THE WOOD CHIPPER PROBLEM, MAKING SENSE OF THE SCRAPS: NEXT GENERATION SEQUENCING ANALYSIS OF CLOSELY RELATED TAILED AND TAIL-LESS ASCIDIAN SPECIES

By

Elijah K. Lowe

A DISSERTATION

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

Computer Science & Engineering

2014

ABSTRACT

THE WOOD CHIPPER PROBLEM, MAKING SENSE OF THE SCRAPS: NEXT GENERATION SEQUENCING ANALYSIS OF CLOSELY RELATED TAILED AND TAIL-LESS ASCIDIAN SPECIES

By

Elijah K. Lowe

Abstract goes here

ACKNOWLEDGMENTS

Acknowledgements go here

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Background

Chordates are a branch of deuterostome that are characterized by a dorsal nervous system, pharyngeal gill slits, and defined by the presence of a notochord. Tunicates are one of the three subphyla of chordates and are grouped because of their outer covering known as a tunic. During development tunicates form a tailed larvae that closely resembles the vertebrate body plan [**?**] and this tadpole larvae is typical of ~3000 tunicates [**?**]. Out of these 3000 species 16 are known to have independently lost their larval tail, with the majority of them being *Molgula* [**?** **?**] and only two tail-less species are found in the Styelidae [**?**]. During this time, known as the free-swimming stage, the elongation and mobility of the tail is depended upon the proper formation of the notochord and muscle cells [**?**]. As a tissue the notochord is closest related to cartilage and serves as the axial skeleton of the embryo in addition to a source patterning signaling [**?**]. In ascidians and in lower vertebrae the improper formation of the notochord leads to severely shortened larva that cannot swim or feed properly [**?** **?** **?**]. 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 and tail loss.

Ascidians are a simpler system to study developmental processes, their development is well studied, they have invariant early cell lineages, a small number of cells [**?**] and there has been no documentation of ascidians developing without an invariant cell lineage [**?**]. They also have rapid embryogenesis, compact genomes, few larval tissue types, simplified

larval body plans and shallow gene networks [**?** **?** **?**]. Although, this study present the first *Molgula* genomes assembled, we use the assembled and annotated genome of *Ciona intestinalis* which serves as the most documented and closest reference for the *Molgula* and other ascidian species [**?** **?** **?**]. In *C. intestinalis* there are 2,600 cells, 36 of them being muscle, 40 of them being notochord and many of these cells have be traced starting at fertilization [**?**]. For these reasons tunicates make great models for both tail development and loss, in addition to several Molgulids independently losing their tail and two of the Molgulids, a tailed and tailless species having the ability to hybridize [**?**]. Although *M. occulta* and *M. oculata* present great systems evolutionarily to study tail development and loss, they have several shortcomings as experimental models; they are only found on the Northern coast of France and have yet to be cultured, they only spawn for one month out of the year, and many of the molecular techniques used in other ascidians have not yet been optimized for these two species.

Genes have been identified by subtractive hybridization screening and microarrays [**?** **?** **?** **?**]. Sequence technology has continued to advance and become cheaper. Technology such as Ion Torrent, Roche 454 and Illumina has made genome or transcriptome wide analysis more readily available for non-model species. These technologies have several advantages over the prior standard mircoarray; they have a wider scope, are more precise and are able to find novel genes [**?**]. With the advances in technology when now sequenced the transcriptomes of both species and their hybrid. This allows us to look at pivotal time points in tail development and compare across closely related species. This type of study has yet to be done.

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. Next-generation sequencing (NGS)

presents a great method of producing observation and hypotheses to be tested experimentally. However, before we can make biological inferences from our data we have to produce a quality assembly. Because of this we first assessed the quality of an efficient low-memory assembly pipeline for our RNA-seq data and identify quality metrics other than N50 and contig length, seeing that they are not best metrics for assessing transcriptome quality [**?**]. 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 analysis the homology between ascidian gene networks, and build more complete transcript module for differential expression[**?**]. In the end we were able to do something really cool.

DELETE TEXT BELOW UNLESS I CAN THINK OF A GOOD/USEFUL WAY TO INCORPORATE(but definitely put in first sentence).

Our study expands on prior knowledge, observing gene expression on a wider scale, the microarrays were done with isolate of the 64-cell which does not look at the whole picture seeing that the notochord develops through interactions with muscle and boundary effects [**?** **?**].

Tail development has been previously studied in ascidians and other chordates, with no one factor being the cause of a improperly form tail or lack of tail.

We started this project with RNA-seq data which presented us with the problem of determining which assembly was the best and what metrics should be used to analysis them.

Experiemential techniques have yet to be adapted to *M. occult* and *M. oculata* because of there short gustation period, not being able to be cultured in lab conditions, although this is being currently developed amongst embryo specific difficulties. Most of the studies for tail development have been done in *C. intestinalis* and *H. roretzi*

Chapter 2

Literature Review

2.1 Ascidian tail development

The notochord is one of the most distinguishing characteristics of chordates. In their adult form ascidians and their vertebrate cousins have no resemblance, however during development they have similar body plans featuring the notochord [**?**]. Ascidians are known for the bilateral and invariant cell cleavage. Their development is well described up to the gastrulation stage [**?** **?** **?**]. Like vertebrate chordates such as xenopus ascidians depend on maternally localized determinants to regulate cell moments and division, however the location and identity of these determinants are different although the development of the early body plans are similar [**?**]. Solitary ascidians notochords typically come from two cell lineages, the primary notochord derive from the "A" blasomere and the secondary notochord comes from the "B" blastomere [**?**] which can be identified by the 4-cell embryonic stage. At the 4-cell stage the blastomeres are labeled in Conklin [**?**] convention; "a" and "A" for the anterior animal and vegetal blastomeres, respectively and "b" and "B" for the posterior animal and vegetal blastomeres, respectively. Although the notochords cells have been traced back to the 4-cell stage, notochord inductions 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—are identifiable and no longer multipotent, while the 2 secondary notochord cells are not restricted until the 110-cell stage [**?** **?** **?** **?**]. Two additional stages of cell

Figure 2.1: **Notochord cells.**The primary notochord cell (blue) also known as the A-lineage are specified at the 64-cell stage. There are a total of 32 primary notochord cell that come from the A7.3 and A7.7 blastomere, and the intercalation of the cell happen in a semi-random stochastic manner.The secondary notochord cells (red) comes from the B8.6 blastomere and are specified at the 110-cell stage, one cell division after the primary notochord cells.

division occur, one at gastrulation and one at neurulation, ending with 40 notochord cells, which is typical of most solitary ascidian tadpole larvae [**?**]. At the onsite of neurulation the notochord begins to form, this process includes 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 then the polarization and intercalate of the cells through a process known as convergence and extension[**?**]. 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 [**?**]. 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, but the ordering of the 32 most anterior is not determinate, cells from both the A7.3 and A7.7 intercalate in a random order (Figure 2.1)[**? ? ? ? ?**]. This process, along with muscle cell are the causes the larval tail to form [**? ? ?**].

