies yielded homologs for an almost identical subset of the *Ciona intestinalis* proteome. While the evolutionary distance between the Molgulids and *C. intestinalis* may be large – the Molgulids are stolidobranch ascidians and are believed to be very divergent from *C. intestinalis*, which is a phlebobranch ascidian [34, 109]—approximately 84% of *Ciona* proteins were found in all assemblies via BLAST, and more than 44% of *Ciona* proteins had putative orthologs in each of our assemblies via reciprocal best hit. Since both transcriptomes are from a limited set of embryonic tissues that do not express all genes, these are surprisingly high numbers! We infer that we have recovered almost all embryonic genes and the majority of genes present in the Molgula genomes.

Read mapping and CEGMA analyses further confirm that the transcriptome assemblies are of high quality and inclusiveness. The assemblies represent 75% or more of the reads from all but one time point, contain complete matches to 90% or more of the conserved eukaryotic gene families in CEGMA, and contain partial matches to 98% or more of the CEGMA families. It is important to note that the CEGMA results are almost certainly biased upwards by the nature of the CEGMA families, which represent many more metabolic and cellular function genes than e.g. animal-specific transcription factors; thus the CEGMA numbers do not directly demonstrate the inclusiveness of the transcriptome families, as they would for a genome assembly [89].

3.4.2 Digital normalization eases assembly without strongly affecting assembly content

One of our goals in this study was explore the impact of digital normalization on the biological interpretation of transcriptome assemblies; while previous studies have shown that digital normalization can make assembly faster and less memory intensive, gene recovery has been less well studied [23, 5]. Here we confirm the computational results: diginorm dramatically reduces the computational cost of Oases assemblies, and also decreases the time and memory requirements for Trinity assemblies.

While digital normalization does alter the number of transcripts significantly, it does not strongly affect either read inclusion or the conserved gene content of the assemblies. Read inclusion by mapping never decreased more than 3% after digital normalization, and in many cases increased. The conserved gene content, measured by a proteome comparison, showed that we recover essentially the same set of proteins with all four treatments on both transcriptomes.

Combined, these results suggest that the varying number of transcripts largely reflect differences in the splice variants reported by different assemblers under different conditions. These results also strongly support the idea that preprocessing with digital normalization does not strongly affect assembly content. We note, however, that the few transcripts not recovered in assemblies of the digitally normalized reads were probably not recovered because the underlying reads were eliminated during digital normalization. This is an area where digital normalization can be improved.

Only a small number (well below 1%) of different homology matches were reported between the various assemblies. Because of this we decided not to merge or otherwise combine the different assemblies: the likely benefits were outweighed by the risk of introducing chimeric transcripts or combining isoforms.

We also note that the variation in number of assembled transcripts due to read preprocessing and choice of assembler despite the similar gene content suggests that traditional genome assembly metrics such as number of transcripts, total bp assembled, and N50 are not useful for transcriptome evaluation as previously suggested [87]. For example, the same exon may be included in multiple splice variants, inflating the total bp assembled; some assemblers may choose to report more isoforms than others even with the same read support; and N50 makes little sense for transcriptomes.

3.4.3 Trinity assemblies are more sensitive to low-abundance k-mers but contain no new conserved genes

The difference in transcript numbers between Trinity and Oases assemblies is stark: for the same data set, with the same treatment, Trinity always produces thousands more transcripts than Oases. Moreover, many more reads can be mapped to the Trinity assemblies —an additional 10% or more, for every stage. Despite this greater inclusion of reads, we see no substantial gain in either CEGMA matches or *Ciona* proteome matches for the Trinity assemblies.

This conundrum can be resolved by examining the k-mer spectra, which show that the Trinity assemblies include many more low-abundance k-mers from the read data set. This demonstrates that Trinity is more sensitive to low-abundance sequences, and may include more isoforms in its assemblies—by design, Trinity attempts to be more sensitive to isoforms than Oases, and focuses particularly on low-coverage isoforms [133, 21, 131]. Those transcripts were indeed the results of Trinity assembling low coverage reads, having an average coverage of 5x compared to 75x.

3.5 Conclusions

We show that transcriptome assembly on two closely related species of Molgulid ascidians produced accurate and high-quality transcriptomes, as determined by several different metrics. Importantly, four different assembly protocols produced transcriptomes that contained nearly identical complements of homologs to the nearest model organism, *Ciona intestinalis*. While variations in isoform content were observed, these variations had little apparent impact on sensitivity of homologous gene recovery. We provide detailed assembly protocols that should enable others to easily achieve *de novo* transcriptome assemblies.

3.6 Acknowledgments

EKL and this research were supported by the National Science Foundation under Cooperative Agreement No. DBI-0939454 (BEACON). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. CTB was supported in part by Agriculture and Food Research Initiative Competitive Grant no. 2010-65205-20361 from the United States Department of Agriculture, National Institute of Food and Agriculture.

Chapter 4

Genome assembly and characterization[1]

4.1 Introduction

Ascidians are marine invertebrates that spend their adult life filter feeding through an incurrent siphon and an outcurrent siphon. Ascidians are evolutionarily interesting because of the phylogenetic position – they are tunicates, the sister group to vertebrates and cephalochordates, with whom they form the chordate phylum. Although ascidians share little morphological resemblance to vertebrates in their adult stage, they do share several features in their larval stage: a notochord, dorsal hollow neural tube, and gill slits during development [136, 7].

The development of ascidians is well documented, and the cell lineage from fertilization to gastrulation has been described thoroughly in *Ciona intestinalis* [84, 85, 82]. Studies of other ascidian species have shown that the majority of the phyla members have an invariant cell lineage and typical development [2]. However, a few solitary ascidians have deviated from the typical developmental program and undergone tail-loss [120, 129]. *M. occulta* and *M. oculata* are two species that are found in the shallow waters for Roscoff, France that closely resemble each other– in their adult stage, they differ only by a white pigment spot

[1] Portions of this section were published in Stolfi et al., [116].

found between the siphons of *M. oculata* (Figure 4.1). These two *Molgula* species, however, have different methods of development—*M. oculata* develops as a typical tadpole larvae and *M. occulta* develops without a tail. The underlying molecular reasons for divergence are unknown.

Many ascidian genes have been studied across a number of ascidians, showing that gene function tends to be orthologous within the phyla [98]. Although genes tend to be expressed in homologous patterns and tissues, the presence of genes are not the same across species. There are a number of cases where a gene that has been shown to be necessary for a phenotype in one species is completely absent in other ascidian species with the same phenotype [60]. Ascidian species are far more divergent than they appear phenotypically. It has been shown that in ascidians with the same phenotype and gene expression, regulatory modules are not necessarily the same [35, 53, 115]. This is often attributed to the conservation of gene regulatory networks (GRNs) and the flexibility of TF binding site distribution in a given enhancer, which contribute to conservation of enhancer function [25]. This regulatory turnover is termed "developmental system/systems drift" (DSD) [128]. This term broadly applies to the divergence in the molecular or morphogenetic basis for the development of identical homologous characters.

Tunicates have even deviated from *hox* patterning and function [37], and here genomics has shed some light on the area. Ascidians are broadcast spawners, which leads to them being highly polymorphic and having rapid rates of evolution [14]. This drives rapid divergence in genomes, as well as change of gene function when compared to other chordates [60]. Through whole genome sequencing and assembly of two closely related species *M. occulta* and *M. oculata*, and the more divergent *M. occidentalis*, we have demonstrated that *M. occidentalis*, *M. occulta*, and *M. oculata* all have different *hox* configurations, and while having invariant

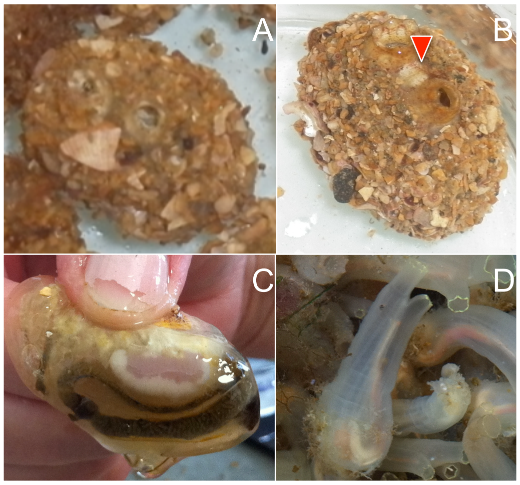


Figure 4.1: **Adult ascidians.** *M. occulta* (A) and *M. oculata* (B) are nearly identical in their adult stage with the white pigment spot (red arrow). Their tunic is covered in sand, since they are found on the sandy sea bottoms. Under their sand covered tunic, the two species differ by the color of their eggs—purple in *M. oculata*, pictured, and an orange-yellowish color in *M. occulta*—which are found just above the kidney complex (C). *C. intestinalis* (D) is the best studied ascidian and has a well-assembled and annotated genome [14, 102].

development, the cis-regulatory elements behind the development have diverged between *Ciona* and *Molgula*.

4.2 Materials and methods

4.2.1 Genomic DNA library preparation and sequencing

Genomic DNA was phenol/chloroform extracted from dissected gonads of *Molgula occulta* (Kupffer) and *Molgula oculata* (Forbes) adults from Roscoff, France, and a *Molgula occidentalis* (Traustedt) adult from Panacea, Florida, USA (Gulf Specimen Marine Lab). Genomic DNA was sheared using an M220 Focused-ultrasonicator (Covaris, Woburn, MA). Sequencing libraries were prepared using KAPA HiFi Library Preparation Kit (KAPA Biosystems, Wilmington, MA) indexed with DNA barcoded adapters (BioO, Austin, TX). Size selection was performed using Agencourt (Beckman-Coulter, Brea, CA) AMPure XP purification beads (300-400 bp fragments), or Sage Science (Beverly, MA) Pippin Prep (650-750 bp and 875-975 bp fragments). For *M. occulta* and *M. occidentalis* libraries, 6 PCR cycles were used. For *M. oculata* libraries, 8 cycles were used for the 300-400 bp library, and 10 cycles were used for the 650-750 and 875-975 bp libraries. Libraries of different species but same insert size ranges were multiplexed for sequencing in three 2x100 PE lanes on a HiSeq 2000 sequencing system (Illumina, San Diego, CA) at the Genomics Sequencing Core Facility, Center for Genomics and Systems Biology at New York University (New York, NY). Thus, each lane was dedicated to a mix of species, specifically barcoded libraries of a given insert size range. Raw sequencing reads were deposited as a BioProject at NCBI under the ID# PRJNA253689.

4.2.2 Genome sequence assembly

All genomes were assembled on Michigan State University High Performance Computing Cluster (`http://contact.icer.msu.edu`). Prior to assembly, read quality was examined using FastQC v0.10.1. Reads were then quality trimmed on both the 5' and 3' end using seqtk trimfq (`https://github.com/lh3/seqtk`) which uses the Phred algorithm to determine the quality of a given base pair. Seqtk trimfq only trims bases, so no reads were discarded. Each library per species was then abundance filtered using 3-pass digital normalization to remove repetitive and erroneous reads [5, 104, 33]. Genome assembly was done using velvet v1.2.08 [141] with k-mer overlap length ('k') ranging from 19 to 69 and scaffolding was done by Velvet, by default. Velvet does not produce separate files for contigs and scaffolds; because Velvet scaffolded conservatively, contigs dominated the assemblies so we refer to both contigs and scaffolds as contigs. CEGMA scores were then computed to evaluate genome completeness [89]. The latest versions of three species' genome assemblies have been deposited on the

ANISEED (Ascidian Network for In Situ Expression and Embryological Data) database for browsing and BLAST searching at http://www.aniseed.cnrs.fr/ [126]. Scripts for genome assembly and CEGMA analysis can be found in the following github repository: https://github.com/ged-lab/2014-molgula-genome-assemblies

4.2.3 Gene identification and alignments

Thirty-nine hox genes were identified in human and downloaded from the NCBI database. These sequences were then BLASTed against each of the three assembled *Molgula* genomes. The alignments were then extracted and BLASTed against the NCBI non-redundant database. *Molgula* aligning sequences were extracted, annotated and placed in the following files, mocc_hox_aa.fa, mocu_hox_aa.fa, and moxi_hox_aa.fa, which are located at https://github.com/ged-lab/2014-elijahlowe-thesis in the directory data/. *Hox1-13* sequences for human, fruit fly, and Amphioxus were download from 'Homeobox Database' http://homeodb.zoo.ox.ac.uk/. These sequences were then joined in a multifasta file with the identified Molgula *hox* genes and used to produce a phylogenetic trees using MAFFT version 7 online rough tree program at http://mafft.cbrc.jp/alignment/server/clustering.html [47, 46]. Additional alignments between the three species were conducted using mVista [73, 18, 135] with *M. oculata* as the anchoring sequence because it shows the most similarity between the three *Molgula* species. The LAGAN alignment algorithm was used with translated anchoring to improve alignment because of evolutionary distances[6].

Species	N50	Mean contig length	Total	Total number of base pairs	CEGMA C[1]	CEGMA P[2]
M. occidentalis	26,298	5,072	51,761	262,547,660	81.45	96.77
M. occulta	13,011	3,233	58,489	189,110,562	77.42	98.79
M. oculata	34,042	6,270	25,497	159,886,716	89.92	99.19

Table 4.1: **Genome assembly statistics.** The contig N50 length, mean contig length, total number of contigs, total number of base pairs and CEGMA scores were collected for each draft assembly. The CEGMA score is a metric of completeness measured against highly conserved eukaryotic genes. Alignments of 70% or greater of the protein length are called complete (C) and all other statistically significant alignments are called partial (P).

4.3 Results

4.3.1 Genome assemblies assessment

Genomes of three Molgula species (*M. occidentalis*, *M. oculata*, and *M. occulta*) were sequenced using next-generation sequencing technology and assembled. A common metric for judging the quality of a genome assembly is the contig N50 length, which is determined such that 50% of the assembly is contained in contigs of this length or greater. We used the contig N50 length to select the best assembly for each species given the varying 'k' parameter (length of k-mer overlap). A 'k' of 39 yields the best assembly for both *M. occidentalis* and *M. occulta*. The best 'k' for *M. oculata* was 61. *M. occidentalis*, *M. occulta*, and *M. oculata* N50 lengths were approximately 26.3 kb, 13 kb, and 34 kb, respectively (Table 4.1).

In addition to N50 lengths, we also used CEGMA (Core Eukaryotic Genes Mapping Approach) scores, in order to evaluate the assemblies' representative completeness [89]. CEGMA reports scores for complete and partial alignments to a subset of core eukaryotic genes. An alignment is considered "complete" if at least 70% of a given protein model aligns to a contig in the assembly, while a partial alignment indicates that a statistically significant portion of the protein model aligns. The partial alignment scores are ~97% or higher for all assemblies. *M. oculata* has the best complete alignment score at ~90%. *M. occidentalis* and

M. occulta have complete alignment scores of 81% and 77% respectively (Table 4.1). These scores indicate that our assemblies contain at least partial sequences for the vast majority of protein-coding genes in the genomes of these species. Various factors make it unreliable to predict genome size and gene density based on assembly metrics alone [3]. Of the handful of sequences we isolated and analyzed, we found that the sizes of introns and upstream regulatory regions were roughly comparable to those from their *Ciona* orthologs. This suggests that the *Molgula* genomes may be as compact as the *C. intestinalis* genome (i.e., ~150-170 Mb, ~16,000 genes) [55, 107, 102].

4.3.2 Gene complexes

Hox genes are a subset of the homeobox genes and are known to be involved with the establishment of morphological identities along the anteroposterior axis of bilaterians and cnidarians [17]. All *hox* genes have a highly conserved 60 amino acid (aa) homeobox sequence [74, 19] and *hox* genes are distinguished by variation of the homeobox domain. There are 4 *HOX* clusters in humans totaling in 39 genes. Within tunicates, *C. intestinalis*, *Halocythia roretzi* and *Oikopleura dioica hox* genes have been characterized. *C. intestinalis* has 9 *hox* genes, *Hox1* through *6*, *Hox10*, and *Hox12-13* [14]. The *hox* gene of *C. intestinalis* were initially found on 5 scaffolds spanning ~980 kb using the draft assembly, with *hox2-4*, *hox5-6* and *hox12-13* being found on the same scaffold and later identified to be two clusters of *hox* genes across two chromosomes [38]. *O. dioica* also has 9 *hox* genes, *hox1-2, hox4*, a duplicate *hox9*, and *hox10-13*, however, none of the genes have been found on the same scaffold, even using a 250 kb window [108].

Eight *hox* genes have be found in *M. occulta* and *M. oculata*, while nine have been found in *M. occidentalis*. The eight found were *Hox1*, *hox2*, *hox3-4*, *hox5*, *hox10* and *hox12-13*, with *hox3* and *hox 4* being found on the same contig in all three species (Figure 4.2). Additionally *hox10*, and *hox12-13* are found on the same contig in *M. oculata* with only *hox12-13* being found on the same contig in *M. occidentalis*. However, it appears that the *hox* genes have been rearranged in *M. oculata*, since *hox10* is downstream of *hox12-13*. *M. occidentalis* had one additional *hox* gene compared to *M. occulta* and *M. oculata*, a duplicate *hox10* gene ~12kb apart found on the same contig. Also, the *M. occidentalis hox2* has a stop codon located in the 3-4 helix, although this may be specific to the animal examined and not the total population. The second *hox10* sequence was not fully sequenced, missing 14 aa of the homeobox domain, and the identity of the two sequences is 52.1% at a nucleotide level, 53.4% at a protein level, and 91.3% identical within the homeobox domain (Figure 1.4). *M. occulta*, *M. oculata*, and *M. occidentalis hox* genes span across 7, 5 and 6 contigs respectively and are 197 kb, 311.7 kb and 279 kb in length, respectively. This is more compact than the *Ciona hox* cluster which exhibits longer than usual intergenic regions, averaging in the 5Mb range, when typically the *hox* genes have 100-120 kb intergenic regions separating them [75]. The *hox* gene in the *Molgula* had far smaller intergenic regions, 10-25 kb in length for the *hox* genes that were found on the same contigs.