Although a tailed larvae is typical of most ascidians, several species with in the Stolidobranchia order have individually undergone tail-loss, many of which fall in the Molgulidae [**? ? ? ?**]. The tail-less—anural—species develop in a similar manner and are indistinguishable from their tailed—urodele—counterparts up to late gastrulation [**? ? ?**]. Anural ascidians lack several urodele features including an intercalated and extended notochord, differenti-

ated muscle cells and the otilith sensory organ. The absence of differentiated muscles cells and intercalated notochord are the cause for the lack of tail in these species [**?** **?**]. The development of several tail-less species have been studied. *M. tectiformis* notochord cells do not divide again after the 10 precursor cells are formed and *M. occulta* stops dividing after 20 cells [**?**]. 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 [**?**]. 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 [**?** **?** **?**]. Seeing that most ascidians have tailed larvae and that the tail can be restored through the use of interspecies hybrids, the lack of tail has been shown to be a loss of function. *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 hybrid produces an embryo with 20 notochord cells like *M. occulta*, however the notochord cells converge and extent like *M. oculata* [**?**]. The ascidian tail has been shown to form in the presence of notochord and the absence of muscles cells [**?**] and the hybrid tail is not flaked by muscles as that of tail species [**?**]. However, also it has been shown that embryos that develop urodele features are batch specific, and only in embryos that express the p58 which is associated with cytoskeleton are urodele features restored [**?** **?**]. It has also been shown that in hybrid embryos in which urodele features were 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* [**?**]. This along

with evidence that with may have been, the ancestral notochord—the axochord—is muscle based [**?**], shows strong evidence for the need of both notochord and muscles cells for the formation of the ascidian tail.

2.2 Brachyury has been shown to be the

Brachyury a T-box transcription factor, has been identified to be essential for notochord development [**?**]. The notochord induction is regulated by the *FGF/MAPK/Ets* signaling cascade [**?**]. Where 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 [**?**]. It was observed from isolation experiments that notochord/nerve cord precursors that loss *FGF9/16/20* competence at the 32-cell stage assume the default nerve cord cell fate, the converse is true for presumptive nerve cord blastomeres that are introduce to *FGF*, they forgo their default nerve cord fate and choose the notochord fate [**?** **?**]. If *FGF9/16/20* is not present at the 32 cell stage competence is lost, *bra* is not induced and the notochord can no longer from [**?** **?**]. This is because *MAPK* is not activated and the induction of *bra* and repression of *FoxB* are not carried out [**?**]. Without the repression of *FoxB TF* the notochord cell fate is repressed through the repression of *bra*. It has been observed in *H. roretzi* that *FoxB* represses the activation of *bra* predominately through the binding of Fox BS1 (GCACTGA*ACAAACA*TACATAG). *FoxB* is activated by *ZicN* and present in both nerve cord and notochords precursors, however is repressed by *MAPK* in the notochord cell lineage at the 64-cell stage [**?**]. *MAPK* is thought to be repressed by *Ephin* which is one of the key differences between notochord and nerve cord determination. At this point *bra* is expressed first weakly in the at the 64-cell stage in the notochord/nerve chord precursors [**?**

] and unlike other chordates, in ascidians *bra* is only expressed in the notochord cells [**?** **?** **?** **?**]. 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 stop earlier than that of *M. oculata*, but produce different results. *Bra* is expressed in the 10 precursor notochord cells in *M. occulta*, another round of cell division occurs which does not in *M. tectiformis*. In these two species of *Molgula* muscle actin became pseudo genes, however the mutation in the muscle actin genes are not the same [**?** **?**]. *Manx* is another gene identified to be important for tell development in *Molgula*, and is lowly expressed in *M. occulta*, and has been shown to restore the hybrid tail [**?** **?**].

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 shown to cause a shortened ascidian tail affecting both the mediolateral intercalation and the elongation of the ascidan tail [**?**]. The *pk* mutant *aimless* produces a truncated tail, however the polarity of the nuclei are present, showing that prickle does not establish polarity with in the cell but polarity between cells, acting in a local manner and perhaps their is a global organizer [**?** **?**]. 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 effects [**?**].

Many of the upstream genes and transcription factors that interact with *bra* has been studied in fair detail, through known-outs, and cell isolation experiments. Not as much detail is known about the downstream genes regulated by *bra*. On larger scale subtractive screening was done to identify genes downstream of *bra*, 39 genes were initially found [**?**]. An attempt to characterize a number of these genes have been made, identifying functions such as extracellular matrix components (*cadherin 8, entactin, fibronectin, laminin alpha1,*

alpha4, and beta1, and thrombospondin, genes involved in cell shape and polarity (*pk, trop, ERM, ACL*), axon guidance (*netrin, semaphorin 3A*), amongst a host of other biological processes [**?** **?** **?**].

Oikopleura which are in the same subphyla—tunicates or urochordates—also develops in a typical chordate manner with a notochord and has a compact genome, however, in Larvacean there are 20 notochord cells [**?** **?**]. *Oikopleura* 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 [**?** **?**]. When comparing gene networks, *Oikopleura* did not exhibit the same mechanism for tail development as *Ciona*, of the 50 bra target genes previously identified cite only 26 of them had orthologs, almost 50% of the genes were not present [**?**]. Of the genes that did show homology, expression ranged from notochord specific to tail including possible notochord, to tissues that were clearly not the notochord.

2.3 Assembling and analyzing data

One of the major advances in science in the past 20 years was the implementation of sequencing technologies. These technologies allowed us to examine problems in ways not previously possible. The first wave was Sanger sequencing in the 1986, but was not broadly used until 10 years later. Another technology, mircoarrays, which became popular starting in the mid '90s, allow us to look at a wide spectrum of genes and understand relative expression within a sample. Kobayashi et al. [**?**] 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 is needed. *C. intestinalis* was sequenced using sanger sequencing, and is well assembled and most well annotated [**?**]. In addition to long reads (Sanger) scaffolding was done using experimental data[**?**]. Sanger sequencing able to sequence whole genomes without the need of prior knowledge to identify novel genes but is costly and time consuming[**?** **?**].

Sanger was the 1st generation of sequencing technologies, and currently both 2nd and 3rd generation are in use, with Roche 454, Ion Torrent, Illumina and PacBio are the most wide spread. These technologies are far easier to produce data and much less costly than sanger sequencing [**?**]. There are many trade-offs for each of the technologies, cost per MB, sequencing time, prep cost, error rate and sequencing bias; 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 better for counting, in addition to PacBio having a high error rate [**?**]. Illumina and Ion Torrent have the best error rates and while Ion Torrent calls more more Single nucleotide polymorphisms, it also calls more false positives. For this reason, amongst other Illumina is the most used because it is the most versatile and preforms the best in general [**?**]. This drop in price and produced many of the assembled genomes within the Tuncata phyla. Outside of this project there are eight tunicate genomes assembled; *C. intestinalis*, *C. savignyi*, *Oikopleura dioica*, *Botryllus schlosseri*, *Halocynthia uranium*, *H. foretzi*, *Phallusia fumigata*, and *P. mammilata*, but no *Molgula* genomes.

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 [**?**], allowing us to study organisms that are ecologically or evolutionarily interesting.

There are now several sequencing technologies, Illumina being one of the most versatile [**?**], that can produce millions of short reads ranging from 75 to 150 bp in length at a low cost [**?**]. 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 [**?** **?** **?**]. It has been shown in some cases that mapping mRNA-seq reads to a reference genome yields better transcriptomes than *de novo*

assemblies, even if the genome is 5-15% divergent [**?**]. However, with many non-model organisms, no nearby reference genome is available.

De novo transcriptome assembly is the only solution for organisms with no evolutionarily close reference genome. Transcriptome assemblers such as Trinity [**?**] and Velvet/Oases [**? ?**] 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 [**?**].

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 essentially 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 [**?**], 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 [**?**] and culturing [**?**] 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 [**?** **?**]. 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 [**?**]. 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 are 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 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 performed well on other data sets [**? ? ?**]. 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>" [**?**].

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 [**?**]. 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

[**?**] 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 3.1); 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 GB of RAM versus 34.14 GB of RAM for the unnormalized reads (Figure 3.1).

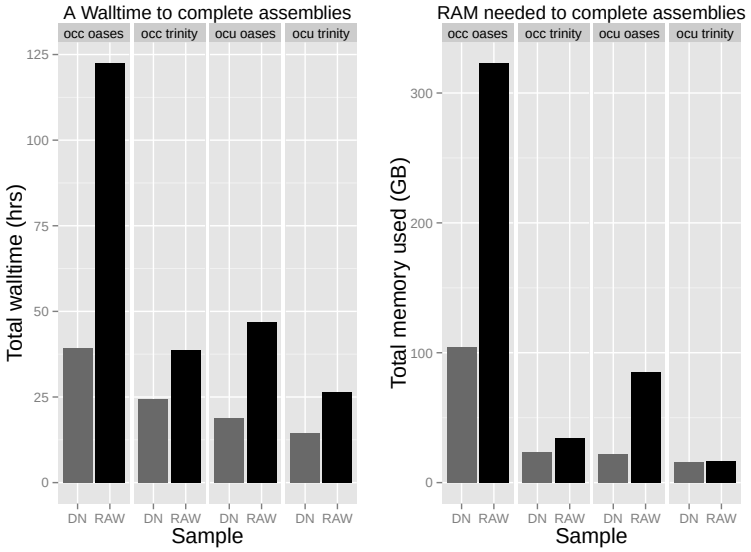


Figure 3.1: **Wall time and memory requirements for assemblies.** Wall time (left) in hours to complete the diginorm (DN) and raw read (RAW) assemblies for both species and assemblers. Oases assembled multiple k's, $21 \leq k \leq 35$ opposed to Trinity that uses only a single k. This is one reason the assembly times differed. (right) 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.

Table 2: Assembly Statistics

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.

3.3.2 Assembly statistics varied by preprocessing approach and assembler

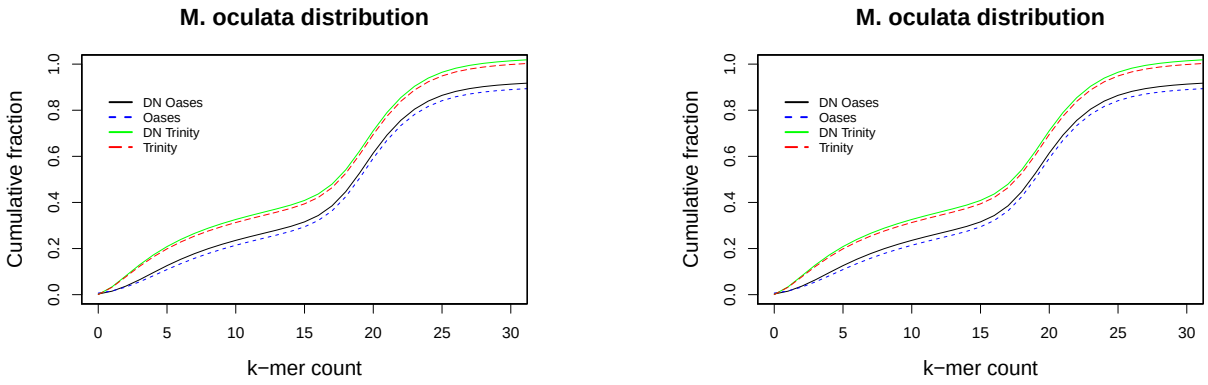
Oases run with the diginormed reads yielded fewer total transcripts than Oases run with the unnormalized reads. The *M. oculata* diginorm assembly produced 300 fewer transcripts, and the *M. occulta* diginorm assembly produced 227 fewer transcripts (Table 3.2). Digital normalization had the opposite 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 [**?**]. 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 is incorporated in the respective assemblies. Both Oases and Trinity assemblies are shown for **??** *M. occulta* k-mer distribution and **??** *M. oculata* k-mer distributions. Trinity had a higher k-mer distribution for both species, reflective of the inclusion of more low abundance reads into the Trinity assemblies.

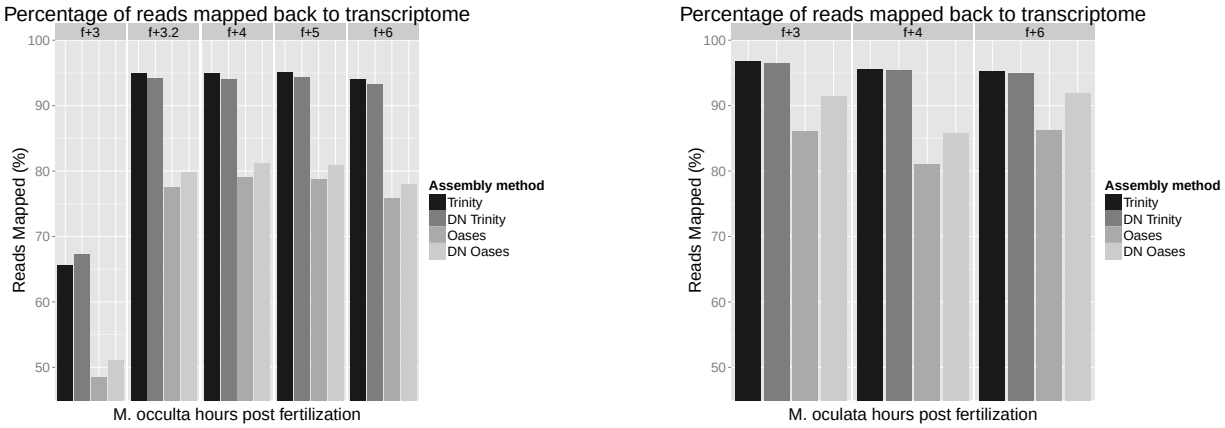
Table 3: K-mer multiplicity

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**??**). 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 raw-reads assembly differ from the diginorm assemblies in more than 3% of their read content.