4.3.3 Divergence of Gene regulatory networks

Our sequencing efforts revealed extreme genetic divergence not only between *Ciona* and *Molgula*, as expected, but even within the Molgulids. For example, in Stolfi et al., [115] we used BLAST to identify the *Molgula* orthologs of *C. intestinalis Mesp*. *C. intestinalis Mesp* is the sole ortholog of vertebrate genes coding for *MesP* and *Mesogenin bHLH* transcription factor family members [102]. VISTA alignment shows high sequence similarity between sequences 5' upstream of the *Mesp* genes from the closely related *M. oculata* and *M. occulta*.

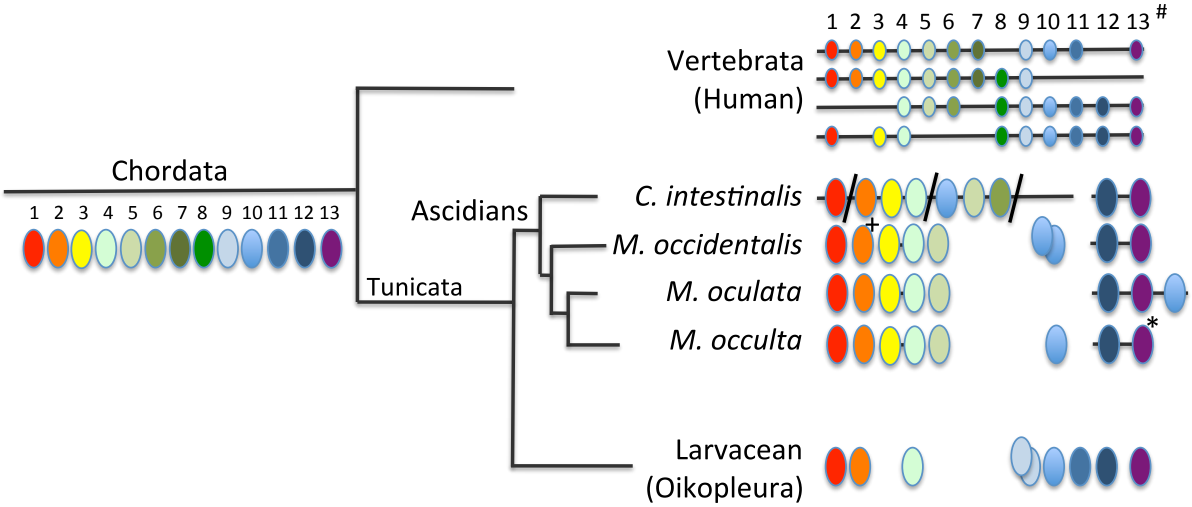


Figure 4.2: **Hox clusters for *M. occulta*, *M. oculata* and *M. occidentalis*** Eight *hox* genes were found in *M. occulta* and *M. oculata*, while nine were found in *M. occidentalis*. *Hox1*, *hox2*, *hox3-4*, *hox5*, *hox10* and *hox12-13* were found in all three *Molgula* species. *Hox3-4* was found on the same contig in all species, with *hox12-13* found on the same contig in *M. occidentalis* and *M. oculata*. **M. occulta hox12-13* is not found on the same contig, but when aligned using mVista, there is high sequence similarity, showing the possible placement of *hox12-13* in *M. occulta*. $^+$*M. occidentalis hox2* gene had a stop codon found in the 3-4 helix. # numbers correspond to gene color, and rearrangements have been found in *Ciona* and *Molgula*.

However, there is no conservation of *Mesp* DNA sequences, coding or non-coding, between *M. oculata/occulta* and *M. occidentalis*, nor between *C. intestinalis* and any of the three *Molgula* species. In previous phylogenetic surveys, *M. occidentalis* has been placed as an early-branching *Molgula* species, often grouped together in a subfamily with species ascribed to the genera Eugyra and Bostrichobranchus instead [24, 34, 129]. Our sequencing results support the view that *M. occidentalis* is highly diverged from other *Molgula* species.

This sequence divergence is also evident when analyzing the *hox* genes. When comparing sequence similarity of the *hox* genes that were found on the same contigs (*hox3-4* and *hox12-13*), only regions clustered around the coding region for *M. occulta* when compared to *M. oculata* showed similarity, and only highly conserved coding domains showed similarity in *M. occidentalis* when compared *M. oculata* (Figure 1.1). This sequence similarity was a lot less obvious when comparing *M. occulta* to *M. occidentalis.* However, because of a lack of synteny outside of coding regions between *M. occidentalis* and *M. oculata* we were able to identify *distal-less*, downstream of *Hox13*, which is expressed in endodermal strand cells in *Ciona.*

4.4 Expression of the *M. occidentalis Mesp* gene marks the B7.5 cells

Ciinte.Mesp specifies the B7.5 cells as the sole progenitors of the cardiopharyngeal lineage [101, 13, 29, 113]. We performed RNA in situ hybridization (ISH) for *M. occidentalis Mesp (Moocci.Mesp)* and found that this gene is also expressed only in the B7.5 cells of *M. occidentalis* embryos (Figure 4.3C). These cells are unequivocally identified due to the perfect conservation of early embryonic cell cleavage patterns in all ascidians. ISH for the *M. oc-*

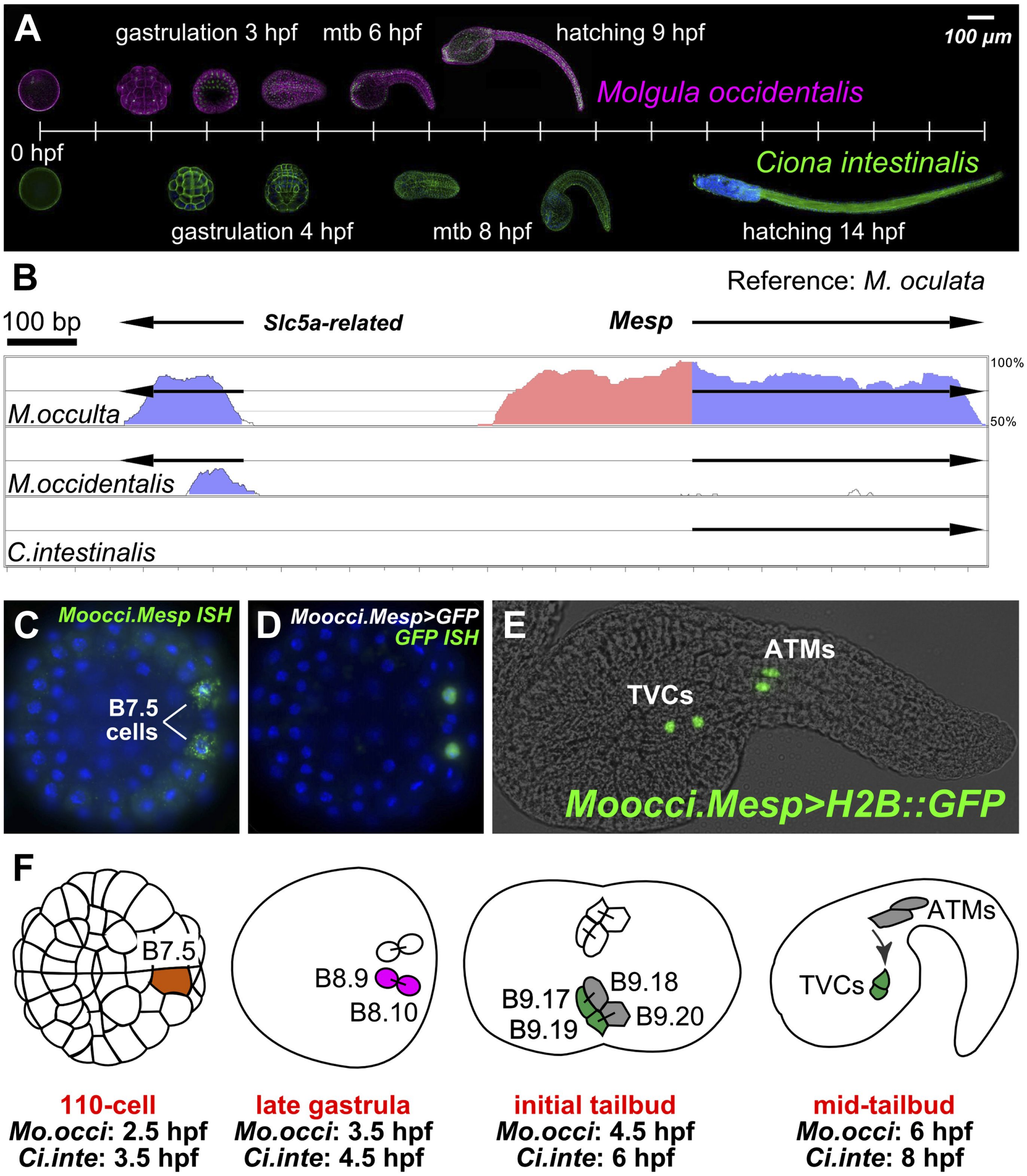


Figure 4.3: **The B7.5 lineage in *M. occidentalis*** (A) Diagram comparing *M. occidentalis* (top) and *C. intestinalis* (bottom) embryogenesis at 24C. Embryos were stained with Alexa Fluor dye-conjugated phalloidin to visualize cell outlines and DAPI to visualize cell nuclei. (B) Diagram of mVISTA ([18]; genome.lbl.gov/vista/) alignment of *M. oculata Mesp (Moocul.Mesp)* locus to orthologs in *M. occulta*, *M. occidentalis*, and *C. intestinalis*. Shaded peaks indicate sequence conservation above 70% over 100-bp windows (blue = protein-coding, pink = non-coding). Arrows indicate direction of transcription of protein-coding genes. Non-coding sequences upstream of *Mesp* are only conserved between *M. oculata* and *M. occulta*. *M. occidentalis* and *C. intestinalis* show considerable divergence even in protein-coding sequences. Note that microsynteny with SLC5A-related gene supports the orthology of these sequences among the Molgulids.

Figure 4.3 (cont'd): (C) In situ hybridization (ISH) for *Moocci.Mesp* in 110-cell stage embryo (vegetal view), showing mRNA detection (green) in B7.5 blastomeres. Nuclei were counter-stained with DAPI (blue). Staging is given by hours post-fertilization (hpf). (D) Vegetal view of a 110-cell stage embryo electroporated with *Moocci.Mesp*>GFP reporter construct. Reporter gene expression was detected by ISH for GFP transcripts (green). Nuclei were stained with DAPI (blue). (E) Lateral view of a mid-tailbud stage embryo electroporated with Moocci. Mesp>Histone2B::GFP reporter construct. GFP fluorescence reveals B7.5 descendants on left side of embryo: two trunk ventral cells (TVCs) and two anterior tail muscles (ATMs). (F) Diagram of B7.5 lineage divisions from 110-cell stage to mid-tailbud stage, inferred from previous C. intestinalis studies. Cells are named according to Conklin's method (Conklin, 1905). The lineage is bilaterally symmetric, but only cells on the left side are indicated and named. Relative staging given for *M. occidentalis (Mo.occi)* and *C. intestinalis (Ci.inte)*. 110-cell and late gastrula: vegetal view. Initial tailbud: dorsal view. Mid-tailbud: lateral view. Anterior pole is on the left in all images and illustrations. DOI: 10.7554/eLife.03728.003

culta Mesp gene (*Mooccu.Mesp*) also revealed conserved expression in this tail-less species (Figure 4.3). We successfully adapted the Ciona electroporation protocol for simultaneous transfection of reporter gene plasmids into hundreds of synchronized *M. occidentalis* embryos (Figure 4.3D,E). We were also able to electroporate *M. occulta* embryos (Figure 4.3). However, only *M. occidentalis* was routinely available to us for in vivo studies, so we focused our experiments on this species. Development of *M. occidentalis* embryos was optimal at 24C and faster than that of *C. intestinalis* (Figure 4.3A). Using electroporation-based transfection, we determined that an ~1.1 kb genomic DNA fragment upstream of *Moocci.Mesp* is able to drive expression of fused reporter genes specifically in the B7.5 cells with no 'leaky' expression in other cells as is commonly observed in *C. intestinalis* (Figure 4.3D; [112]). This faithful recapitulation of *Moocci.Mesp* expression and the persistence of GFP allows for visualization of the descendants of B7.5 long after endogenous Moocci.Mesp transcription has ceased (Figure 4.3; [13]). At the tailbud stage, we find that each B7.5 blastomere gives rise to four grand-daughter cells (Figure 4.3E). The two anterior B7.5 grand-daughter cells on either side of the bilaterally symmetric embryo migrate anteriorly and are termed the trunk ventral cells (TVCs) due to their final position in the *C. intestinalis* and *H. roretzi* embryos [82]. Their posterior sister cells remain in the tail and become anterior tail muscles (ATMs). As far as we can tell, B7.5 lineage ontogeny is perfectly conserved between *M. occidentalis* and *C. intestinalis* (Figure 4.3F).

4.5 Divergence of Mesp cis-regulatory sequence function between *M. occidentalis* and *C. intestinalis*

Given the obvious parallels between *C. intestinalis* and *M. occidentalis* cardiopharyngeal development, we expected transcriptional regulatory mechanisms to also be highly conserved between the two species. We tested this assumption by electroporating *C. intestinalis* reporter constructs into *M. occidentalis* embryos, and vice-versa. We observed that a *Ciinte.Mesp* reporter construct [13], when electroporated into *M. occidentalis* embryos, drives relatively weak reporter gene expression in B7.5 with substantial leaky expression in other tissues (Figure 4.4 (cont'd)A). Conversely, the Moocci.Mesp enhancer fails to drive any reporter gene expression when electroporated into *C. intestinalis* embryos (Figure 4.4 (cont'd)B), despite recapitulating robust B7.5-specific expression in *M. occidentalis* embryos (Figure 4.3D,E). These data suggest acute DSD of transcriptional regulatory mechanisms underlying otherwise identical *Mesp* expression patterns. More specifically, the trans-regulatory environment of the B7.5 blastomeres has diverged between Molgula and Ciona, and compensatory changes in the respective *Mesp* cis-regulatory sequences must have rendered these unable to function adequately outside of that milieu.

The observation of *Mesp* incompatibility or unintelligibility, lead to the examination of Foxf cis-elements, who is also involved in TVC specification, both revealed divergent cisregulatiory logic for underlying identical gene expression patterns. This sparked an interest to see if there was a general trend of cis-regulatory unintelligibility between *C. intestinalis* and *M. occidentalis*. This revealed that *M. occidentalis Tbx6* we found that the Ciinte.Hand-r reporter can drive reporter gene expression in *M. occidentalis* TVCs. Thus, unlike the case of Foxf, there is an asymmetric intelligibility of Hand-r TVC enhancers between *M. occidentalis* and *C. intestinalis*. Moreover, a *M. oculata Hand-r* TVC enhancer is functional in *M. occidentalis* but not in *C. intestinalis*. Taken together, these data suggest that differences in enhancer logic may have accumulated over the course of the deep evolutionary history between Molgula and Ciona but not between *M. occidentalis* and *M. oculata*, and that some enhancers may have evolved asymmetrically in the two branches, retaining greater pan-ascidian "intelligibility" in one or the other.

4.6 Discussion

Three *Molgula*—*M. occulta*, *M. oculata* and *M. occidentalis*—species have been sequenced and assembled; these are the first of any molgulids to have assembled genomes. Developmentally the three species are very similar up to the gastrula stage, where *M. occulta* diverge from the typical solitary ascidian body plan and develop without a tail [2, 82].