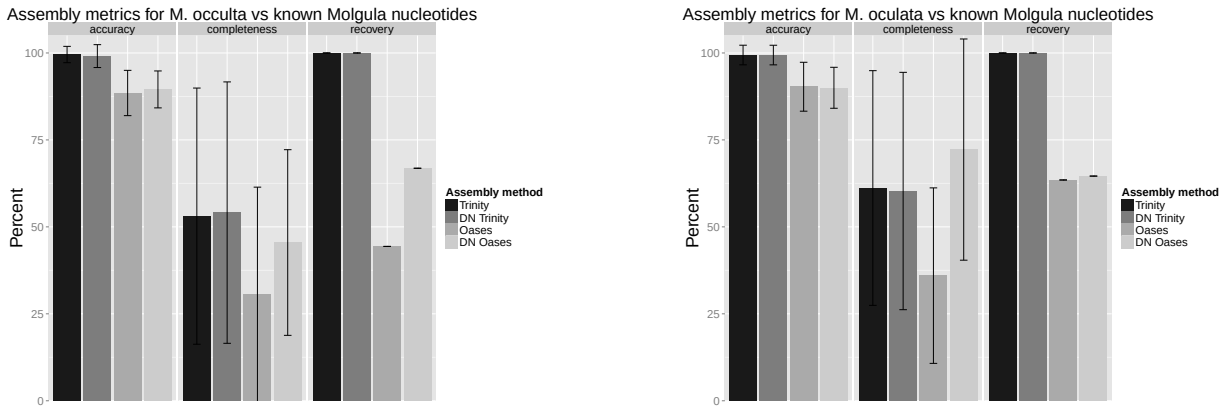
(a) *M. occulta* (b) *M. oculata*

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 were 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 Trinity assemble 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 [**?** **?**]. 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 [**?**]. 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, not coding, so we would not expect high completeness.)



(a) *M. occulta*

(b) *M. oculata*

Figure 3.4: **Accuracy, completeness and recovery rate against know Molgula sequences.** The NCBI has 178 Molgula sequence in its database. Transcripts were searched against these sequences using BLASTN with a cut-off of e-12. Trinity assemblies performed the best, recovering all known sequences. *M. occulta* unnormalized assembled 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.

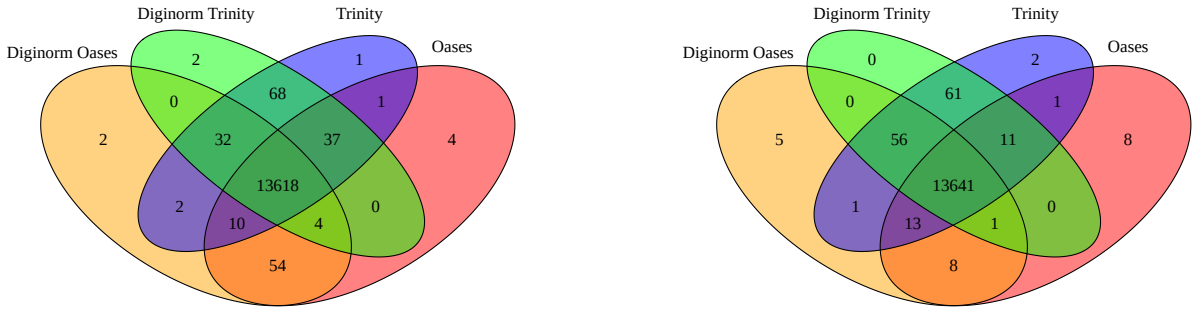
3.3.6 Both unnormalized and normalized assemblies recovered many of the same transcripts

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 [**?**]. 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 homologue with *C. intestinalis* via BLAST for *M. occulta* (left) and *M. oculata* (right). Each oval represent the total number of homologs 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 [**?**]. 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 [**? ?**]—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 [**?**].

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 [**?** **?**]. 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 [**?**]. 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 [**?** **?** **?**]. 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.

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

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 phylogentic position—urochordates—the sister group to vertebrates and cephalochordates, together they form the chordate phylum. Although ascidans share no resemblance to vertebrates in their adult stage, they share several features, a notochord, dorsal hollow neural tube, and gill slits during development in addition to 18S placement [**?** **?**]. The development of ascidians are well documented, the cell lineage from fertilization to gastrulation has be followed in *Ciona intestinalis* [**?** **?** **?**]. Studies of other ascidians species have shown that the majority of phyla has an invariant cell lineage and typical development [**?**]. Only very few solitary ascidians have deviated and undergone tail-loss [**?** **?**]. *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 found between the siphons of *M. oculata* (Figure 4.1). These two *Molgula* species, however, have different methods of development—*M. oculata* developing a typical tadpole larvae and *M. occulta* developing without a tail.

Many of the genes have been studied across a number of ascidian, showing that gene function tends to be orthologous within the phyla [**?**]. Although genes tend to be expressed in homolguous patterns and tissues, the presence of genes are not the same across species. There are a number of case where a gene that has been shown to be necessary for a phenotype is completely absence an other ascidian species [**?**]. It has also been shown that in ascidians with the same phenotype and gene expression, regulatory modules are not necessarily the same [**?** **?** **?**]. Ascidian species are far more divergent than the appear phenotypically. Urochordates have even deviated from *hox* patterning and function [**?**], genomics has shed some light on the area. Ascidians are board cast spawners, which leads to them being highly polymorhpic and having rapid rates of evolution [**?**]. This is the cause fairly divergence genomes outside of coding regions, and change of gene function when compared to other chordates [**?**]. We will demonstrate this divergence using two closely related species *M. occulta* and *M .oculata*, and the more divergent *M. occidentalis*. Whole genome sequencing and assembly has given us a better picture of what is going on evolutionary for close and divergent species and allow us to characterize gene networks, identify regulatory elements and get a better understanding of mechanism of development in ascidians.

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). Sequenc-

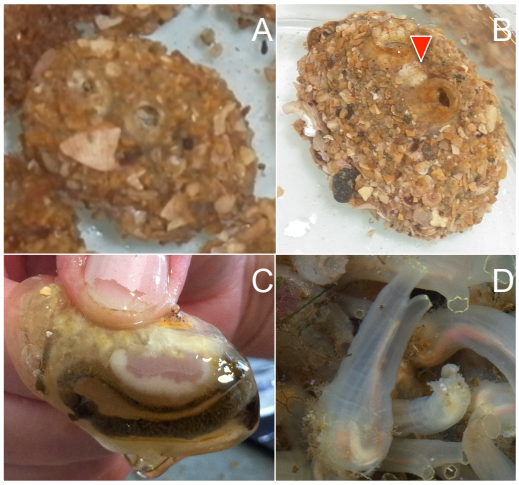


Figure 4.1: **Adult ascidians.** *M. occulta* (A) and *M. oculata* (B) are nearly identical in their adult stage with the white pigment spot (red arrow). Their tunic is covered in sand, seeing that they are found on the sandy sea bottoms. Under there sand covered tunic, the two species differ by the color of their eggs—purple in *M. oculata*, pictured and an orange-yellowish color in *M. occulta*—found just above the kidney complex (C). *C. intestinalis* (D) is one of the more studies ascidians and has a assembled genome.

ing 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 100 100 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 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 [**?** **?** **?**]. Genome assembly was done using velvet v1.2.08 [**?**] 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 [**?**]. 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/ [**?**]. Scripts for genome assembly and CEGMA analysis can be found in the following github repository: https://github.com/elijahlowe/molgula_genome_assemblies.git

4.2.3 Gene identification and alignments

Thirty-nine hox genes were identified in hunman and downloaded from the NCBI database. These sequences were then BLAST against each of the three assembled *Molgula* genomes. The alignments were then extracted and BLAST against the NCBI non-redundant database. *Molgula* alignments 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/elijahlowe/el /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 [**?** **?**]. Additional alignments between the three species were conducted using mVista [**?** **?** **?**] 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[**?**].

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 [**?**]. CEMGA reports scores for complete and partial alignments to a subset of core eukaryotic genes. An alignment

Table 1: Assembly Statistics

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 scores is a metric of completeness measured against highly Conserved eukaryotic genes. Alignments of 70% or greater of the protein length are called complete (C[1]) and all other statistically significant alignments are called partial (P[2]).

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 [**?**]. 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, [**? ? ?**].

4.3.2 Gene complexes

Hox genes are as subset of the homeobox genes and known to be involved with the establishment of morphological identities along the anteroposterior axis of bilaterians and cnidarians [**?**]. All *hox* genes have a highly conserved 60 amino acid (aa) homeobox sequence [**? ?**]

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* [**?**]. The *hox* gene of *C. intestinalis* where 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 [**?**]. *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 [**?**].

Eight *hox* genes have be found in *M. occulta* and *M. oculata*, while nine have been found in *M. occidentalis*. *Hox1*, *hox2*, *hox3-4*, *hox5*, *hox10* and *hox12-13*, with *hox3-4* 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*, *hox10* is downstream of *hox12-13*. *M. occidentalis* had one additional *hox* gene compared to *M. occulta* and *M. oculata*, there appears to be a duplicate *hox10* gene ~12kb apart found on the same contig. The second *hox10* sequence was not fully sequenced, missing 14 aa of the homeobox domain, the identify of the two sequences were 52.1% at a nucleotide level, 53.4% at a protein level, and 91.3% identical within the homeobox domain (Figure A.3). *M. occulta*, *M. oculata*, and *M. occidentalis hox* genes span across 7, 6 and 5 contigs respectively and spends xxx kb, 333 kb and sass kb, respectively. This show to be more compact that *Ciona* which exhibits longer and than usual introns between *hox* genes, averaging in the 5Mb range, when typically the *hox* genes have 100-120 kb separating them [**?**]. spends $3x$ as many base pair, but these show to be 3x that of human. *Mocci.hox2* has a stop codon located in the 3-4 helix.

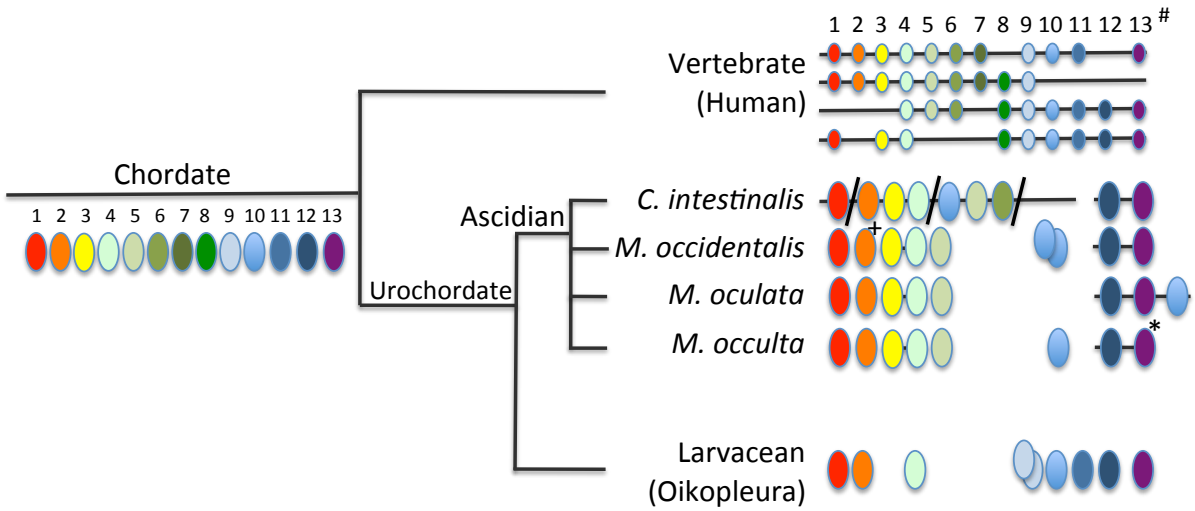


Figure 4.2: **Hox clusters for *M. occulta*, *M. oculata* and *M. occidentalis*** Eight *hox* genes were found in *M. occulta* and *M. oculata*, while nine were found in *M. occidentalis*. *Hox1*, *hox2*, *hox3-4*, *hox5*, *hox10* and *hox12-13* were found in all three *Molgula* species. *Hox3-4* were found on the same contig in all species, with *hox12-13* being found on the same contig in *M. occidentalis* and *M. oculata*. **M. occulta hox12-13* are not found on the same contig, but when aligned using mVista, this is high sequence similar, showing the possible placement for 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, rearrangement has be found in *Ciona* and *Molgua*.

4.3.3 Divergence of GRN

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., [**?**] we used BLAST to identify the *Molgula* orthologs of *C. intestinalis Mesp* (*Ciinte.Mesp*, as per the proposed tunicate gene nomenclature rules, see Stolfi et al., [**?**]). *Ciinte.Mesp* is the sole ortholog of vertebrate genes coding for *MesP* and *Mesogenin bHLH* transcription factor family members [**?**]. VISTA alignment shows high sequence similarity between sequences 5' upstream of the *Mesp* genes from the closely related *M. oculata* and *M. occulta*. 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 [**?** **?** **?**]. Our sequencing results support the view that *M. occidentalis* is highly diverged from other *Molgula* spp.

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 coding sequences showed similarity in *M. occidentalis* when compared *M. oculata* (Figure 5.2). 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*, which stains endodermal strand cells in *Ciona* downstream of *Hox13*.

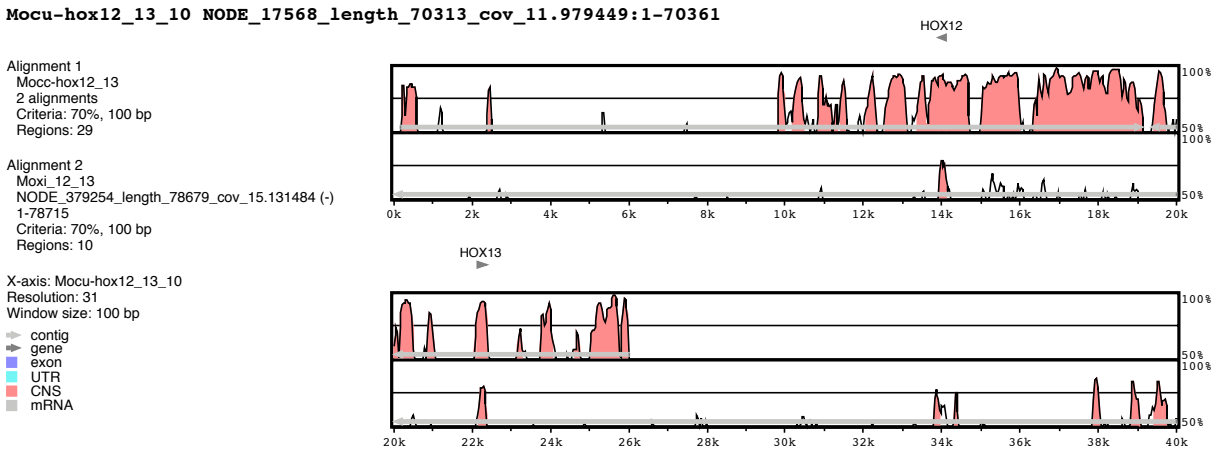


Figure 4.3: **Alignment for *hox12-13* in *M. occulta*, *M. oculata* and *M. occidentalis*** The contig containing *hox12-13* for *M. occidentalis* and *M. oculata*, along with the two contigs containing *hox12* and *hox13* for *M. occulta*. *M. oculata* was used as the anchor sequence because it showed the most similar 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 similar in coding regions. Grey arrows show the direction of the contig.