In vertebrate and other bilaterians the *hox* genes has been shown to be important for patterning along the anterior-posterior axis [17, 70]. The same has not been shown in ascidians, where *hox* has more of a tissue specific role [37]. *Ciona* has 10 *hox* genes and is missing *hox7-9* and *11*, with *hox10* and *12* being the only two genes to show morphological effects when knocked down. *Hox10* is involved in the regulation of the motor neuron differentiation and *hox12* is involved in tail development, through the elongation of the posterior most

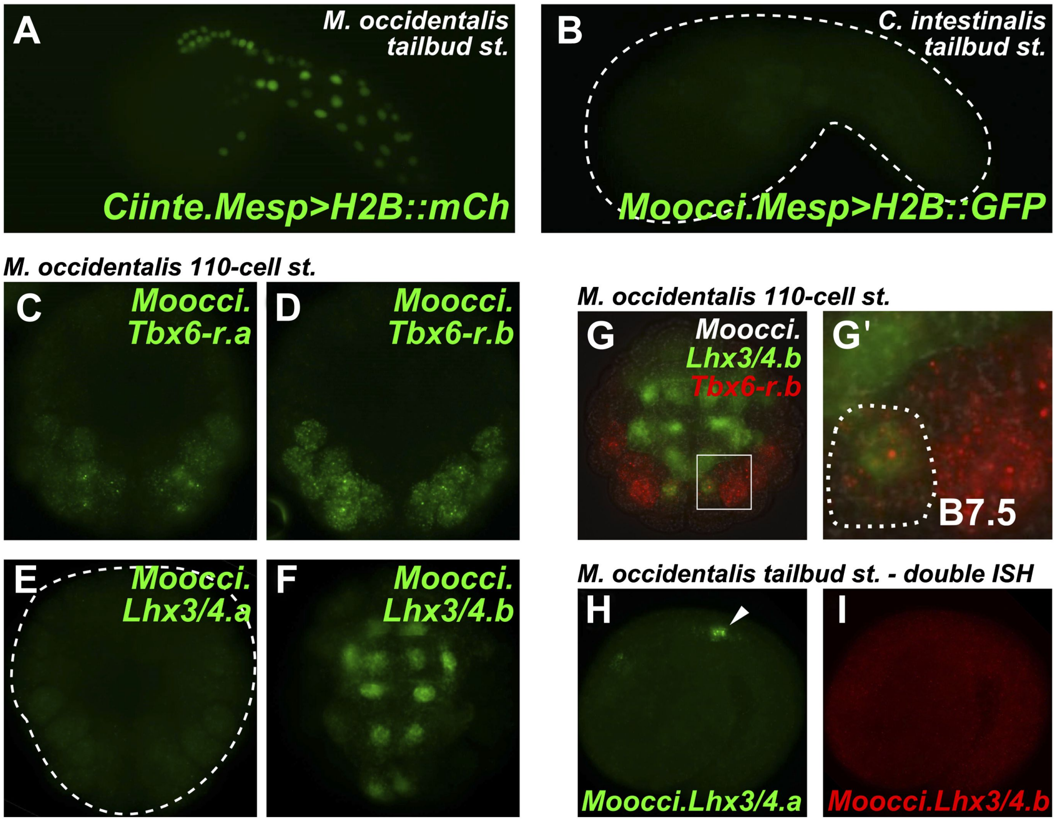


Figure 4.4: **The B7.5 lineage in *M. occidentalis*** Cross-species reporter construct assays reveal multiple cases of cis-regulatory unintelligibility. (A) *C. intestinalis* embryo electroporated with Ciinte.Tbx6-r.b>GFP reporter construct [9], which drives GFP expression in tail muscles and the B7.5 lineage cells (arrowheads), recapitulating endogenous Ciinte.Tbx6-r.b expression. (B) *M. occidentalis* embryo electroporated with Moocci.Tbx6-r.b>GFP reporter construct, which recapitulates expression in tail muscle cells including B7.5 lineage cells (arrowheads). (C) *C. intestinalis* embryo electroporated with Moocci.Tbx6-r.b>GFP, which drives expression in B-line tail muscle and mesenchyme cells but is excluded from the B7.5 lineage. (D) C. intestinalis embryo electroporated with Ciinte.Hand-r>H2B::GFP reporter (Davidson and Levine, 2003), which drives H2B::GFP expression in anterior endoderm, A7.6 lineage, and TVCs (arrowheads), recapitulating endogenous Ciinte.Hand-r expression. (E) *M. occidentalis* embryo electroporated with Moccci.Hand-r>H2B::GFP construct, which recapitulates the same expression pattern. (F) *C. intestinalis* embryo electroporated with Moocci.Hand-r>H2B::GFP, which drives expression in endoderm and A7.6 lineage, but not in B7.5. (G) *M. occidentalis* embryo electroporated with Ciinte.Ebf neuron-specific YFP (green) and H2B::mCherry (red) reporter constructs electroporated [114], which drive very weak expression in a limited subset of motor ganglion neurons (green and red).

Figure 4.4 (cont'd): (H) *C. intestinalis* embryo electroporated with a Moocci.Ebf>YFP reporter, which drives robust reporter gene expression in several brain, motor ganglion, and nerve cord neurons. (I) *C. intestinalis* embryos electroporated with Ciinte.Sox1/2/3 (left) and Moocci.Sox1/2/3 (right) H2B::mCherry reporter constructs, both of which recapitulate Sox1/2/3 expression in ectoderm. Panels A-F are lateral views of tailbud embryos, panels G is a dorsal view of a tailbud embryo, panel H is a dorso-lateral view of a hatched larva, and panel I shows vegetal views of mid-gastrula stage embryos. Anterior is to the right, except in Panel I, in which anterior is to the top. DOI: 10.7554/eLife.03728.023

section of the tail and of the epidermal cells at the tail tip [37]. We observed the absence of *hox6* in all three *Molgula* species, which is not surprising, seeing that *hox6* is also missing in *O. dioica* and no expression is detected in *C. intestinalis* through whole mount in situ hybridization at any stage of development [38, 105]. No two of the *Molgula* species show the same *hox* cluster and all show a strong divergence outside of coding regions, even more so in *M. occidentalis.* There is a duplication of *M. occidentalis hox10* which could lead to a split in function since *Ciona Hox10* is expressed in two region during the mid-tailbud stage—a small region of the anterior nerve cord, and a small area of the posterior ventral endoderm and adjacent tissue [38]. It is proposed that ascidians evolved their simple body plans and rapid embryogenesis through extensive genomic rearrangement and gene loss, with specific loss of several *hox* genes [36]. *Hox12-13* are not found on the same contig in *M. occulta*, however when aligned with mVista there appears to be a strong case for synteny (Figure 1.1), so it is possible they are clustered, but the contigs are not joined because the *M. occulta* genome assembly is too fragmented. *Hox* function does not have the same level and so far their has not been a consistence between *hox* clusters in no Tunicate studied, showing that *hox* is tends do be divergent within the tunicates.

In *Ciona dll* stains in the endodermal strand cell during early-tailbud embryos, showing positive signals in two cells of the endodermal strand [8] which derives from the primary muscle or mesenchyme lineage [96], while in Drosophila embryos *dll* is required for the gene pathway for limb formation in the thoracic segment [130]. Because of the lack of homology outside of coding regions, we were able to identify *distal-less* (*dll*) downstream of *hox13* in both *M. occidentalis* and *M. oculata*, but we did not find it in *M. occulta*. Further investigation is needed, as this could potentially be tied to failed muscle cell differentiation in *M. occulta*.

From our initial survey of a handful of enhancers from *C. intestinalis*, *M. occidentalis*, and *M. oculata*, we encountered several instances of either mutual unintelligibility, or asymmetric intelligibility of enhancers. These results add to the mounting evidence suggesting that acute and pervasive DSD may have occurred over the course of ascidian evolution, obfuscated by the identical cell lineages and highly conserved gene expression patterns of ascidian embryos [118]. The multiple examples of cis-regulatory unintelligibility we identified were rather unexpected given (a) the extremely conserved pattern of expression of orthologous genes from Molgula and Ciona [124] and (b) previous observations of mutual intelligibility of enhancers between C. intestinalis and H. roretzi (e.g., Otx), and between C. intestinalis and the more closely related C. savignyi [44, 4]. Large-scale, quantitative cross-species assays and detailed GRN studies will illuminate factors that may contribute to conservation or divergence in regulatory mechanisms.

4.7 Conclusion

Hox genes function is not conserved within the tunicates, which have also undergone substantial *hox* gene loss. All central *hox* genes are missing in all three *Molgula* species. Studying these three molgula species have compounded the evidence that Tunicate has lost both *hox7* and *hox8* after diverging from the vertebrate and cephalochordate lineages, while ascidians have lost *hox9* and *hox11* after diverging from the larvaceans. It appears because *hox* does not act as an anterior-posterior organizer in ascidians, *hox* genes are more easily lost in ascidians and other genes have taken over this role.

Although tunicates have a well defined and invariant cleavage, the mechanism behind development is not always the same from species to species. We speculate that a high frequency of compensatory changes, required for the rapidly evolving ascidians to accommodate the constraints imposed by their invariant embryonic cell lineages and highly compact genomes, has given rise to a preponderance of cross-species cis-regulatory unintelligibility, following the DSD/SPC model. This perfect storm of intrinsic factors may be the key to explaining the dichotomy observed between highly conserved embryos and divergent cis-regulatory structure/function in ascidians.

Chapter 5

Differential expression analysis of tail loss in an invertebrate chordate

5.1 Introduction

Chordates are composed of three subphyla—vertebra, tunicata and cephalochordata—that all share several characteristics, with the notochord being the key characteristic (from which the phylum name comes). The tail development of larvaceans such as *Oikopleura dioica* and several species of ascidians and tunicates has been well studied [39, 79, 53]. Larvaceans form tailed larvae with a hollow dorsal notochord and keep their tail throughout their adult life stage. Ascidians generally form their tail in a similar manner before undergoing the process of metamorphosis, in which the larval tail is absorbed in to the trunk region [88]. A typical ascidian larvae tail forms through the convergence, intercalation and extension of the notochord and the differentiation of the posterior muscle cells [117]. When fully formed the ascidian notochord contains 40 cells, flanked by three rows of muscle cells. The ancestral notochord or notochord-like structure is believed to have been muscle based, which is perhaps the reason that the tail formation is often contingent upon proper development of both notochord and muscles [57]. This idea is supported by the observation that the primary and secondary notochord and muscle lineage are derived from the same blastomere, and the ascidian tail needs both the notochord and differentiated muscle to form a larval tail [82, 16].

Of the ~3000 species of ascidians fewer than 20 do not form a tail, with the majority being Molgulidae species [2, 34]. This likely represents several independent instances of evolutionary loss of the tail, and introduces the question of why are the Molgulidae so susceptible to tail-loss. Although the mechanism behind tail-loss differs by species, a common characteristic is the lack of a notochord that intercalates and extends, as well as a less differentiated central nervous system (CNS) structures and tail muscles [117]. *Molgula bleizi* notochord cells converge to the midline, and began to extend, but cells never properly intercalate and the tail formation stops before it is fully formed [42]. In *M. bleizi* there is an early down-regulation of *brachyury* (*bra*)—a key notochord inducer—and larva-specific muscle actin genes have become pseudo genes [42]. A similar situation is observed in *Molgula occulta*: muscle actins have become pseudogenes, independently of their conversion in *M. bleizi*.

M. occulta and *Molgula oculata* are two closely related species, who in their adult form are virtually identical, with the exception of a white pigment spot between the two siphons of the tailed species, *M. oculata*. During development the species are indistinguishable up to the gastrula stage. In late gastrula when the notochord and muscle progenitor begin to move posteriorly [123] and the notochord begins to form, the morphological divergence becomes evident. Through subtractive screening Swalla and Jeffery [121] identified *manx*, a zinc finger transcription factor (TF) and the cytoskeletal protein p58 as down-regulated in *M. occulta* relative to *M. oculata* [119], and this was shown to be one of the causes of the tail loss in *M. occulta*. There are several steps that take place to form the notochord and tail: first, the notochord cells move mediolaterally to the midline; and next, the cells polarize and intercalate, changing their shape and extending posteriorly [49, 43, 111]. This process is known as convergence and extension. Although the two species have different developmental programs, crossing the tail-less *M. occulta* eggs with sperm of the tailed *M. oculata* produces a hybrid with 20 notochord cells like the tail-less species, but in which the cells intercalate and extend like the tailed species. In the hybrids the expression of *manx* and p58 are restored, and antisense phosphorothiated oligodeoxynucleotide *manx* in the hybrids have shown that zygotic *manx* is necessary for tail formation [121].

Several key tail development genes have been identified as present in the *M. occulta* genome, but expressed at low levels during embryogenesis, and when expressed, some of these genes were shown to restore features in the hybrid [120, 39, 123, 121]. With advances in high throughput sequencing technologies, gene expression of *M. occulta*, *M. oculata*, and hybrid species can be analyzed on a whole transcriptome level [22, 91]. mRNA of three different developmental stages for *M. occulta*, *M. oculata*, and their hybrid has been sequenced and assembled at Michigan State University (MSU). These three transcriptomes were used to assess the expression levels of known notochord genes downstream of *bra* using *C. intestinalis* data from the NCBI database (`ncbi.nlm.nih.gov`). BLAST searches were done against known notochord genes, and several of them were selected for further analysis, including *FGF9/16/20*, *prickle (pk)*, and several other downstream genes. These genes were then used to construct a putative *brachyury* gene regulatory network for both *M. occulta* and *M. oculata.*

5.2 Methods

5.2.1 Sample collection, sequencing and assembly

RNA was extracted from all three *Molgula* species using the methods discussed in Lowe et al. [66]. RNAs for the gastrula (3hpf), neurula (4hpf) and mid-tailbud (6hpf) stages were extracted for both *M. occulta* and *M. oculata*, with a replicate for the gastrula and a sample for early-tailbud stage sequenced from *M. occulta*. DNA was extracted from the gonads of an individual adult specimen for *M. occulta*, and *M. oculata*. Two paired-end jumping libraries were collected for each species ranging from ~300bp to ~950bp. Further details about extraction methods and libraries can be found in Stolfi et al., [115]. Sequencing for *M. occulta* and *M. oculata* RNA were conducted at the Michigan State University, while all other sequencing was done at New York University. All libraries were paired-end, with 75 base pair (bp) reads for the sequencing done at MSU and 100 bp reads for the NYU sequencing.

Genome assemblies were conducted using 3-pass digital normalization [5] and assembled using Velvet[141]. Other assemblers were tested, however, Velvet produced the best results with the least fragmented assemblies. Assemblies were initially done with $21 \geq k \geq 71$, for intervals of 10. We selected the 'k' value with the highest N50, and then the genomes were reassembled for a k±10 with a step size of 2 for the selected assembly, and the best N50 was chosen. A k of 31, and 49 were select for *M. occulta*, and *M. oculata*, respectively.

Both *de novo* and reference based assemblies were used to create gene models. Reads were mapped to their respective genomes using bowtie2 and tophat to identify genes and alternative splicing variants [56, 127]. The accepted.bam files were then sorted and indexed using samtools [61]. The sorted bam files were then processed using cufflinks and cuffmerge to generated consensus gtf annotation files. The digitally normalized trinity *de novo* assembled transcripts from Lowe et al. [66], were aligned to their respective genomes using BLAT [23]. The cufflinks/cuffmerged gtf files were then converted into bed files and merged with the annotation files from the mapped *de novo* assembly aligned using gimme (`https://github.com/likit/gimme`). Gimme joins gene models using a graph based method to develop more complete transcripts. The gimme gene models were then converted to gff format using the script bed2gff in the gimme utils folder in order to extract the transcripts from the genome in a multi fast file. Transcripts were then extracted using “gffread -w transcripts.fa -g /path/to/genome.fa transcripts.gtf” which is included in the cufflinks package. The extracted transcripts were partitioned into transcript families and annotated using the khmer suite and steps found in the eel-pond protocol (`https://khmer-protocols.readthedocs.org/`). *Ciona intestinalis* was used as an annotation reference, and the sequences were retrieved as discussed in Chapter 3.

5.2.2 Gene counts and differential expression analysis

Reads were mapped to transcripts from the gimme gene models for their respective species. Reads from the hybrids were mapped onto both the *M. occulta* and *M. oculata* transcripts, since the hybrids are F1 hybrids and should contain an allele from each parent. Read counts were generated using eXpress [94]. eXpress gives the option of “Total counts” and “Effective counts”, which reports the number of reads mapped per transcript and the normalized counts based on transcript length, respectively. Because EdgeR uses unnormalized reads, “Total counts” were used. Counts for hybrid reads mapped to *M. occulta* and *M. oculata* were combined to calculate total expression at a given stage. A replicate was only available for one of the samples, 3hpf, and because of this 5hpf was treated as a replicate for 6hpf. These time points correspond to early and mid-tail bud stages in the tailed ascidian.

Differential expression was calculated using the Bioconductor EdgeR package because of its ability to work with minimal replicates [95]. We ran RBH for *M. occulta* and *M. oculata* in order to identify orthologous transcripts to conduct allele specific differential expression analysis on the hybrid. We used both estimateGLMCommonDisp and estimateGLMTagwiseDisp to calculate dispersions. There is only a replicate for *M. occulta* gastrula stage (F+3), because of this we used *M. occulta* early tailbud stage as a replicate for the mid tailbud stage. These replicates were used to calculate dispersions for all samples. Exact-test with p = 0.05 was used to determine differential expression. Transcripts with a false discovery rate (FDR) of 0.05 were called as differentially expressed.

5.3 Results

5.3.1 *M. occulta* and *M. oculata* have strong overlap in gene expression

C. intestinalis is the closest ascidian species with a well annotated genome, and so we used *Ciona* proteins obtained from the NCBI to annotate the genomes of both *M. occulta* and *M. oculata*. Reciprocal best hit (RBH) blast with an e-value of 1e-3 was done with the *M. occulta* and *M. oculata* transcriptomes against *C. intestinalis* for the annotation of both *Molgula* species. We are aware this is a low threshold for homology, however, information for these species are not known and we wanted to gain as much insight as possible for genes and gene function. Moreover, the reciprocal best hit criterion is extremely stringent.

The gene models for *M. occulta* and *M. oculata* produced 42,365 and 40,775 sequences total, respectively. Precisely 8,627 *M. occulta* transcripts were annotated as orthologs and 22,700 transcripts were annotated as showing homology to *C. intestinalis*. Similar annotation numbers were produced with *M. oculata*: 8,677 showed orthology, and 22,583 showed homology. *M. occulta* and *M. oculata* have a high overlap in number of translated transcripts that showed any level of homology with *C. intestinalis* proteins from the NCBI database.