4.4 Discussion

Three *Molgula*—*M. occulta*, *M. oculata* and *M. occidentalis*—species have been sequenced and assembled, these are the first of any molgulides 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 develops without a tail [**?** **?**]. In vertebrate and other bilitarians the *hox* genes has shown to be important for patterning along the anterior-posterior axis [**?** **?**]. The same has not be shown in ascidans, *hox* has more of a tissue specific role [**?**]. *Ciona* has 10 *hox* genes and is missing *hox7-9* and *11*, with *hox10* and *12* being the only 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 section of the tail and of the epidermal cells at the tail tip [**?**]. 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 express is detected in *C. intestinalis* through Whole Mount In Situ Hybridization at any stages of development [**?** **?**]. No two of the *Molgula* species show the same *hox* pattern and show a strong divergence outside of coding regions, even more so in *M. occidentalis*. There is a duplicate in *M. occidentalis hox10* which could lead to a split in function seeing that *Ciona Hox10* is expressed in two region during the mid-tailbud stage—a small region of anterior the nerve cord, and a small area of the posterior ventral endoderm and adjacent tissue [**?**]. It is proposed that ascidians evolved there simple body plans and rapid embryogensis through extensive genomic rearrangement and gene loss, some of those genes being hox genes [**?**]. *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 5.2), so its possible they are clustered, but the contigs are not joined because the genome is more fragmented.

Because of the lack of homology outside of coding regions, we were able to identify *distal-less* (*dll*) location downstream of *hox13* in both *M. occidentalis* and *M. oculata*, this was not the case for *M. occulta*. In *Ciona dll* stains in the endodermal strand cell during early-tailbud embryos, showing positive signals in two cells of the endodermal strand [**?**] which derives from the primary muscle or mesenchyme lineage [**?**], in Drosophila embryos *dll* is required for the gene pathway for limb formation in the thoracic segment [**?**].

4.5 Conculsion

Hox genes function is not conserved within the urochordates, and is coordinated with gene loss. All central *hox* genes are miss in all three *Molgula* species. The lack of having an overall organizer has made the developmental process more roboust [**?**]

Chapter 5

Tail loss?

5.1 Introduction

Ascidians are an interesting system to study because of the relationship to the vertebrates. They form a tailed larvae with a hollow dorsal notochord before undergoing the process of metamorphosis. A typical ascidian larvae tail forms form through the convergence, intercalation and extension of the notochord. When fully formed the ascidian notochord contains 40 cells, larvavearn notochord form in a similar manner, however only contains 20 cells. The ancestral notochord or notochord-like structure is believed to have been muscle based. This is perhaps the reason behind the tail formation being tied to both notochord and muscles, which come from the same cell lineage. Both the primary and secondary notochord and muscle linOf the ~3000 species of ascidians less than 20 have been identified as undergoing tail-loss. Although each case of tail loss has happened independent of each other, many of the species tend to fall in closely related clades and many of them are Molgula. M. occulta, M. blezi . Although the mechanism behind tail-loss differs by species, a common characteristic is the lack of a notochord that intercalates and extends. M. bliezi notochord cells converge to the midline, and began to extend, however, cells never properly intercalated and the tail formation stop before it is fully formed. One reason behind this is the early down-regulation of *bra* and another is the muscle actin becoming pseudo genes. It is thought that the early form of the notochord where not cartilage based but were made of muscle.

In their adult form the two species 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 indistiguisable up to the gasturalation stage. It is at late gastrula when the notochord and muscle cells begin to move posteriorly. There are several steps that take place to form the notochord and tail. First the notochord cells moveolaterally to the midline. Next the cells polarize and intercalate, changing their shape and extending posteriorly. this process is known as convergence and extension. genes necessary for tail formation that is missing from M. occulta has been found in other ascidians, for instance, macho-1 has been found the tail-less M. tectiformis. —- With advances in high throughput sequencing technologies, gene expression of M. occulta, M. oculata, and hybrid species can be analyzed (Gyoja et al. 2007; Pickrell et al. 2010). The transcriptomes of three different developmental stages of M. occulta, M. oculata, and hybrids have been sequenced at Michigan State University. The three transcriptomes were used to identify the presence or absence of known notochord genes downstream of bra using C. intestinalis data from the NCBI database. BLAST searches were with known notochord genes, and several of them were selected for further analysis. FGF9/16/20, prickle (pk), and several other downstream brachyury factors?noto6, leprecan, merlin, and noto17?were analyzed for presence, temporal and spatial expression using in situ hybridizations. In addition to focusing on the notochord genes an EdgeR differential expression analysis was done to identify other genes that are involved in tail development.

Simple body plan with a small number of cells [**? ? ?**], rapid embryo development.

Induced at the 32 cell stage by *FGF9/16/20* muscle and notochord come from the same cell lineage. Hybrids are tails are resorted in embryos that contain p58, which stains in the muscle

—-

5.2 Methods

5.2.1 Sample collection, sequencing and assembly

DNA was extracted from the gonads of an individual adult specimen for M. occidentalis, M. occulta, and M. oculata. Paired in jumping libraries were collected for each sample ranging from ~300bp to ~950bp. further details about extraction methods and libraries can be found in Stolfi et al., []. RNA was extracted from all three molgula species using the methods discussed in Lowe et al., []. Sequencing for M. occulta and M. oculata was conducted at the Michigan State University

Genome assembly was conducted using 3-pass digital normalization[] and assembled using Velvet[]. Assemblies were done with 21 Both de novo and reference based assembly were used when creating gene models. reads were mapped to their respective genomes using bra and top hat to identify genes and alternative splicing variants. the accepted.bam hits ere then sorted and indexed using samtools. the sorted bam files where then processed using cufflinks and cuff merge to generated consenus gtf annotation files. the de novo assembled transcripts from Lowe et al., were aligned to their respective genomes using BLAT. the cufflinks/cuffmerged gtf files were then covered into bed file and the read mapped annotation and de novo assembled aligned annotation files were merged using gimme. Gimme joins gene models using a graph based method to develop more complete transcripts. the gimme gene models were then covertedd to gff format using the crript bed2gff in the gimme utils folder in order to extract the transcripts from the genome in a multi pasta file. transcripts were then partitioned into transcript families using khmer partitioning tool.

5.2.2 Gene counts and differential expression analysis

reads are mapped to transcripts from the gimme gene modules for their respective species. Hybrid reads are mapped onto the M. occulta and M. oculata genomes as well, seeing that they are F1 hybrids and should contain an allele from each parent. Read counts were generated using express []. Effective counts, which normalizes counts based on transcript length were used because transcript length will differ across species. A replicate was only provided for one of the samples, 3hpf, because of this various dispersion calculation methods were used. the one replicate was used to calculated statistics, in addition to 5hpf being treated as a replicate for 6hpf. These time points represent what would be early and mid-tail bud stages in the urodele ascidian.