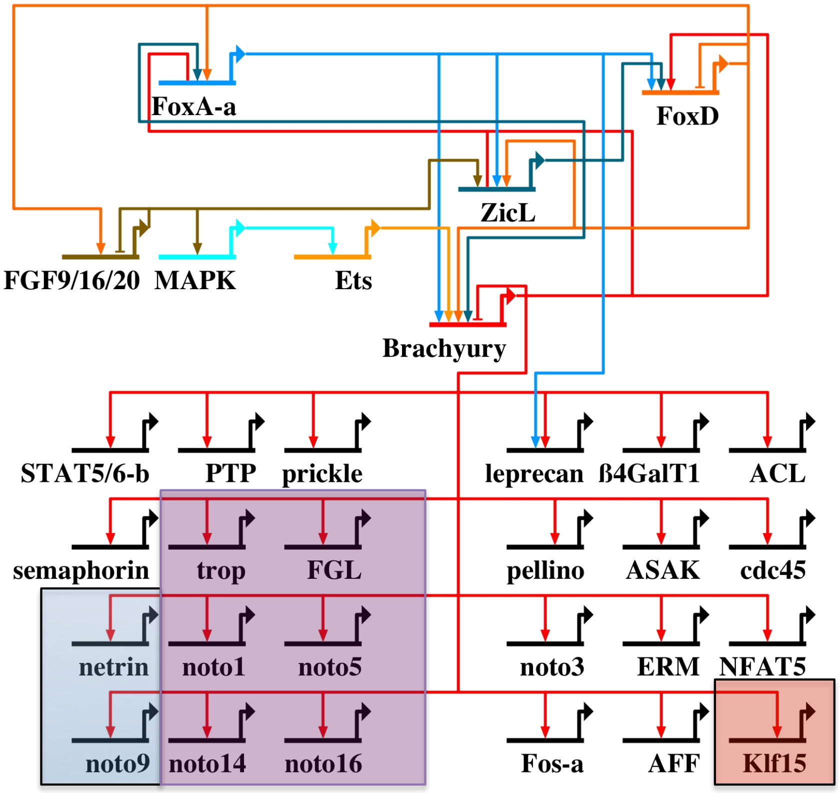


Figure 5.1: ***Brachyury gene regulatory network.*** *Bra* is a key notochord inducer, and without its expression neither the notochord nor the chordate tail forms. Downstream genes have been identified in various studies and of these 67 genes 11 were missing from the transcriptomes of both species. The missing genes were *cofilin*, *entactin* (*nidogen-2-like*), *fibrinogen-like protein* (*FGL*), *fibronectin*, *multidom*, *noto1*, *noto5*, *noto14*, *noto16*, and *tropomyosin*. *netrin*, and *noto9* are missing in *M. occulta* and *Klf15* is missing in *M. oculata*. (Blue) missing in *M. occulta*, (Red) missing in *M. oculata*, and (Purple) missing in both species. This GRN was built from previous studies [31, 30, 32, 54, 53], using the BioTapestry software [65].

Of the 16,414 *Ciona* proteins, *M. occulta* had BLAST hits for 83.6% and 86.5% had hits in *M. oculata*. *M. occulta* had hits for 453 proteins that were not found in *M. oculata* and *M. oculata* had hits for 921 transcripts that did not have hits in *M. occulta*, yielding an overlap of 97% in assembled homologs.

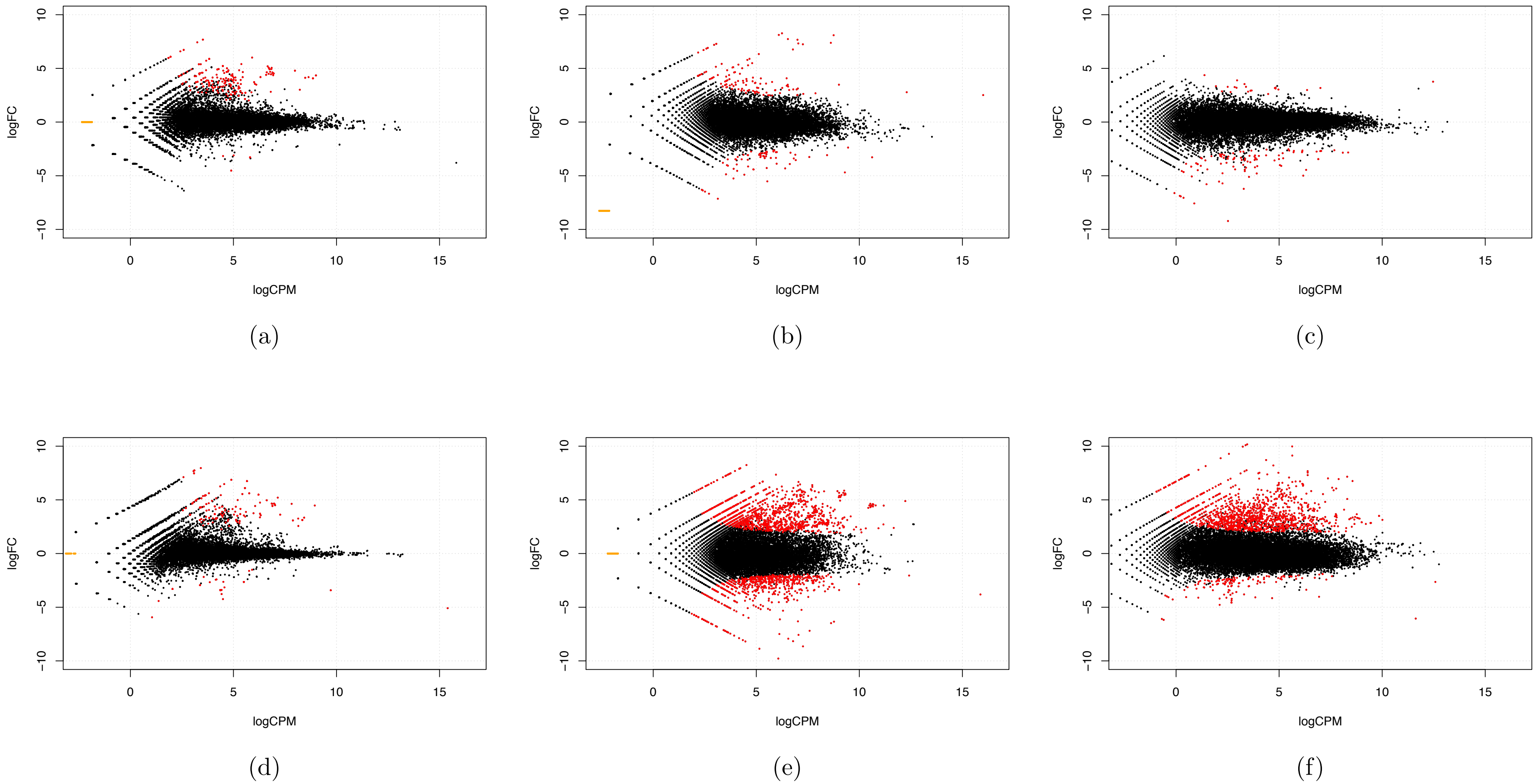
5.3.2 Notochord gene network

Next, we examined genes associated with notochord development in *C. intestinalis* to investigate the molecular development of the tail. Gene candidates were identified as being involved in tail formation and notochord development through previous analysis of genes downstream of *bra* [31, 30, 32, 54, 53]. From these studies many potential notochord genes were identified; we compiled a list of 67 genes and identified those expressed during the gastrula, neurula or tailbud stages within the transcriptomes of both *M. occulta* and *M. oculata* (Figure 5.1). Of the 67 genes, 11 were not expressed in the transcriptomes of either species: *cofilin*, *entactin* (*nidogen-2-like*), *fibrinogen-like protein* (*FGL*), *fibronectin*, *multidom*, *noto1*, *noto5*, *noto14*, *noto16*, and *tropomyosin.* The remaining genes without orthologous sequences were *netrin*, and *noto9* in *M. occulta* and *Klf15* in *M. oculata.* In *Ciona*, *netrin* is expressed in the notochord and the central nervous system and is associated with axon guidance [30]. *Klf15* was detected in the notochord, but there is currently no known information regarding its function [90]. Taken together, this demonstrates a strong overlap in the presence of the number of genes associated in notochord formation in both *M. occulta* and *M. oculata.*

5.3.3 Differential expression between neurula and tailbud appears to be key factor in tail development

Species	Condition	Number of transcripts that show...		
		Up-regulation	Down-regulation	No differential expression
M. occulta	from Gastrula to Neurula	260	8	20197
	from Neurula to Tailbud	1	4	20460
M. oculata	from Gastrula to Neurula	119	66	20280
	from Neurula to Tailbud	1170	626	18669
Hybrid	from Gastrula to Neurula	21	99	20345
	from Neurula to Tailbud	1270	129	19066

Table 5.1: Differential expression: Species *vs* time



(a)

(b)

(c)

(d)

(e)

(f)

Figure 5.2: **Differential expression of homologous transcripts.** Differential expression in *M. occulta* for (a) gastrula vs neurula, and (d) neurula vs tailbud, *M. oculata* for (b) gastrula vs neurula, and (e) neurula vs tailbud, and the hybrid for (c) gastrula vs neurula, and (f) neurula vs tailbud. Genes in red are differentially expressed with a FDR <0.05

Only *M. occulta* embryos that expressed *manx* and p58 produced hybrids with urodele features [39, 121]. To identify other candidate genes whose differential expression in *M. occulta* may contribute to the tail-less condition, we sequenced and assembled three developmental stages—gastrula, neurula, and tailbud—across *M. occulta*, *M. oculata* and their interspecies hybrid. The tail-less species, *M. occulta*, showed the highest level of differential expression (FDR <0.05), with 260 (97%) of the identified differentially expressed transcripts up-regulated in the neurula stage relative to the gastrula stage. (Table 5.1). *M. oculata* also had more transcripts up-regulated (65%) than down-regulated at the neurula stage relative to the gastrula. Hybrids did not follow this trend; the majority of differentially-expressed hybrid genes are down-regulated at this stage (82%).

When comparing the tailbud stage to the neurula stage there was essentially no significant differential expression observed in *M. occulta.* A total of 5 genes were said to be differentially expressed (FDR <0.05). However a major shift in differential expression was seen in both *M. oculata* and the hybrid at the tailbud stage relative to the neurula stage (Figure 5.2). There were 1170 and 1270 transcripts up-regulated, and transcripts and 129 transcripts down-regulated in *M. oculata* and the hybrid respectively. That equates to a 10x increase in differentially expressed transcripts in both hybrid and *M. oculata* tailbud embryos relative to the respective neurula stage embryos.

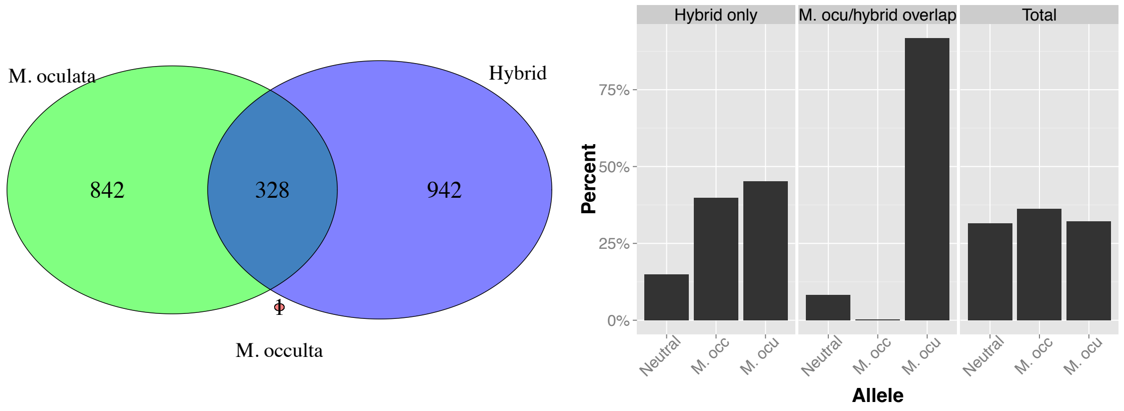
5.3.4 Overlap between hybrid and M. oculata alleles

When comparing transcript expression from neurula to tailbud in *M. oculata* and the hybrid there is a total of 2,440 transcripts up-regulated collectively. Of these 2,440 transcripts, 328 (13%) overlap (Figure 5.3a). There were no transcripts in *M. occulta* identified as differentially expressed that overlapped with either *M. oculata* or their hybrid. Of the transcripts that were down regulated, there was only an overlap of 5 transcripts (0.06%) between *M. oculata* and the hybrid, and again none with *M. occulta*. We further examined their allele specific expression to determine if the *M. oculata* alleles represented the majority of this gene expression in the hybrids, or if there was some rescue of expression of the *M. occulta* alleles. When analyzing the 328 up-regulated transcripts that overlapped between *M. oculata* and the hybrid, there was large skew for expression from the *M. oculata* allele: 91.7% of expression came from *M. oculata* (Figure 5.3b). A similar but far less dramatic trend was observed for the transcripts up-regulated in hybrid but not overlapping with *M. oculata*, with more expression coming from the *M. oculata* allele. The allele specific differential expression at tailbud was 32% *M. oculata*, 36% *M. occulta* and 32% hybrid.

5.4 Discussion

Several genes identified as expressed in the notochord of *C. intestinalis* [30, 32] were observed to be missing in both *M. occulta* and *M. oculata*; of these genes all were also identified as being missing in the *O. dioica* [53]. The lack of expression of these gene is the Molgulids and *O. dioica* implies these genes are not necessary for the development of a fully functional notochord, CNS, or muscles in the ancestral chordate.

Here we present a differential expression analysis of two closely related *Molgula* species, *M. occulta* and *M. oculata*, using high throughput sequencing technology. We were able to create gene models using both a mapping based approach and *de novo* assembly approach, and then combined the assemblies for better transcript models. We showed that from neurula to the tailbud stage there is a 10 fold increase in transcripts that are identified as differen-



(a)

(b)

Figure 5.3: **Upregulated transcripts overlap between hybrid and *M. oculata*.** When comparing the gastrula and neurula time point for *M. occulta*, *M. oculata*, and their hybrid, both *M. oculata* and the hybrid showed differential expression in at least 7% of their transcripts with the majority being up-regulated. (5.3a) There is a 15% overlap in overexpressed genes *M. oculata* and the hybrid, for the transcripts overexpressed when comparing gastrula to neural expression, and there is no overlap with *M. occulta*. (5.3b) When looking at the allelic express for the same condition, there is a strong skew in expression coming from the tailed allele for the unregulated transcripts that overlap between *M. oculata* and the hybrid. The highest percentage of transcripts of allelic expression in the hybrid also comes from *M. oculata* for up-regulated transcripts that do not overlap with *M. oculata*, but not to as great an effect. For genes' overall allelic expression, the majority comes from *M. occulta*, 36%, with 32% coming from *M. oculata*.

tially expressed (p=0.05) in both *M. oculata* and the hybrid. Using this same condition we were able to show that there is almost no differential expression in the the tail-less *M. occulta*. This result correlate with the embryo morphologies during this transition; in *M. oculata*, the processes of tail formation are occurring, which may explain the increased transcriptional activity. In contrast in *M. occulta* there are no noticeable morphological changes occurring. In hybrid embryos, this dynamic regulation of gene expression at the tailbud stage is rescued. Because hybrids show partially rescued tails and CNS structures, I propose that these transcripts may have an important role in the formation of the tail. Furthermore, of those transcripts upregulated at the tailbud stage in the hybrid but not in *M. occulta*, it appears that expression is being restored from the *M. oculata* allele. This suggests that the relative lack of differential gene expression in the neurula-to-tailbud transition in *M. occulta* may be due to loss-of-function of cis-regulatory elements controlling the expression of key genes involved in tail and CNS formation.

5.5 Conclusion

We have used two closely related Molgulids with the ability to hybridize to study three key time points in tail development: gastrulation, neurulation and tailbud. This study has shown that from neurula to tailbud a number of transcripts are up-regulated, in both *M. oculata* and the hybrid. Not only are transcripts up-regulated in this condition, with no significant differential expression in the tail-less species, but there is also an overlap between the differentially expressed genes in the *M. oculata* and the hybrid. Since both *M. occulta* and *M. oculata* have a strong overlap in genes implicated in notochord development, this differential expression could be the possible cause of loss of the tail and notochord and CNS in *M. occulta*.

The hybrids shed light on the fact that cis-regulatory elements are one of the key causes for the lack of urodele features in the tail-less *M. occulta*. Previously, it has been shown that *Molgula* have a high turnover in sequences of cis-regulatory elements, diverging in specific binding sites, while conserving orthologous gene expression [115]. Perhaps, this cis-regulatory turnover has led to the Molgulidae becoming more susceptible to the deactivation of binding sites, which in turn contributes to the loss of tails in several species.

Chapter 6

Conclusions and Discussion

The chordate body plan is conserved throughout the phyla with only a few species deviating from the conserved body plan; at some point in their life, most chordates develop a tadpole larvae containing a hollow dorsal neural tube, and a postanal tail containing a notochord flanked by muscle cells. However, it has been documented that 16 species out of the approximately 3000 species of tunicates—one of the three subphyla of chordates—have independently undergone tail loss, with the majority being within the molgulids [2, 34]. Because they have both tailed and tail-less species, molgulids are useful models for studying changes in the development of body plans. Having closely related species with altered development allows us to look at the short-term modifications that occur during the evolution of alternate body plans, showing us that larval development can evolve rapidly. Of particular interest, two molgulids have the ability to hybridize, which allows us to examine the mechanisms of evolutionary change at an allele-specific level.

6.1 Evaluating a lightweight transcriptome assembly pipeline on two closely related ascidian species

The drop in sequencing price has aided in our efforts to understand body plan development by allowing us to quickly and inexpensively obtain transcriptome and genome sequences, along with temporal expression profiles. However, for non-model systems with no existing genomic sequence, we must first assemble the transcriptomes and genomes. The methodology for assembly is not unambiguous: there are a number of different steps—quality trimming, filtering, and choice of assembler(s), with several programs at each step and no clear choice. Studies have been done to compare assemblers, but no one assembler is clearly the best choice [10]. Many times assembly methods are chosen on the usability of software and the availability of resources.

Factors that are most limiting for assemblies are memory, and in some cases time. The redundant nature of sequencing data allows for the removal of redundant reads with approaches like digital normalization, but at what cost? Here in efforts to assemble the first two *molgula* transcriptomes, we evaluate the cost of filtering sequencing reads for redundancy. We show that abundance filtering quality trimmed assembly reads enables transcriptome assembly with a reduction in both memory requirement and assembly time, while retaining essentially the same information content (e.g. number of genes, average gene length, and homology to the closest sequence species). We have demonstrated that the Oases and Trinity assemblers return similar results, both suitable for downstream analysis using the full or redundancy filtered dataset. Our pipelines are available so we also provide methodology to be used by future researchers. In addition assembly metrics to evaluate assemblies are also an important step in the protocol. One of the standard metrics for evaluating assemblies is the N50, but this is designed for genomic evaluation and does not clearly translate to transcriptomes because of isoforms. For example, the same exon may be included in multiple splice variants, inflating the total bp assembled; some assemblers may choose to report more isoforms than others even with the same read support; and "total length" makes little sense for transcriptomes. In contrast, homology and programs such as CEGMA are more informative and useful for downstream analysis, because they measure the information in the transcriptome and cover the annotation step as well. When measuring recovered homologies and CEGMA measurements we find that both assemblers and both treatments compare well.

6.2 Change in gene function and cis-regulatory binding sites

In tunicates *hox* genes are not important for patterning along the anterior-posterior axis, as they are in vertebrate and other bilaterians [17, 70, 37]. This limited function in the *hox* genes is evident from the lack of noticeable phenotypic defects in the knockdown of *C. intestinalis hox* genes, with the exception of *hox10* and *hox12*. *Hox10* is involved in neuronal development and *hox12* is involved in the formation of the posterior most portion of the the tunicate larval tail [37]. In addition, tunicates have undergone loss and rearrangement within the *hox* clusters, unlike most animals that have been studied [36]. There have yet to be two tunicate species found with the same *hox* cluster configuration - there are typically changes in ordering, duplications, and which hox genes are present (figure 4.2). Only *C. intestinalis* has enough scaffolding to completely identify the structure and intergenic spacing between the *hox* genes. However, of the *hox* gene we found on the same scaffolds in *M. occulta*, *M. oculata* and *M. occidentalis* the average intergenic spacing is far smaller than in *Ciona*, ranging from 10-25 kb in length. Without additionally scaffolding we cannot fully examine the rearrangements within the *hox* clusters of ascidians based on the molgulid sequences. Perhaps soon after diverging from vertebrae and cephalochordates, the *hox* cluster genes were co-opted into other GRNs, leading to higher lability than in species where they retained their stereotyped roles.

We have also identified a number of enhancers between *C. intestinalis*, *M. occidentalis*, and *M. oculata* that were either mutually unintelligible, or asymmetrically intelligible. We propose that a high frequency of compensatory changes are required for the rapidly evolving ascidians to maintain their invariant embryonic cell lineages in the presence of their compact genomes. These compensatory changes have caused the enhancers to diverge within the ascidians while allowing them to develop with their typical body plan.

6.3 Differential expression analysis of tail loss in an invertebrate chordate

Studying closely related organisms gives us insight into the underlying evolutionary mechanisms of divergence and underlying development. Indirect and direct developing sea urchins have been studied showing that change in axial cleavage patterns and early cell fate leads to divergent larval body plans that can exhibit similar adult body plans [138, 27]. In contrast, ascidians retain their typical cell division and invariant cell fate in direct, indirect, tailed and tail-less ascidians [41, 69].

We have shown that evolutionary drift has occurred in the cis-regulatory modules of developing ascidian embryos between *C. intestinalis* and *M. occidentalis*, and these enhancers are not always capable of driving expression in other ascidian species. It appears that the same has happened between the tailed *M. oculata* and tail-less *M. occulta*, however those cis-mechanisms are restored in the (*M. oculata x M. occulta*) interspecific hybrids, probably allowing transcription factors to bind their targets at a higher affinity and restoring the necessary level of expression to develop the urodele features. Further research is needed to identify and test cis-elements, but as of now we have a better understanding of possible mechanisms of the restoration of the larval tail in hybrids, which we can at least hypothesize is driven by recovery of cis-regulatory modules.

6.4 Conclusion

The identification of regulatory genes expressed differently in the urodele and anural species is needed to understand the molecular mechanism underlying the evolutionary transition from urodele to anural development. When examining the tailed and tail-less *Molgula*—*M. oculata* and *M. occulta*—there is a strong overlap (91%) in the expression of genes that can be annotated using the *C. intestinalis* proteome. There is also a strong overlap in the expression of genes associated with notochord development, which is the key structure-defining element not present in *M. occulta*. We observed that it was not the lack of genes in the transcriptome, but the expression of said genes at a sufficient level, which led to the tail-less phenotype. One of the shortcomings of our study is the lack of replicates from all stages; ideally we would like at least three per developmental stage for each organism and because of this there are likely to be more false positives and false negatives in our analysis than there would be with more replicates. However, we were able to identify differentially expressed genes using the available replicates to estimate the dispersion for each of the samples, and able to identify a subset of genes involved the development of the urodele features by examining the correlation between two independent datasets, the tailed *M. oculata* and the hybrid. For further examination replicates are needed, along with addition sequencing of other tailed and tail-less molgula genomes and transcriptomes. This would give us stronger insights into why the molgula are able to easily undergo tail-loss and the toolkit for molgula tail development.

Here we were able to see in the hybrids that the restored expression came from the tailed allele and we proposed that cis-regulatory modules are the cause of this. The restored expression of genes involved with the development of urodele features are acting in an allele specific manner, while expression for other genes (total) has a more balanced expression from each allele. The tailed allele restored the enhancer binding sites; presumably TF were able to bind at a higher affinity and restore necessary level of expression to generate urodele features, including the formation of the notochord and hence the larval tail. To confirm this hypothesis we would first perform a ChIP-seq analysis to identify direct binding targets, and correlate them with the differentially expressed genes. A clear starting point is the gene *bra* which is known to be important for notochord development, and *Tbx2/3* which is downstream of *bra*, and involved in convergence and extension [45]. *Tbx2/3* was also identified in our differential expression analysis. From this gene-targeted analysis we can examine the divergence between *M. occulta* and *M. oculata* enhancers, examine their binding affinities, and test the found sites using transgenesis to see if we can restore the *M. occulta*'s tail.

Our hypothesis depends heavily on the assumption that the gene regulatory network (GRN) is conserved within the tunicates. Comparing the GRN of distantly related organisms can identify the genes necessary to the development of similar phenotypes and body plans. Currently we plan to expand our analysis by comparing the *molgula* to the ten tunicate genomes assemblies found on the aniseed website (`http://www.aniseed.cnrs.fr/`), and available chordate genomes found on NCBI (`http://www.ncbi.nlm.nih.gov/`) and ensembl (`ensembl.org`). With the information we currently have we will look more deeply into the sequence divergence of both coding and non coding regions. To examine the noncoding regions additional genomic sequencing is needed for gap filling and scaffolding. However, it is also necessary to determine if the GRNs have conserve structure by e.g. confirming the spatial expression patterns for the network. Next steps on this project would be to identify relevant transcription factors by annotating the genes and then selecting differentially expressed genes associated with DNA binding functions. Once we identify our candidate TFs we would proceed with whole mount in situ hybridization to confirm conserved spatial patterning.

APPENDICES

Appendix A

Supplemental Figures

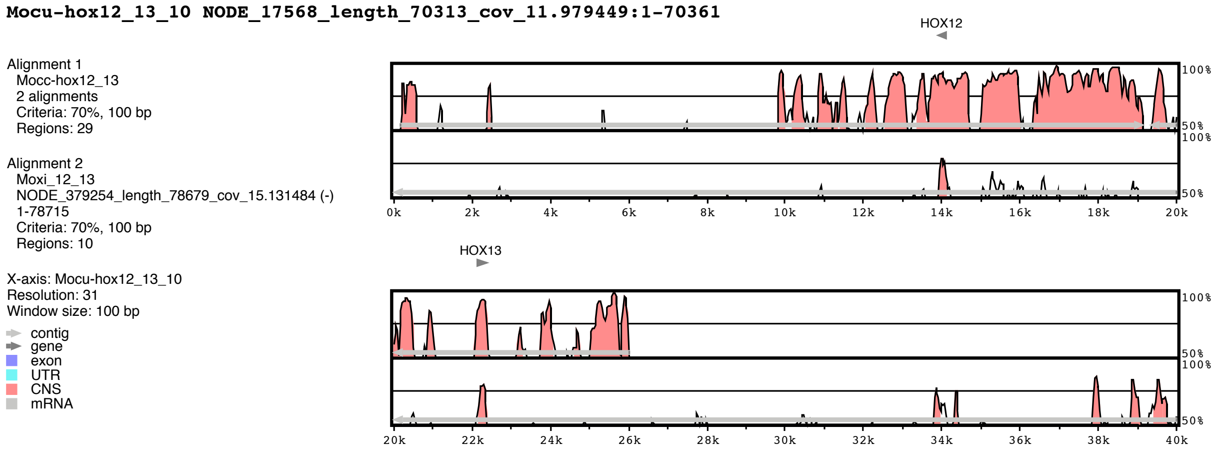
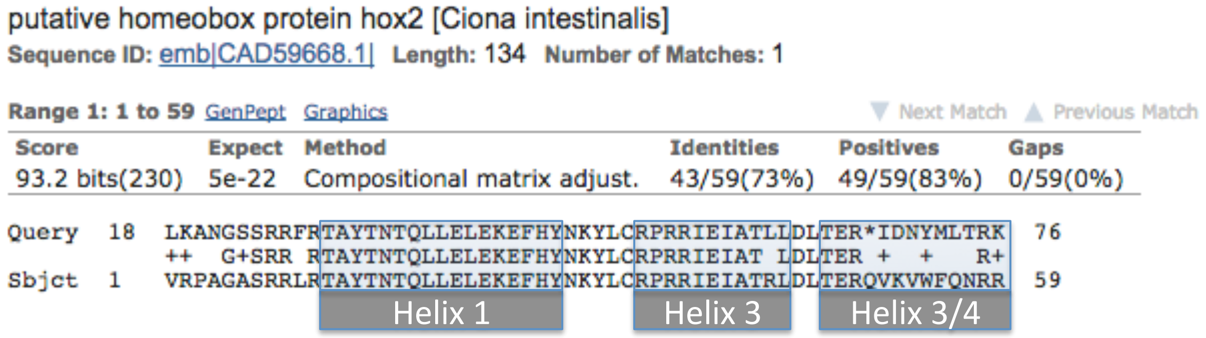


Figure A.1: **Alignment for *hox12-13* in *M. occulta*, *M. oculata* and *M. occidentalis*** The contig containing *hox12-13* for *M. occidentalis* and *M. oculata*, along with the two contigs containing *hox12* and *hox13* for *M. occulta*. *M. oculata* was used as the anchor sequence because it showed the most similarity between the three species. Outside of the coding regions and its flanking area, there is very little sequence similarity between the species, and *M. occidentalis* exclusively shows similarity in coding regions. Grey arrows show the direction of the contig.



putative homeobox protein hox2 [Ciona intestinalis]

Sequence ID: emb|CAD59668.1| Length: 134 Number of Matches: 1

Range 1: 1 to 59 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
93.2 bits(230)	5e-22	Compositional matrix adjust.	43/59(73%)	49/59(83%)	0/59(0%)

```
Query  18  LKANGSSRRFRTAYTNTQLLELEKEFHYNKYLCRPRRIEIATLLDLTER*IDNYMLTRK  76
           ++  G+SRR RTAYTNTQLLELEKEFHYNKYLCRPRRIEIAT LDLTER +  +   R+
Sbjct  1   VRPAGASRRLRTAYTNTQLLELEKEFHYNKYLCRPRRIEIATRLDLTERQVKVWFQNRR  59
```

Figure A.2: **Alignment of *M. occidentalis hox2* genes with *Ciona* show premature stop codon.** The *M. occidentalis hox2* gene has a stop codon in the 3' region, inside of the 3/4 helix. *hox2* knockdowns in *Ciona* did not show any phenotypic difference, so the function of *hox2* may not be important in *M. occidentalis*