Notochord genes have been identified using subtractive hybridization, mircoarrays and

5.3 Results

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

C. intestinalis is the closest ascidian species with a well annotated gene, because of these reason it was used to annotate the genomes of both *M. occulta* and *M. oculata*.

—

5.3.2 Notochord gene network

Reciprocal best hit (RBH) blast with an e-value of 1e-6, were done with the M. occulta and M. oculata transcriptomes against C. intestinalis for the annotation of each species.

M. occulta and M. oculata have a high overlap in number of translated transcripts that showed homology with C. intestinalis proteins from the NCBI database. Of the 16 thousand proteins found in the NCBI database, both molgulid species recovered 84% of the proteins. M. occulta had an additional 202 transcripts that were not found in M. oculata and M. oculata had an additional 250 transcripts that did not have hits in M. occulta, overlapping by 97%. Next, we examined genes associated with notochord development in C. intestinalis to better analyze the molecular development of the tail. Seventy-two genes identified as being involved in notochord development (Kugler et al 2008; Hotta et al 2000; Kugler et al. 2011; Jose?-Edwards 2011) were BLAST against each species. Four genes?not1, not4, not5, and snail?were not found in either of the species. This could have been due to true absence or sequence divergence in the transcripts. The remaining 68 genes were shared by both species with the exception of col5a, which was missing from M. occulta. Several genes found in the Ciona notochord GRN were not discovered in either M. occulta or M. oculata. Noto1, noto5, noto14, noto16, and ... were all missing from the transcriptomes of M. occulta and M. oculata. Additionally Kihf5 was not found in M. oculata but showed homology in M. occulta. M. occulta did not have 3 genes that were found in M. oculata, netrin, noto9 and something.

5.3.3 Prelliminary results: EdgeR differential expression analysis by stage

Gene expression at the gastrula stage does not show a great deal of fold-change between M. occulta, M. oculata and the hybrid. The majority of the genes had a fold-change of less than 5 fold, and most of the genes clustered inside of that 5-fold window (Figure 3a).

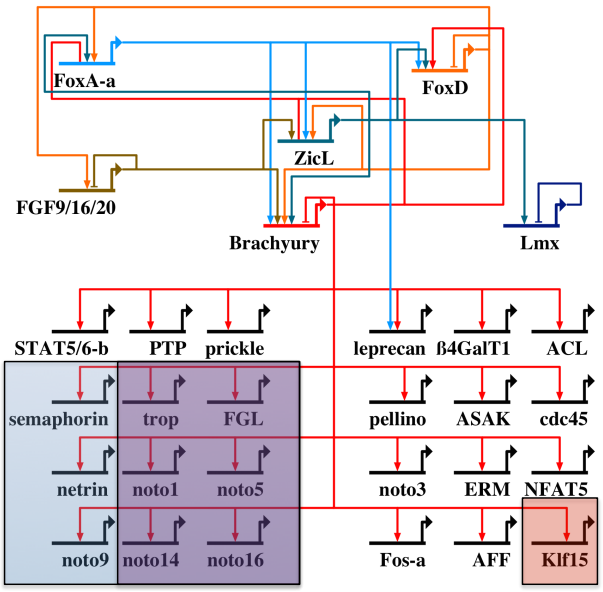


Figure 5.1: **Hox cluster for *Hox 10, 12-13* in *M. occulta*, *M. oculata* and *M. occidentalis***

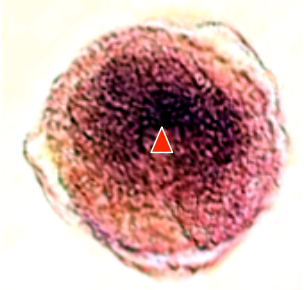


Figure 5.2: **Hox cluster for *Hox 10, 12-13* in *M. occulta*, *M. oculata* and *M. occidentalis***

The fact that the majority of the genes do not show significant fold-change is not surprising because M. occulta and M. oculata are very similar at this stage of development (Swalla and Jeffery 1990). When comparing each species to the hybrid, most of the genes clustered within a 5-fold-change window. During neurulation M. occulta versus M. oculata differential expression pattern is similar to gastrula stage, clustering within a 5 fold-change window. However, there is a drastic different in the fact that 300 of the genes have a 10 fold-change in expression (Figure 3d). The change in expression between the two species mirrors what was observed in morphological studies (Swalla and Jeffery 1990). M. occulta has normal urodele?tailed?development up to gastrula and begins to diverge at neurulation. Of the genes that were higher expressed in M. oculata compared to M. occulta, those same genes were higher expressed in hybrids versus M. occulta. Those genes do not show significant differential expression when comparing M. oculata to hybrid. There are 20 transcripts that show a 10 fold-change increase in expression in hybrid compared to M. oculata.

Chapter 6

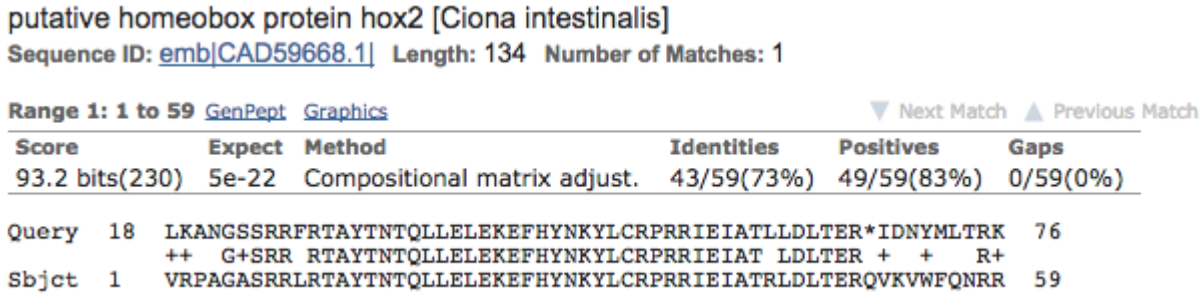
Conclusions

APPENDIX

ascidian notochord tunicate NGS rna-seq de novo assembly mapped based assembly urdele anural M. occulta M. oculata M. occidentalis C. intestinalis

Appendix A

Supplemental figures


putative homeobox protein hox2 [Ciona intestinalis]
Sequence ID: emb|CAD59668.1| Length: 134 Number of Matches: 1

Range 1: 1 to 59 GenPept Graphics ▼ Next Match ▲ Previous Match

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.1: **Alignment of *M. occidentalis hox2* genes alignment with *Ciona* show premature stop codon.** Two copies of *hox10* were found in *M. occidentalis* ~12 kb apart on the same contig.