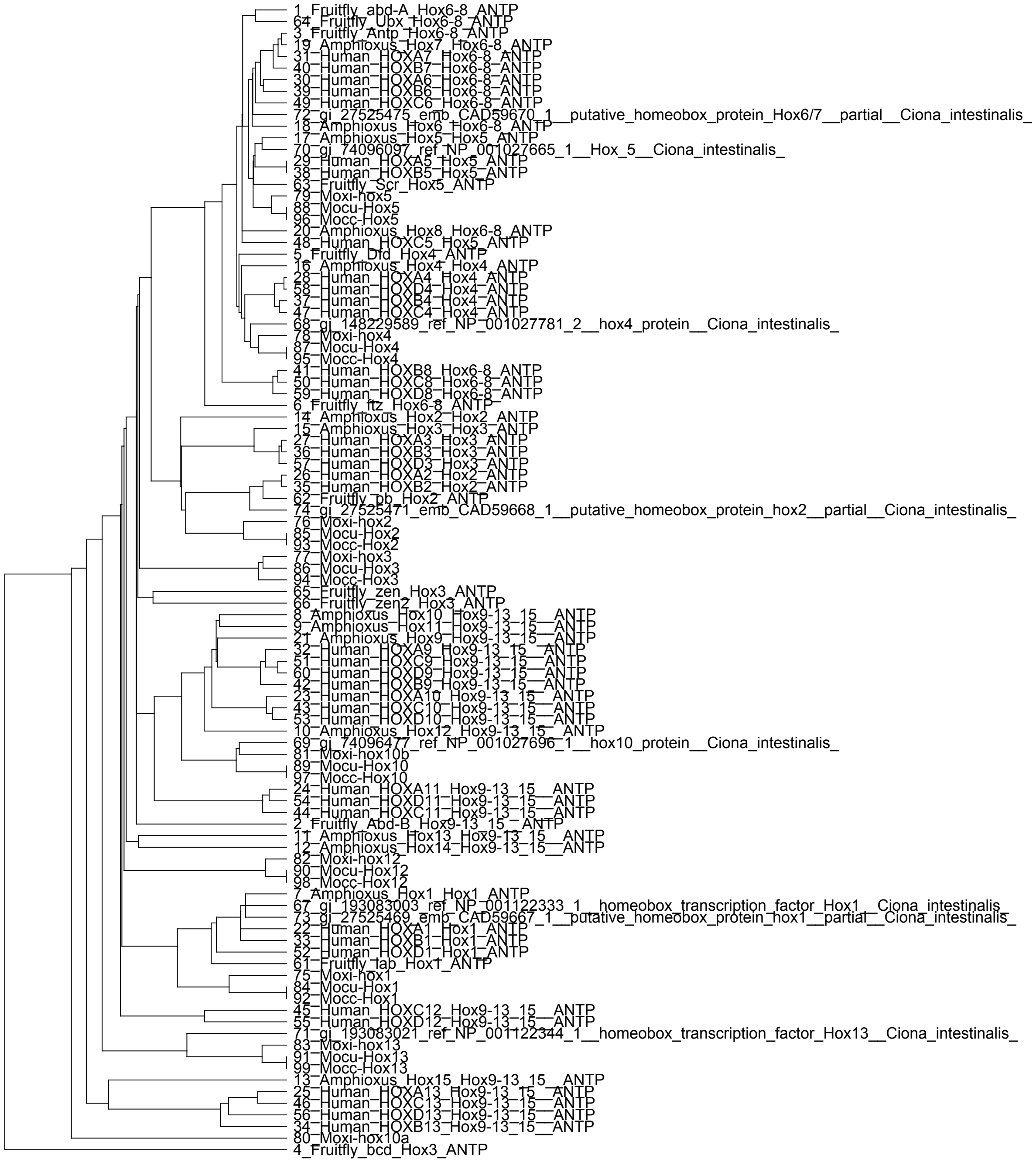

1_Fruitfly_abd-A_Hox6-8_ANTP
64_Fruitfly_Ubx_Hox6-8_ANTP
3_Fruitfly_Antp_Hox6-8_ANTP
19_Amphioxus_Hox7_Hox6-8_ANTP
31_Human_HOXA7_Hox6-8_ANTP
40_Human_HOXB7_Hox6-8_ANTP
30_Human_HOXA6_Hox6-8_ANTP
39_Human_HOXB6_Hox6-8_ANTP
49_Human_HOXC6_Hox6-8_ANTP
72_gi_27525475_emb_CAD59670_1__putative_homeobox_protein_Hox6/7__partial__Ciona_intestinalis_
18_Amphioxus_Hox6_Hox6-8_ANTP
17_Amphioxus_Hox5_Hox5_ANTP
70_gi_74096097_ref_NP_001027665_1__Hox_5__Ciona_intestinalis_
29_Human_HOXA5_Hox5_ANTP
38_Human_HOXB5_Hox5_ANTP
63_Fruitfly_Scr_Hox5_ANTP
79_Moxi-hox5
88_Mocu-Hox5
96_Mocc-Hox5
20_Amphioxus_Hox8_Hox6-8_ANTP
48_Human_HOXC5_Hox5_ANTP
5_Fruitfly_Dfd_Hox4_ANTP
16_Amphioxus_Hox4_Hox4_ANTP
28_Human_HOXA4_Hox4_ANTP
58_Human_HOXD4_Hox4_ANTP
37_Human_HOXB4_Hox4_ANTP
47_Human_HOXC4_Hox4_ANTP
68_gi_148229589_ref_NP_001027781_2__hox4_protein__Ciona_intestinalis_
78_Moxi-hox4
87_Mocu-Hox4
95_Mocc-Hox4
41_Human_HOXB8_Hox6-8_ANTP
50_Human_HOXC8_Hox6-8_ANTP
59_Human_HOXD8_Hox6-8_ANTP
6_Fruitfly_ftz_Hox6-8_ANTP
14_Amphioxus_Hox2_Hox2_ANTP
15_Amphioxus_Hox3_Hox3_ANTP
27_Human_HOXA3_Hox3_ANTP
36_Human_HOXB3_Hox3_ANTP
57_Human_HOXD3_Hox3_ANTP
26_Human_HOXA2_Hox2_ANTP
35_Human_HOXB2_Hox2_ANTP
62_Fruitfly_pb_Hox2_ANTP
74_gi_27525471_emb_CAD59668_1__putative_homeobox_protein_hox2__partial__Ciona_intestinalis_
76_Moxi-hox2
85_Mocu-Hox2
93_Mocc-Hox2
77_Moxi-hox3
86_Mocu-Hox3
94_Mocc-Hox3
65_Fruitfly_zen_Hox3_ANTP
66_Fruitfly_zen2_Hox3_ANTP
8_Amphioxus_Hox10_Hox9-13_15__ANTP
9_Amphioxus_Hox11_Hox9-13_15__ANTP
21_Amphioxus_Hox9_Hox9-13_15__ANTP
32_Human_HOXA9_Hox9-13_15__ANTP
51_Human_HOXC9_Hox9-13_15__ANTP
60_Human_HOXD9_Hox9-13_15__ANTP
42_Human_HOXB9_Hox9-13_15__ANTP
23_Human_HOXA10_Hox9-13_15__ANTP
43_Human_HOXC10_Hox9-13_15__ANTP
53_Human_HOXD10_Hox9-13_15__ANTP
10_Amphioxus_Hox12_Hox9-13_15__ANTP
69_gi_74096477_ref_NP_001027696_1__hox10_protein__Ciona_intestinalis_
81_Moxi-hox10b
89_Mocu-Hox10
97_Mocc-Hox10
24_Human_HOXA11_Hox9-13_15__ANTP
54_Human_HOXD11_Hox9-13_15__ANTP
44_Human_HOXC11_Hox9-13_15__ANTP
2_Fruitfly_Abd-B_Hox9-13_15__ANTP
11_Amphioxus_Hox13_Hox9-13_15__ANTP
12_Amphioxus_Hox14_Hox9-13_15__ANTP
82_Moxi-hox12
90_Mocu-Hox12
98_Mocc-Hox12
7_Amphioxus_Hox1_Hox1_ANTP
67_gi_193083003_ref_NP_001122333_1__homeobox_transcription_factor_Hox1__Ciona_intestinalis_
73_gi_27525469_emb_CAD59667_1__putative_homeobox_protein_hox1__partial__Ciona_intestinalis_
22_Human_HOXA1_Hox1_ANTP
33_Human_HOXB1_Hox1_ANTP
52_Human_HOXD1_Hox1_ANTP
61_Fruitfly_lab_Hox1_ANTP
75_Moxi-hox1
84_Mocu-Hox1
92_Mocc-Hox1
45_Human_HOXC12_Hox9-13_15__ANTP
55_Human_HOXD12_Hox9-13_15__ANTP
71_gi_193083021_ref_NP_001122344_1__homeobox_transcription_factor_Hox13__Ciona_intestinalis_
83_Moxi-hox13
91_Mocu-Hox13
99_Mocc-Hox13
13_Amphioxus_Hox15_Hox9-13_15__ANTP
25_Human_HOXA13_Hox9-13_15__ANTP
46_Human_HOXC13_Hox9-13_15__ANTP
56_Human_HOXD13_Hox9-13_15__ANTP
34_Human_HOXB13_Hox9-13_15__ANTP
80_Moxi-hox10a
4_Fruitfly_bcd_Hox3_ANTP

Figure A.3: **Alignment of *hox* genes.** Alignments of the aa homeobox sequences from all the Molgula species, show that they group with their respective orthologs. All but one of the *M. occidentalis* cluster properly, but *M. occidentalis hox10* full homeobox sequence was not fully assembled, so this is possibly the reason for poor clustering.

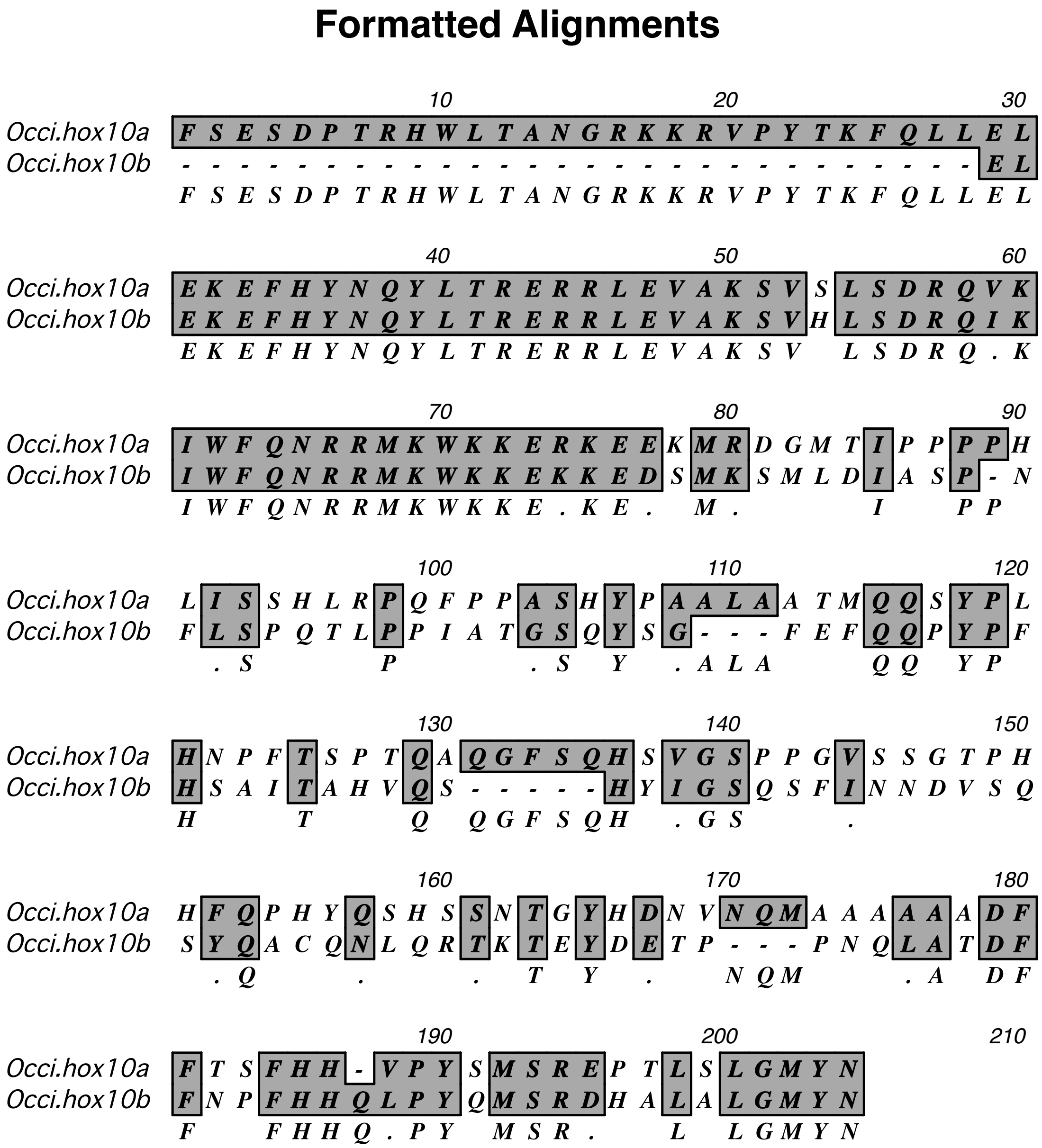

Formatted Alignments
10 20 30
Occi.hox10a F S E S D P T R H W L T A N G R K K R V P Y T K F Q L L E L
Occi.hox10b - E L
F S E S D P T R H W L T A N G R K K R V P Y T K F Q L L E L
40 50 60
Occi.hox10a E K E F H Y N Q Y L T R E R R L E V A K S V S L S D R Q V K
Occi.hox10b E K E F H Y N Q Y L T R E R R L E V A K S V H L S D R Q I K
E K E F H Y N Q Y L T R E R R L E V A K S V L S D R Q . K
70 80 90
Occi.hox10a I W F Q N R R M K W K K E R K E E K M R D G M T I P P P P H
Occi.hox10b I W F Q N R R M K W K K E K K E D S M K S M L D I A S P - N
I W F Q N R R M K W K K E . K E . M . I P P
100 110 120
Occi.hox10a L I S S H L R P Q F P P A S H Y P A A L A A T M Q Q S Y P L
Occi.hox10b F L S P Q T L P P I A T G S Q Y S G - - - F E F Q Q P Y P F
. S P . S Y . A L A Q Q Y P
130 140 150
Occi.hox10a H N P F T S P T Q A Q G F S Q H S V G S P P G V S S G T P H
Occi.hox10b H S A I T A H V Q S - - - - - H Y I G S Q S F I N N D V S Q
H T Q Q G F S Q H . G S .
160 170 180
Occi.hox10a H F Q P H Y Q S H S S N T G Y H D N V N Q M A A A A A A D F
Occi.hox10b S Y Q A C Q N L Q R T K T E Y D E T P - - - P N Q L A T D F
. Q . . T Y . N Q M . A D F
190 200 210
Occi.hox10a F T S F H H - V P Y S M S R E P T L S L G M Y N
Occi.hox10b F N P F H H Q L P Y Q M S R D H A L A L G M Y N
F F H H Q . P Y M S R . L L G M Y N

Figure A.4: **Alignment of *M. occidentalis* duplicate *hox10* genes** Two copies of *hox10* were found in *M. occidentalis* ~12 kb apart on the same contig.

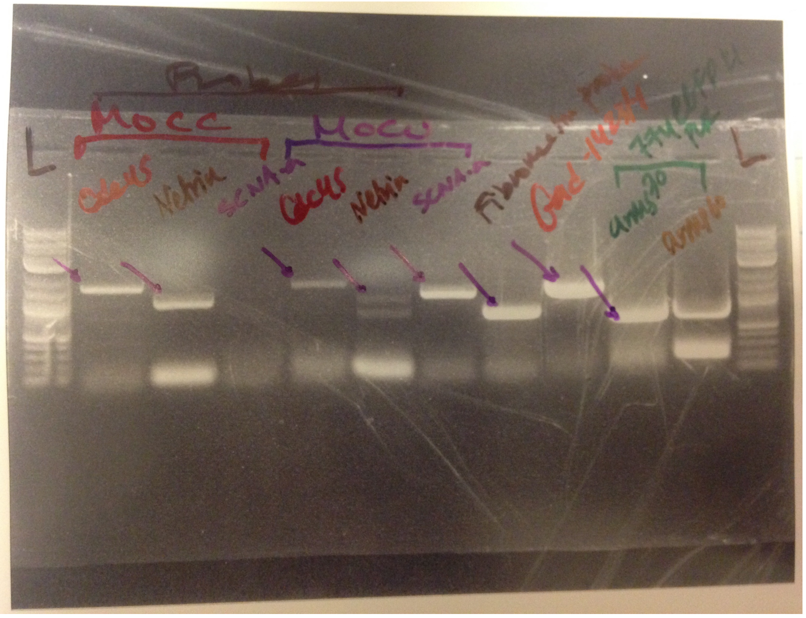


Figure A.5: **Gel electrophoresis of cdc45, netrin and controls** *Netrin* was not found in *M. occulta's* transcriptome but was recovered via PCR from a cDNA library. However, this library has a wider range of developmental stages, so it is still possible *netrin* is not expressed at the right stage for tail development.

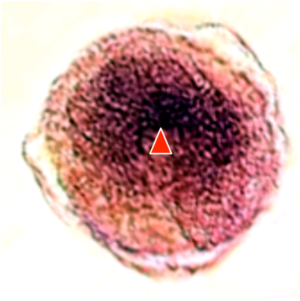


Figure A.6: **Whole Mount In Situ Hybridization of *prickle* in *M. occulta*** The PCP gene *pk* was shown to be expressed in the notochord in a similar pattern to *C. intestinalis*

Appendix B

Contributions

B.1 Chapeter 3: Evaluating a lightweight transcriptome assembly pipeline on two closely related ascidian species

The experimental design for the RNA sequence analysis was decided by C. Titus Brown (CTB) and Billie J. Swalla (BJS). The fertilization and collection of RNA samples were done by BJS and myself. Library preparation for Illumina sequencing was done by Kanchan Pavangadkar. All downstream analysis was done by myself. The idea for evaluating the de novo assembly pipeline came about through conversations with CTB, as well as methods for evaluating the transcriptome assemblies. Writing was done by me with edits from BJS and CTB.

B.2 Chapter 4: Genome assembly and characterization

The experimental design for the DNA sequencing was decided by CTB, BJS, and Lionel Christiaen (LC). Embryonic samples were collected by BJS, LC, Alberto Stolfi (AS), Claudia

Racioppi (CR) and myself. Genome assembly was conducted by me. Ideas for examining divegand were developed by AS, orthologous sequences were identified through the use of RBH blast and done so by me because the sequences were not yet annotated on the Aniseed database. The *Hox* analysis was done by me and was initiated by a question from David Arnosti. Writing was done by AS, myself with edits from BJS, CTB, CR, and LC.

B.3 Chapter 5: Differential expression analysis of tail loss in an invertebrate chordate

The experimental design is the same for the "Evaluating a lightweight transcriptome assembly pipeline on two closely related ascidian species" chapter (3). Transcript models were assembled by me. Read mapping for differential expression analysis was conducted by me, and the analysis of hybrid expression counts was the idea of CTB. The differential expression analysis was conducted by me, including the idea to focus on the overlapping upregulated gene in *M. oculata* and the interspecific hybrid. Annotation of the overlapping upregulated genes was done by AS, Anna Di Gregorio and myself. Writing was done by me with edits from AS, CR, and CTB.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] S. Bassham and J. Postlethwait. Brachyury (t) expression in embryos of a larvacean urochordate, oikopleura dioica, and the ancestral role of t. *Developmental Biology*, 220(2):322–332, Apr. 2000.

[2] N. J. BERRILL. Studies in tunicate developnent. *Society*, 219:281–346, 1931.

[3] K. R. Bradnam, J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J. A. Chapman, G. Chapuis, R. Chikhi, H. Chitsaz, W.-C. Chou, J. Corbeil, C. D. Fabbro, T. R. Docking, R. Durbin, D. Earl, S. Emrich, P. Fedotov, N. A. Fonseca, G. Ganapathy, R. A. Gibbs, S. Gnerre, . Godzaridis, S. Goldstein, M. Haimel, G. Hall, D. Haussler, J. B. Hiatt, I. Y. Ho, J. Howard, M. Hunt, S. D. Jackman, D. B. Jaffe, E. D. Jarvis, H. Jiang, S. Kazakov, P. J. Kersey, J. O. Kitzman, J. R. Knight, S. Koren, T.-W. Lam, D. Lavenier, F. Laviolette, Y. Li, Z. Li, B. Liu, Y. Liu, R. Luo, I. MacCallum, M. D. MacManes, N. Maillet, S. Melnikov, D. Naquin, Z. Ning, T. D. Otto, B. Paten, O. S. Paulo, A. M. Phillippy, F. Pina-Martins, M. Place, D. Przybylski, X. Qin, C. Qu, F. J. Ribeiro, S. Richards, D. S. Rokhsar, J. G. Ruby, S. Scalabrin, M. C. Schatz, D. C. Schwartz, A. Sergushichev, T. Sharpe, T. I. Shaw, J. Shendure, Y. Shi, J. T. Simpson, H. Song, F. Tsarev, F. Vezzi, R. Vicedomini, B. M. Vieira, J. Wang, K. C. Worley, S. Yin, S.-M. Yiu, J. Yuan, G. Zhang, H. Zhang, S. Zhou, and I. F. Korf. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, 2(1):10, July 2013.

[4] C. D. Brown, D. S. Johnson, and A. Sidow. Functional architecture and evolution of transcriptional elements that drive gene coexpression. *Science (New York, N.Y.)*, 317(5844):1557–1560, Sept. 2007.

[5] C. T. Brown, A. Howe, Q. Zhang, A. B. Pyrkosz, and T. H. Brom. A reference-free algorithm for computational normalization of shotgun sequencing data. arXiv e-print 1203.4802, Mar. 2012.

[6] M. Brudno, C. B. Do, G. M. Cooper, M. F. Kim, E. Davydov, NISC Comparative Sequencing Program, E. D. Green, A. Sidow, and S. Batzoglou. LAGAN and multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. *Genome Research*, 13(4):721–731, Apr. 2003.

[7] C. B. Cameron, J. R. Garey, and B. J. Swalla. Evolution of the chordate body plan: New insights from phylogenetic analyses of deuterostome phyla. *Proceedings of the National Academy of Sciences*, 97(9):4469–4474, Apr. 2000.

[8] A. Caracciolo, A. Di Gregorio, F. Aniello, R. Di Lauro, and M. Branno. Identification and developmental expression of three distal-less homeobox containing genes in the ascidian ciona intestinalis. *Mechanisms of Development*, 99(12):173–176, Dec. 2000.

[9] L. Christiaen, A. Stolfi, B. Davidson, and M. Levine. Spatio-temporal intersection of lhx3 and tbx6 defines the cardiac field through synergistic activation of mesp. *Developmental Biology*, 328(2):552–560, Apr. 2009.

[10] K. Clarke, Y. Yang, R. Marsh, L. Xie, and K. K. Zhang. Comparative analysis of de novo transcriptome assembly. *Science China Life Sciences*, 56(2):156–162, Feb. 2013.

[11] E. G. Conklin. *The organization and cell-lineage of the ascidian egg.* Philadelphia : [Academy of Natural Sciences], 1905.

[12] J. C. Corbo, M. Levine, and R. W. Zeller. Characterization of a notochord-specific enhancer from the brachyury promoter region of the ascidian, ciona intestinalis. *Development (Cambridge, England)*, 124(3):589–602, Feb. 1997.

[13] B. Davidson, W. Shi, and M. Levine. Uncoupling heart cell specification and migration in the simple chordate ciona intestinalis. *Development (Cambridge, England)*, 132(21):4811–4818, Nov. 2005.

[14] P. Dehal, Y. Satou, R. K. Campbell, J. Chapman, B. Degnan, A. D. Tomaso, B. Davidson, A. D. Gregorio, M. Gelpke, D. M. Goodstein, N. Harafuji, K. E. M. Hastings, I. Ho, K. Hotta, W. Huang, T. Kawashima, P. Lemaire, D. Martinez, I. A. Meinertzhagen, S. Necula, M. Nonaka, N. Putnam, S. Rash, H. Saiga, M. Satake, A. Terry, L. Yamada, H.-G. Wang, S. Awazu, K. Azumi, J. Boore, M. Branno, S. Chin-bow, R. DeSantis, S. Doyle, P. Francino, D. N. Keys, S. Haga, H. Hayashi, K. Hino, K. S. Imai, K. Inaba, S. Kano, K. Kobayashi, M. Kobayashi, B.-I. Lee, K. W. Makabe, C. Manohar, G. Matassi, M. Medina, Y. Mochizuki, S. Mount, T. Morishita, S. Miura, A. Nakayama, S. Nishizaka, H. Nomoto, F. Ohta, K. Oishi, I. Rigoutsos, M. Sano, A. Sasaki, Y. Sasakura, E. Shoguchi, T. Shin-i, A. Spagnuolo, D. Stainier, M. M. Suzuki, O. Tassy, N. Takatori, M. Tokuoka, K. Yagi, F. Yoshizaki, S. Wada, C. Zhang, P. D. Hyatt, F. Larimer, C. Detter, N. Doggett, T. Glavina, T. Hawkins, P. Richardson, S. Lucas, Y. Kohara, M. Levine, N. Satoh, and D. S. Rokhsar. The draft genome of ciona intestinalis: Insights into chordate and vertebrate origins. *Science*, 298(5601):2157–2167, Dec. 2002.

[15] F. Denoeud, S. Henriet, S. Mungpakdee, J.-M. Aury, C. D. Silva, H. Brinkmann, J. Mikhaleva, L. C. Olsen, C. Jubin, C. Caestro, J.-M. Bouquet, G. Danks, J. Poulain, C. Campsteijn, M. Adamski, I. Cross, F. Yadetie, M. Muffato, A. Louis, S. Butcher, G. Tsagkogeorga, A. Konrad, S. Singh, M. F. Jensen, E. H. Cong, H. Eikeseth-Otteraa, B. Noel, V. Anthouard, B. M. Porcel, R. Kachouri-Lafond, A. Nishino, M. Ugolini, P. Chourrout, H. Nishida, R. Aasland, S. Huzurbazar, E. Westhof, F. Delsuc, H. Lehrach, R. Reinhardt, J. Weissenbach, S. W. Roy, F. Artiguenave, J. H. Postlethwait, J. R. Manak, E. M. Thompson, O. Jaillon, L. D. Pasquier, P. Boudinot, D. A. Liberles, J.-N. Volff, H. Philippe, B. Lenhard, H. R. Crollius, P. Wincker, and

D. Chourrout. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. *Science*, 330(6009):1381–1385, Dec. 2010.

[16] A. Di Gregorio, R. M. Harland, M. Levine, and E. S. Casey. Tail morphogenesis in the ascidian, ciona intestinalis, requires cooperation between notochord and muscle. *Developmental biology*, 244(2):385–95, Apr. 2002.

[17] J. R. Finnerty. The origins of axial patterning in the metazoa: how old is bilateral symmetry? *The International Journal of Developmental Biology*, 47(7-8):523–529, 2003.

[18] K. A. Frazer, L. Pachter, A. Poliakov, E. M. Rubin, and I. Dubchak. VISTA: computational tools for comparative genomics. *Nucleic Acids Research*, 32(Web Server issue):W273–279, July 2004.

[19] W. J. Gehring, Y. Q. Qian, M. Billeter, K. Furukubo-Tokunaga, A. F. Schier, D. Resendez-Perez, M. Affolter, G. Otting, and K. Wthrich. Homeodomain-DNA recognition. *Cell*, 78(2):211–223, July 1994.

[20] T. C. Glenn. Field guide to next-generation DNA sequencers. *Molecular Ecology Resources*, 11(5):759–769, Sept. 2011.

[21] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. a. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. Full-length transcriptome assembly from RNA-seq data without a reference genome. *Nature biotechnology*, 29(7):644–52, July 2011.

[22] F. Gyoja, Y. Satou, T. Shin-i, Y. Kohara, B. J. Swalla, and N. Satoh. Analysis of large scale expression sequenced tags (ESTs) from the anural ascidian, molgula tectiformis. *Developmental biology*, 307(2):460–82, July 2007.

[23] B. J. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P. D. Blood, J. Bowden, M. B. Couger, D. Eccles, B. Li, M. Lieber, M. D. MacManes, M. Ott, J. Orvis, N. Pochet, F. Strozzi, N. Weeks, R. Westerman, T. William, C. N. Dewey, R. Henschel, R. D. LeDuc, N. Friedman, and A. Regev. De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. *Nat. Protocols*, 8(8):1494–1512, Aug. 2013.

[24] K. A. Hadfield, B. J. Swalla, and W. R. Jeffery. Multiple origins of anural development in ascidians inferred from rDNA sequences. *Journal of Molecular Evolution*, 40(4):413–427, Apr. 1995.

[25] E. E. Hare, B. K. Peterson, and M. B. Eisen. A careful look at binding site reorganization in the even-skipped enhancers of drosophila and sepsids. *PLoS genetics*, 4(11):e1000268, Nov. 2008.

[26] H. Hashimoto, T. Enomoto, G. Kumano, and H. Nishida. The transcription factor FoxB mediates temporal loss of cellular competence for notochord induction in ascidian embryos. *Development*, 138(14):3091–3091, June 2011.

[27] J. J. Henry and R. A. Raff. Evolutionary change in the process of dorsoventral axis determination in the direct developing sea urchin, heliocidaris erythrogramma. *Developmental Biology*, 141(1):55–69, Sept. 1990.

[28] B. G. Herrmann and A. Kispert. The t genes in embryogenesis. *Trends in Genetics*, 10(8):280–286, Aug. 1994.

[29] T. Hirano and H. Nishida. Developmental fates of larval tissues after metamorphosis in ascidian halocynthia roretzi. i. origin of mesodermal tissues of the juvenile. *Developmental Biology*, 192(2):199–210, Dec. 1997.

[30] K. Hotta, H. Takahashi, T. Asakura, B. Saitoh, N. Takatori, Y. Satou, and N. Satoh. Characterization of brachyury-downstream notochord genes in the ciona intestinalis embryo. *Developmental biology*, 224(1):69–80, Aug. 2000.

[31] K. Hotta, H. Takahashi, a. Erives, M. Levine, and N. Satoh. Temporal expression patterns of 39 brachyury-downstream genes associated with notochord formation in the ciona intestinalis embryo. *Development, growth & differentiation*, 41(6):657–64, Dec. 1999.

[32] K. Hotta, S. Yamada, N. Ueno, N. Satoh, and H. Takahashi. Brachyury-downstream notochord genes and convergent extension in ciona intestinalis embryos. *Development, Growth & Differentiation*, 49(5):373–382, June 2007.

[33] A. C. Howe, J. K. Jansson, S. A. Malfatti, S. G. Tringe, J. M. Tiedje, and C. T. Brown. Tackling soil diversity with the assembly of large, complex metagenomes. *Proceedings of the National Academy of Sciences*, 111(13):4904–4909, Apr. 2014.

[34] J. L. Huber, K. B. da Silva, W. R. Bates, and B. J. Swalla. The evolution of anural larvae in molgulid ascidians. *Seminars in cell & developmental biology*, 11(6):419–26, Dec. 2000.

[35] C. Hudson, M. Ba, C. Rouvire, and H. Yasuo. Divergent mechanisms specify chordate motoneurons: evidence from ascidians. *Development (Cambridge, England)*, 138(8):1643–52, Apr. 2011.

[36] T. Ikuta and H. Saiga. Organization of hox genes in ascidians: present, past, and future. *Developmental Dynamics: An Official Publication of the American Association of Anatomists*, 233(2):382–389, June 2005.

[37] T. Ikuta, N. Satoh, and H. Saiga. Limited functions of hox genes in the larval development of the ascidian ciona intestinalis. *Development (Cambridge, England)*, 137(9):1505–1513, May 2010.

[38] T. Ikuta, N. Yoshida, N. Satoh, and H. Saiga. Ciona intestinalis hox gene cluster: Its dispersed structure and residual colinear expression in development. *Proceedings of the National Academy of Sciences of the United States of America*, 101(42):15118–15123, Oct. 2004.

[39] Jeffery, R. billie, and J. Swalla. Factors necessary for restoring an evolutionary change in an anural ascidian embryo. *Developmental biology*, 153:194–205, 1992.

[40] W. R. Jeffery. Minireview ascidian gene-expression profiles. *Genome biology*, 3(10):1–4, 2002.

[41] W. R. Jeffery and B. J. Swalla. An evolutionary change in the muscle lineage of an anural ascidian embryo is restored by interspecific hybridization with a urodele ascidian. *Developmental Biology*, 337:328–337, 1991.

[42] W. R. Jeffery, B. J. Swalla, N. Ewing, and T. Kusakabe. Evolution of the ascidian anural larva: evidence from embryos and molecules. *Molecular biology and evolution*, 16(5):646–54, May 1999.

[43] D. Jiang, E. M. Munro, W. C. Smith, S. Barbara, and F. Harbor. Ascidian prickle regulates both mediolateral and anterior-posterior cell polarity of notochord cells. *Current biology*, 15:79–85, 2005.

[44] D. S. Johnson, B. Davidson, C. D. Brown, W. C. Smith, and A. Sidow. Noncoding regulatory sequences of ciona exhibit strong correspondence between evolutionary constraint and functional importance. *Genome Research*, 14(12):2448–2456, Dec. 2004.

[45] L. Katikala, H. Aihara, Y. J. Passamaneck, S. Gazdoiu, D. S. Jos-Edwards, J. E. Kugler, I. Oda-Ishii, J. H. Imai, Y. Nibu, and A. Di Gregorio. Functional brachyury binding sites establish a temporal read-out of gene expression in the ciona notochord. *PLoS Biol*, 11(10):e1001697, Oct. 2013.

[46] K. Katoh and D. M. Standley. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, 30(4):772–780, Apr. 2013.

[47] K. Katoh and H. Toh. PartTree: an algorithm to build an approximate tree from a large number of unaligned sequences. *Bioinformatics (Oxford, England)*, 23(3):372–374, Feb. 2007.

[48] A. I. Kavka and J. B. A. Green. Tales of tails: Brachyury and the t-box genes. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1333(2):F73–F84, Oct. 1997.

[49] R. Keller, L. Davidson, a. Edlund, T. Elul, M. Ezin, D. Shook, and P. Skoglund. Mechanisms of convergence and extension by cell intercalation. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 355(1399):897–922, July 2000.

[50] K. Kobayashi, L. Yamada, and N. Satoh. Differential gene expression in notochord and nerve cord fate segregation in the ciona intestinalis embryo. *Genesis*, 51(9):647–659, 2013.

[51] M. J. Kourakis, W. Reeves, E. Newman-Smith, B. Maury, S. Abdul-Wajid, and W. C. Smith. A one-dimensional model of PCP signaling: Polarized cell behavior in the notochord of the ascidian ciona. *Developmental Biology*, 395(1):120–130, Nov. 2014.

[52] A. Kubo, N. Suzuki, X. Yuan, K. Nakai, N. Satoh, K. S. Imai, and Y. Satou. Genomic cis-regulatory networks in the early ciona intestinalis embryo. *Development (Cambridge, England)*, 137(10):1613–1623, May 2010.

[53] J. E. Kugler, P. Kerner, J.-M. Bouquet, D. Jiang, and A. Di Gregorio. Evolutionary changes in the notochord genetic toolkit: a comparative analysis of notochord genes in the ascidian ciona and the larvacean oikopleura. *BMC evolutionary biology*, 11(1):21, Jan. 2011.

[54] J. E. Kugler, Y. J. Passamaneck, T. G. Feldman, J. Beh, T. W. Regnier, and A. Di Gregorio. Evolutionary conservation of vertebrate notochord genes in the ascidian ciona intestinalis. *Genesis (New York, N.Y. : 2000)*, 46(11):697–710, Nov. 2008.

[55] C. D. Laird. Chromatid structure: relationship between DNA content and nucleotide sequence diversity. *Chromosoma*, 32(4):378–406, Dec. 1971.

[56] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature Methods*, 9(4):357–359, Apr. 2012.

[57] A. Lauri, T. Brunet, M. Handberg-Thorsager, A. H. L. Fischer, O. Simakov, P. R. H. Steinmetz, R. Tomer, P. J. Keller, and D. Arendt. Development of the annelid axochord: Insights into notochord evolution. *Science*, 345(6202):1365–1368, Sept. 2014.

[58] P. Lemaire. Unfolding a chordate developmental program, one cell at a time: invariant cell lineages, short-range inductions and evolutionary plasticity in ascidians. *Developmental biology*, 332(1):48–60, Aug. 2009.

[59] P. Lemaire. Evolutionary crossroads in developmental biology: the tunicates. *Development*, 138(11):2143–2152, May 2011.

[60] P. Lemaire, W. C. Smith, and H. Nishida. Ascidians and the plasticity of the chordate developmental program. *Current biology : CB*, 18(14):R620–31, July 2008.

[61] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079, Aug. 2009.

[62] W. Li and A. Godzik. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659, July 2006.

[63] L. Liu, Y. Li, S. Li, N. Hu, Y. He, R. Pong, D. Lin, L. Lu, and M. Law. Comparison of next-generation sequencing systems. *BioMed Research International*, 2012:e251364, July 2012.

[64] M. Lohse, A. M. Bolger, A. Nagel, A. R. Fernie, J. E. Lunn, M. Stitt, and B. Usadel. RobiNA: a user-friendly, integrated software solution for RNA-seq-based transcriptomics. *Nucleic Acids Research*, 40(W1):W622–W627, June 2012.

[65] W. J. R. Longabaugh, E. H. Davidson, and H. Bolouri. Visualization, documentation, analysis, and communication of large-scale gene regulatory networks. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1789(4):363–374, Apr. 2009.

[66] E. K. Lowe, B. J. Swalla, and C. T. Brown. Evaluating a lightweight transcriptome assembly pipeline on two closely related ascidian species. *PeerJ PrePrints*, 2:e505v1, 2014.

[67] M. D. Macmanes. On the optimal trimming of high-throughput mRNA sequence data. *Frontiers in Genetics*, 5:13, 2014.

[68] M. E. Maliska, M. W. Pennell, and B. J. Swalla. Developmental mode influences diversification in ascidians. *Biology Letters*, 9(3):20130068, June 2013. Ascidian species (Tunicata: Ascidiacea) usually have tailed, hatching tadpole larvae. In several lineages, species have evolved larvae that completely lack any tail tissues and are unable to disperse actively. Some tailless species hatch, but some do not hatch before going through metamorphosis. We show here that ascidian species with the highest speciation rates are those with the largest range sizes and tailed hatching larval development. We use methods for examining diversification in binary characters across a posterior distribution of trees, and show that mode of larval development predicts geographical range sizes. Conversely, we find that species with the least dispersive larval development (tailless, non-hatching) have the lowest speciation rates and smallest geographical ranges. Our speciation rate results are contrary to findings from sea urchins and snails examined in the fossil record, and further work is necessary to reconcile these disparate results.