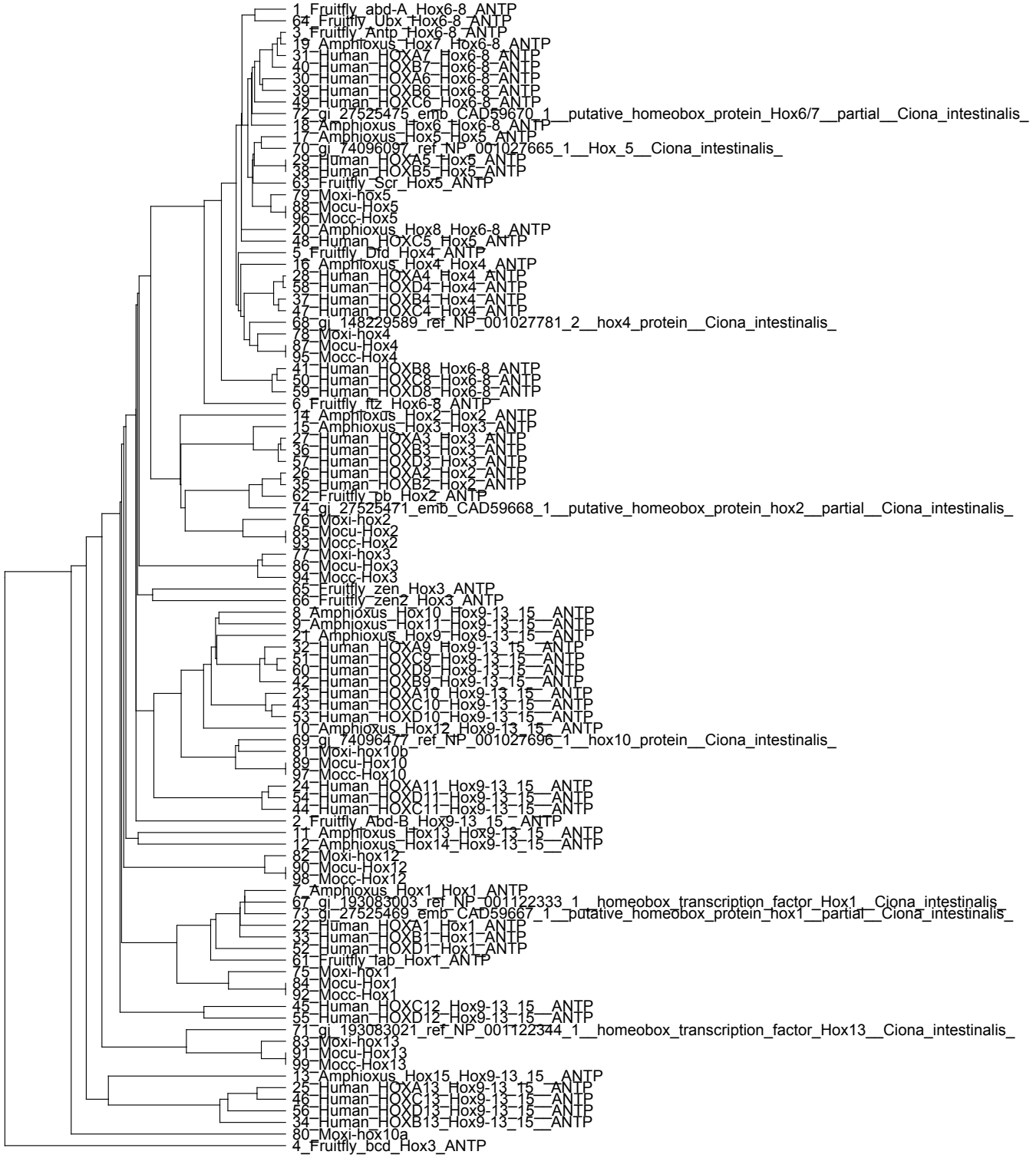


Figure A.2: **Alignment of *hox* genes.** Two copies of *hox10* were found in *M. occidentalis* ~12 kb apart on the same contig.

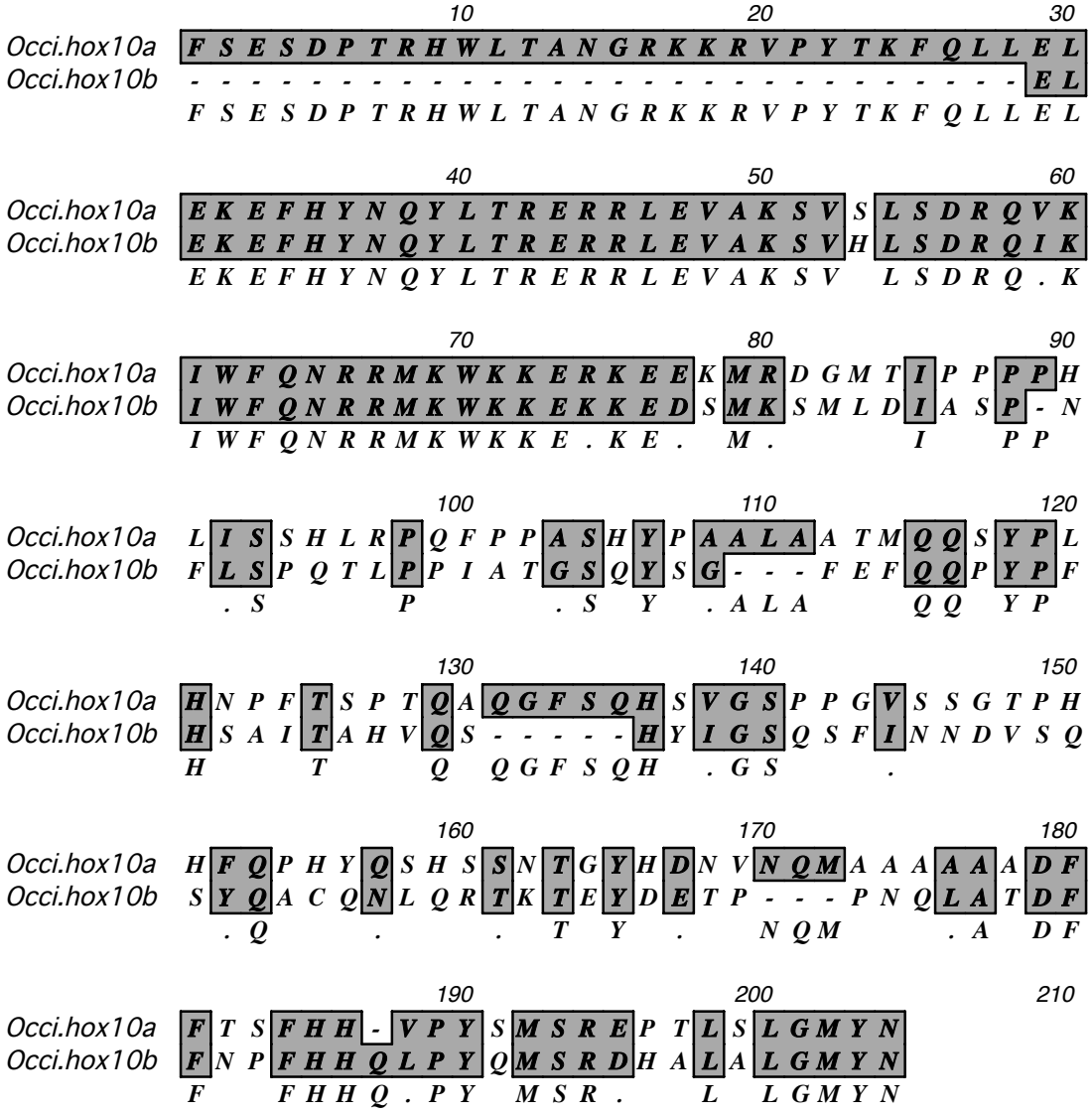


Figure A.3: **Alignment of *M. occidentalis* duplicate *hox10* genes** Two copies of *hox10* were found in *M. occidentalis* ~12 kb apart on the same contig.