[69] M. E. Maliska and B. J. Swalla. Molgula pugetiensis is a pacific tailless ascidian within the roscovita clade of molgulids. *The Biological Bulletin*, 219(3):277–282, Dec. 2010.

[70] M. Mallo and C. R. Alonso. The regulation of hox gene expression during animal. *Development*, 140(19):3951–3963, Oct. 2013.

[71] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517, Sept. 2008.

[72] J. a. Martin and Z. Wang. Next-generation transcriptome assembly. *Nature reviews. Genetics*, 12(10):671–82, Oct. 2011.

[73] C. Mayor, M. Brudno, J. R. Schwartz, A. Poliakov, E. M. Rubin, K. A. Frazer, L. S. Pachter, and I. Dubchak. VISTA : visualizing global DNA sequence alignments of arbitrary length. *Bioinformatics*, 16(11):1046–1047, Nov. 2000.

[74] W. McGinnis, R. L. Garber, J. Wirz, A. Kuroiwa, and W. J. Gehring. A homologous protein-coding sequence in drosophila homeotic genes and its conservation in other metazoans. *Cell*, 37(2):403–408, June 1984.

[75] W. McGinnis and R. Krumlauf. Homeobox genes and axial patterning. *Cell*, 68(2):283–302, Jan. 1992.

[76] M. L. Metzker. Emerging technologies in DNA sequencing. *Genome Research*, 15(12):1767–1776, Dec. 2005.

[77] T. Minokawa, K. Yagi, K. W. Makabe, and H. Nishida. Binary specification of nerve cord and notochord cell fates in ascidian embryos. *Development*, 128(11):2007–2017, June 2001.

[78] D. Miyamoto and R. Crowther. Formation of the notochord in living ascidian embryos. *Journal of Embryology and Experimental Morphology*, VOL. 86:1–17, 1985.

[79] Y. Nakatani, R. Moody, and W. C. Smith. Mutations affecting tail and notochord development in the ascidian ciona savignyi. *Development*, 126(15):3293–3301, Aug. 1999.

[80] Y. Nakatani and H. Nishida. Duration of competence and inducing capacity of blastomeres in notochord induction during ascidian embryogenesis. *Development, Growth & Differentiation*, 41(4):449–453, Aug. 1999.

[81] Y. Nakatani, H. Yasuo, N. Satoh, and H. Nishida. Basic fibroblast growth factor induces notochord formation and the expression of as-t, a brachyury homolog, during ascidian embryogenesis. *Development (Cambridge, England)*, 122(7):2023–2031, July 1996.

[82] H. Nishida. Cell lineage analysis in ascidian embryos by intracellular injection of a tracer enzyme: III. up to the tissue restricted stage. *Developmental Biology*, 121(2):526–541, June 1987.

[83] H. Nishida. Development of the appendicularian oikopleura dioica: culture, genome, and cell lineages. *Development, Growth & Differentiation*, 50 Suppl 1:S239–256, June 2008.

[84] H. Nishida and N. Satoh. Cell lineage analysis in ascidian embryos by intracellular injection of a tracer enzyme: I. up to the eight-cell stage. *Developmental Biology*, 99(2):382–394, Oct. 1983.

[85] H. Nishida and N. Satoh. Cell lineage analysis in ascidian embryos by intracellular injection of a tracer enzyme: II. the 16- and 32-cell stages. *Developmental Biology*, 110(2):440–454, Aug. 1985.

[86] H. Nishida and T. Stach. Cell lineages and fate maps in tunicates: Conservation and modification. *Zoological Science*, 31(10):645–652, Oct. 2014.

[87] S. T. ONeil and S. J. Emrich. Assessing de novo transcriptome assembly metrics for consistency and utility. *BMC Genomics*, 14(1):465, July 2013.

[88] M. Paris and V. Laudet. The history of a developmental stage: metamorphosis in chordates. *Genesis (New York, N.Y. : 2000)*, 46(11):657–72, Nov. 2008.

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

[90] Y. J. Passamaneck, L. Katikala, L. Perrone, M. P. Dunn, I. Oda-Ishii, and A. Di Gregorio. Direct activation of a notochord cis-regulatory module by brachyury and FoxA in the ascidian ciona intestinalis. *Development (Cambridge, England)*, 136(21):3679–89, Nov. 2009.

[91] J. K. Pickrell, J. C. Marioni, A. A. Pai, J. F. Degner, B. E. Engelhardt, E. Nkadori, J.-B. Veyrieras, M. Stephens, Y. Gilad, and J. K. Pritchard. variation with RNA sequencing. *Nature*, 464(7289):768–772, 2010.

[92] M. Pop. Genome assembly reborn: recent computational challenges. *Briefings in Bioinformatics*, 10(4):354–366, July 2009.

[93] M. A. Quail, M. Smith, P. Coupland, T. D. Otto, S. R. Harris, T. R. Connor, A. Bertoni, H. P. Swerdlow, and Y. Gu. A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC genomics*, 13:341, 2012.

[94] A. Roberts and L. Pachter. Streaming fragment assignment for real-time analysis of sequencing experiments. *Nature Methods*, 10(1):71–73, Jan. 2013.

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

[96] E. F. Rossomando and S. Alexander. *Morphogenesis: An Analysis of the Development of Biological Form: An Analysis of the Development of Biological Form*. CRC Press, Apr. 1992.

[97] N. Satoh. Ascidian embryos as a model system to analyze expression and function of developmental genes. *Differentiation; Research in Biological Diversity*, 68(1):1–12, Aug. 2001.

[98] N. Satoh. The ascidian tadpole larva: comparative molecular development and genomics. *Nature reviews. Genetics*, 4(4):285–95, Apr. 2003.

[99] N. Satoh and M. Levine. Surfing with the tunicates into the post-genome era. *Genes & development*, 19(20):2407–11, Oct. 2005.

[100] N. Satoh, Y. Satou, B. Davidson, and M. Levine. Ciona intestinalis: an emerging model for whole-genome analyses. *Trends in genetics : TIG*, 19(7):376–81, July 2003.

[101] Y. Satou, K. S. Imai, and N. Satoh. The ascidian mesp gene specifies heart precursor cells. *Development (Cambridge, England)*, 131(11):2533–2541, June 2004.

[102] Y. Satou, K. Mineta, M. Ogasawara, Y. Sasakura, E. Shoguchi, K. Ueno, L. Yamada, J. Matsumoto, J. Wasserscheid, K. Dewar, G. B. Wiley, S. L. Macmil, B. A. Roe, R. W. Zeller, K. E. M. Hastings, P. Lemaire, E. Lindquist, T. Endo, K. Hotta, and K. Inaba. Improved genome assembly and evidence-based global gene model set for the chordate ciona intestinalis: new insight into intron and operon populations. *Genome Biology*, 9(10):R152, 2008.

[103] M. H. Schulz, D. R. Zerbino, M. Vingron, and E. Birney. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics (Oxford, England)*, 28(8):1086–92, Apr. 2012.

[104] E. M. Schwarz, P. K. Korhonen, B. E. Campbell, N. D. Young, A. R. Jex, A. Jabbar, R. S. Hall, A. Mondal, A. C. Howe, J. Pell, A. Hofmann, P. R. Boag, X.-Q. Zhu, T. R. Gregory, A. Loukas, B. A. Williams, I. Antoshechkin, C. T. Brown, P. W. Sternberg, and R. B. Gasser. The genome and developmental transcriptome of the strongylid nematode haemonchus contortus. *Genome Biology*, 14(8):R89, Aug. 2013.

[105] H.-C. Seo, R. B. Edvardsen, A. D. Maeland, M. Bjordal, M. F. Jensen, A. Hansen, M. Flaat, J. Weissenbach, H. Lehrach, P. Wincker, R. Reinhardt, and D. Chourrout. Hox cluster disintegration with persistent anteroposterior order of expression in oikopleura dioica. *Nature*, 431(7004):67–71, Sept. 2004.

[106] H.-C. Seo, M. Kube, R. B. Edvardsen, M. F. Jensen, A. Beck, E. Spriet, G. Gorsky, E. M. Thompson, H. Lehrach, R. Reinhardt, and D. Chourrout. Miniature genome in the marine chordate oikopleura dioica. *Science*, 294(5551):2506–2506, Dec. 2001.

[107] M. W. Simmen, S. Leitgeb, V. H. Clark, S. J. Jones, and A. Bird. Gene number in an invertebrate chordate, ciona intestinalis. *Proceedings of the National Academy of Sciences of the United States of America*, 95(8):4437–4440, Apr. 1998.

[108] W. T. Siok, C. A. Perfetti, Z. Jin, and L. H. Tan. Biological abnormality of impaired reading is constrained by culture. *Nature*, 431(7004):71–76, Sept. 2004.

[109] T. Stach and J. M. Turbeville. Phylogeny of tunicata inferred from molecular and morphological characters. *Molecular Phylogenetics and Evolution*, 25(3):408–428, Dec. 2002.

[110] J. Stapley, J. Reger, P. G. D. Feulner, C. Smadja, J. Galindo, R. Ekblom, C. Bennison, A. D. Ball, A. P. Beckerman, and J. Slate. Adaptation genomics: the next generation. *Trends in ecology & evolution*, 25(12):705–12, Dec. 2010.

[111] D. L. Stemple. Structure and function of the notochord: an essential organ for chordate development. *Development (Cambridge, England)*, 132(11):2503–12, June 2005.

[112] A. Stolfi and L. Christiaen. Genetic and genomic toolbox of the chordate ciona intestinalis. *Genetics*, 192(1):55–66, Sept. 2012.

[113] A. Stolfi, T. B. Gainous, J. J. Young, A. Mori, M. Levine, and L. Christiaen. Early chordate origins of the vertebrate second heart field. *Science (New York, N.Y.)*, 329(5991):565–568, July 2010.

[114] A. Stolfi and M. Levine. Neuronal subtype specification in the spinal cord of a protovertebrate. *Development (Cambridge, England)*, 138(5):995–1004, Mar. 2011.

[115] A. Stolfi, E. K. Lowe, C. Racioppi, F. Ristoratore, C. T. Brown, B. J. Swalla, and L. Christiaen. Divergent mechanisms regulate conserved cardiopharyngeal development and gene expression in distantly related ascidians. *eLife*, 3:e03728, 2014.

[116] A. Stolfi, Y. Sasakura, D. Chalopin, Y. Satou, L. Christiaen, C. Dantec, T. Endo, M. Naville, H. Nishida, B. J. Swalla, J.-N. Volff, A. Voskoboynik, D. Dauga, and P. Lemaire. Guidelines for the nomenclature of genetic elements in tunicate genomes. *Genesis (New York, N.Y.: 2000)*, Sept. 2014.

[117] B. J. Swalla. Mechanisms of gastrulation and tail formation in ascidians. *Microscopy research and technique*, 26(4):274–84, 1993.

[118] B. J. Swalla. Protochordate gastrulation: lancelets and ascidians. gastrulation. *Cold Spring Harbor: Cold Spring Harbor Press*, 49:139, 2004.

[119] B. J. Swalla, M. R. Badgett, and W. R. Jeffery. Identification of a cytoskeletal protein localized in the myoplasm of ascidian eggs: localization is modified during anural development. *Development (Cambridge, England)*, 111(2):425–436, Feb. 1991.

[120] B. J. Swalla and W. R. Jeffery. Interspecific hybridization between an anural and urodele ascidian: differential expression of urodele features suggests multiple mechanisms control anural development. *Developmental biology*, 142(2):319–34, Dec. 1990.

[121] B. J. Swalla and W. R. Jeffery. Requirement of the manx gene for expression of chordate features in a tailless ascidian larva. *Science (New York, N.Y.)*, 274(5290):1205–8, Nov. 1996.

[122] B. J. Swalla, M. a. Just, E. L. Pederson, and W. R. Jeffery. A multigene locus containing the manx and bobcat genes is required for development of chordate features in the ascidian tadpole larva. *Development (Cambridge, England)*, 126(8):1643–53, Apr. 1999.

[123] B. J. Swalla, K. W. Makabe, N. Satoh, and W. R. Jeffery. Novel genes expressed differentially in ascidians with alternate modes of development. *Development*, 318:307–318, 1993.

[124] N. Takada, N. Satoh, and B. J. Swalla. Expression of tbx6, a muscle lineage t-box gene, in the tailless embryo of the ascidian molgula tectiformis. *Development Genes and Evolution*, 212(7):354–356, Aug. 2002.

[125] N. Takada, J. York, J. M. Davis, B. Schumpert, H. Yasuo, N. Satoh, and B. J. Swalla. Brachyury expression in tailless molgulid ascidian embryos. *Evolution & development*, 4(3):205–11, 2002.

[126] O. Tassy, D. Dauga, F. Daian, D. Sobral, F. Robin, P. Khoueiry, D. Salgado, V. Fox, D. Caillol, R. Schiappa, B. Laporte, A. Rios, G. Luxardi, T. Kusakabe, J.-S. Joly, S. Darras, L. Christiaen, M. Contensin, H. Auger, C. Lamy, C. Hudson, U. Rothbaecher, M. Gilchrist, K. Makabe, K. Hotta, S. Fujiwara, N. Satoh, Y. Satou, and P. Lemaire. The ANISEED database: Digital representation, formalization and elucidation of a chordate developmental program. *Genome research*, pages 1459–1468, July 2010.

[127] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H. Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. *Nature Protocols*, 7(3):562–578, Mar. 2012.

[128] J. R. True and E. S. Haag. Developmental system drift and flexibility in evolutionary trajectories. *Evolution & Development*, 3(2):109–119, Apr. 2001.

[129] G. Tsagkogeorga, X. Turon, R. R. Hopcroft, M.-K. Tilak, T. Feldstein, N. Shenkar, Y. Loya, D. Huchon, E. J. Douzery, and F. Delsuc. An updated 18s rRNA phylogeny of tunicates based on mixture and secondary structure models. *BMC Evolutionary Biology*, 9(1):187, Aug. 2009.

[130] G. Vachon, B. Cohen, C. Pfeifle, M. E. McGuffin, J. Botas, and S. M. Cohen. Homeotic genes of the bithorax complex repress limb development in the abdomen of the drosophila embryo through the target gene distal-less. *Cell*, 71(3):437–450, Oct. 1992.

[131] S. M. Van Belleghem, D. Roelofs, J. Van Houdt, and F. Hendrickx. De novo transcriptome assembly and SNP discovery in the wing polymorphic salt marsh beetle pogonus chalceus (coleoptera, carabidae). *PLoS ONE*, 7(8):e42605, Aug. 2012.

[132] M. T. Veeman, Y. Nakatani, C. Hendrickson, V. Ericson, C. Lin, and W. C. Smith. Chongmague reveals an essential role for laminin-mediated boundary formation in chordate convergence and extension movements. *Development (Cambridge, England)*, 135(1):33–41, Jan. 2008.

[133] N. Vijay, J. W. Poelstra, A. Knstner, and J. B. W. Wolf. Challenges and strategies in transcriptome assembly and differential gene expression quantification. a comprehensive in silico assessment of RNA-seq experiments. *Molecular ecology*, 46:620–634, Sept. 2012.

[134] J. P. Vinson, D. B. Jaffe, K. O'Neill, E. K. Karlsson, N. Stange-Thomann, S. Anderson, J. P. Mesirov, N. Satoh, Y. Satou, C. Nusbaum, B. Birren, J. E. Galagan, and E. S. Lander. Assembly of polymorphic genomes: algorithms and application to ciona savignyi. *Genome research*, 15(8):1127–1135, Aug. 2005.

[135] A. Visel, S. Minovitsky, I. Dubchak, and L. a. Pennacchio. VISTA enhancer browsera database of tissue-specific human enhancers. *Nucleic acids research*, 35(Database issue):D88–92, Jan. 2007.

[136] H. Wada and N. Satoh. Details of the evolutionary history from invertebrates to vertebrates, as deduced from the sequences of 18s rDNA. *Proceedings of the National Academy of Sciences of the United States of America*, 91(5):1801–1804, Mar. 1994.

[137] Z. Wang, M. Gerstein, and M. Snyder. RNA-seq: a revolutionary tool for transcriptomics. *Nature reviews. Genetics*, 10(1):57–63, Jan. 2009.

[138] G. A. Wray and R. A. Raff. Evolutionary modification of cell lineage in the direct-developing sea urchin heliocidaris erythrogramma. *Developmental Biology*, 132(2):458–470, Apr. 1989.

[139] H. Yasuo and N. Satoh. An ascidian homolog of the mouse brachyury (t) gene is expressed exclusively in notochord cells at the fate restricted stage. *Development, Growth & Differentiation*, 36(1):9–18, Feb. 1994.

[140] H. Yasuo and N. Satoh. Conservation of the developmental role of brachyury in notochord formation in a urochordate, the ascidian halocynthia roretzi. *Developmental Biology*, 200(2):158–170, Aug. 1998.

[141] D. R. Zerbino and E. Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, 18(5):821–9, May 2008.

[142] J. Zhang, R. Chiodini, A. Badr, and G. Zhang. The impact of next-generation sequencing on genomics. *Journal of genetics and genomics*, 38(3):95–109, Mar. 2011.