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 [14] and this tadpole larvae is typical of ˜3000 tunicates *cite*. Out of these 3000 species 16 are known to have independently lost their larval tail, with the majority of them being *Molgula* [1, 49]. 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 [40]. 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 [16]. In ascidians and in lower vertebrae the improper formation of the notochord leads to severely shortened larva that cannot swim or feed properly [6, 17, 46]. 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 [21] and there has been no documentation of ascidians developing without an invariant cell lineage [22]. Although, this study present the first *Molgula* genomes assemblied, there the assembled and annotated genome of *Ciona intestinalis* which serves as a In *Ciona 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. Tunicates have a small number of cells compared to vertebrates, they also have rapid embryogenesis, compact genomes, few larval tissue types, simplified larval body plans and shallow gene networks [4, 14, 5]. For all of 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 [15].

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 for find novel genes []. With the advances in technology when can now sequence the transcriptomes of both species and their hybrid. This always use to look at pivotal time points in tail development and compare across closely related species. This type of study has yet to be done. Genes have been identified by both hybridization and subtractive screening (hotta, swalla, satoh, etc).

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.

Tail development and lost has been studied on We started this project with the mRNA of the two Molgula species and their hybrid. mRNA was collected for three time points in all

genes have been idenitified in C. intestinalis, holycythia rorzi using substrative methods and microarrays. In molgula a global view has yet to be examined. However, several genes have been identified, using the tailed, tail-less and the hybrid.

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

Ascidians are known for the bilateral and invariant cell cleavage. Their development well described up to the gastrulation stage[34, 35, 33]. 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 [22]. In ascidians the first cell division is coordinated by β-catenin which activates the vegetal gene and restrict GATA4/5 [] and determines the axis of division []. The notochord is one of the most distinguishing characteristics of chordates. Solitary ascidians notochords typically come from two cell lineages, the primary notochord coming from the "A" blasomere and the secondary notochord comes from the "B" blastomere [34]. At this stage the blastomeres are labeled in Conklin [3] 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 this point, notochord inductions does not occur until the 32-cell stage, where the notochord/nerve chord precursors are activated by fiber growth factor (FGF) and without FGF activation the cells lose competency and the notochord can no longer from [32, 31]. By the 64-cell stage there are 10 notochord cell precursors, the 8 primary precursor notochord cells are identifiable and no longer multipotent, while the 2

secondary notochord cells are not restricted until the 110-cell stage [35, 60, 61, 20]. Two addition stages of cell division occur, one at gastrulation and one at neurulation, ending with 40 notochord cells, which is typical of most solitary ascidian tadpole larvae [3]. 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 polarization and intercalate mediolaterally to the midline through a process known as convergence and extension where the cells [47]. 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 [35]. 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 [34, 35, 30, 47**?**]. This process, along with muscle cell are the causes the larval tail to form [30, 13, 47]. 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 [1, 16, 12, 27]. The tail-less—anural—species develop in a similar manner and are indistinguishable from the tailed—urodele—counterparts up to late gastrulation [1, 49, 13]. Anural ascidians lack several urodele features including a converged and extended notochord, 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 [30, 49]. *M. tectiformis* notochord cells do not divide again after the 10 precursor cells are formed and *M. occulta* stops dividing after 20 cells [16]. The same occurs in *M. bleizi*, however after the 20 notochord cells are formed, the embryo attempts to make a tail but never does so [52]. 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 [16, 53, 46]. 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* (urodel) and *M. occulta* (anural) 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, and no other crosses are known to produce hybrids. 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* [49]. The ascidian tail has been shown to form in the presence of notochord and the absence of muscles cells [30] and the hybrid tail is not flaked by muscles as that of tail species [52], however in hybrids embryos that express the p58 which is associated with cytoskeleton develop urodele features. Hybrid embryos that develop urodele features are batch specific and features are only restored in embryos that express p58 [48, 13]. It was also 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 [15].

2.2 Brachyury has been shown to be the

The induction of the notochord begins at the 32 cell stage by fibroblast growth factor (FGF) in the A6.2 and A6.4 notochord/nerve cord precursors[39] after the 7th cleavage. FGF transducer FGF receptor, Ras, MEK and MAPK. MAPK promotes Ets which promotes

Bra at the 64 cell stage. It was observed from isolation experiments that notochord/nerve cord precursors that loss FGF competence at the 32 cell stage assume the default nerve cord cell fate [29]If FGF is not present at the 32 cell stage competence is loss and *bra* is not induced. This is because is be MAPK which is down stream in the cascade is not activated and the induction of bra and repression of FoxB are not carried out [10]. And in the absence of bra notochord cell become nerve chord cells (Yasuo and Satoh 1998 Consevation of the developmental role of bra).As stated above the notochord is specific at the 64 cell stage. At this point *brachyury* is expressed first weakly in the at the 64-cell stage in the notochord/nerve chord precursors [60] and unlike other chordates, in ascidians *bra* is only expressed in the notochord cells [59, 4, 11, 53]. Without *bra* the ascidian tail does not form. 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 [52, 16]. *Manx* is another gene identified to be important for tell development in *Molgula*, however, not in all ascidians. *Manx* is lowly expressed in *M. occulta*, and has been shown to restore the hybrid tail, but there is no homolog for *manx* in *C. intestinalis* [50, 51].

It was shown in *H. roretzi* that *FoxB* is 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 [10]. MAPK is thought to be repressed by Ephin which is one of the key differences between notochord and nerve cord

determination. Ephin and FoxB have redundant roles in the repression of the notochord fate, but differ in that ephin is spatial and FoxB is mediates temporal restriction of *Bra* induction.

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[17]. 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 [17**?**]. 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" [55].

On larger scale subtractive screening was done to identify genes downstream of *bra*, 39 genes were initially found.

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 in Oikopleura [18] of those genes expression ranged from notochord specific to tail including possible notochord, to tissues that were clearly not the notochord.

——

there are 3 major pathways in chordates: FGF, BMP and Nodal.

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 were Sanger sequencing in the 1986, but was not used until 10 years later. and mircoarrays which became popular starting in the mid '90s. Mircoarrays allow us to look at a wide spectrum of genes and understand relative expression within a sample. Sanger sequencing allowed us to sequence whole genomes and

Outside of this project there are four tunicata genomes assembled; *C. intestinalis*, *C. savignyi*, *Oikopleura dioica*, *Botryllus schlosseri*, *Halocynthia uranium*, *H. foretzi*, *Phallusia fumigata*, and *P. mammilata*, but no *Molgula* genomes. Of those genome *C. intestinalis* is the best assembled and most well annotated. In addition to long reads (Sanger) scaffolding was done using experimental data.

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.

Currently both 2^{nd} and 3^{rd} generation technologies are in use, with Roche 454, Ion Torrent, Illumina and PacBio are the most wide spread. 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 [7**?**]. This drop in price and produced many of the assembled genomes within the

Tuncata phyla.

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

There are now several sequencing technologies, Illumina being one of the most versatile [7], that can produce millions of short reads ranging from 75 to 150 bp in length at a low cost [63]. 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 [38, 57, 45]. 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 [56]. 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 [8] and Velvet/Oases [62, 42] 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 [38].

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 [9], 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 [49] and culturing [51] 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 [25, 26]. 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 [2]. 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 [56, 8, 42]. 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>" [24].

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 [19]. 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

[23] 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).

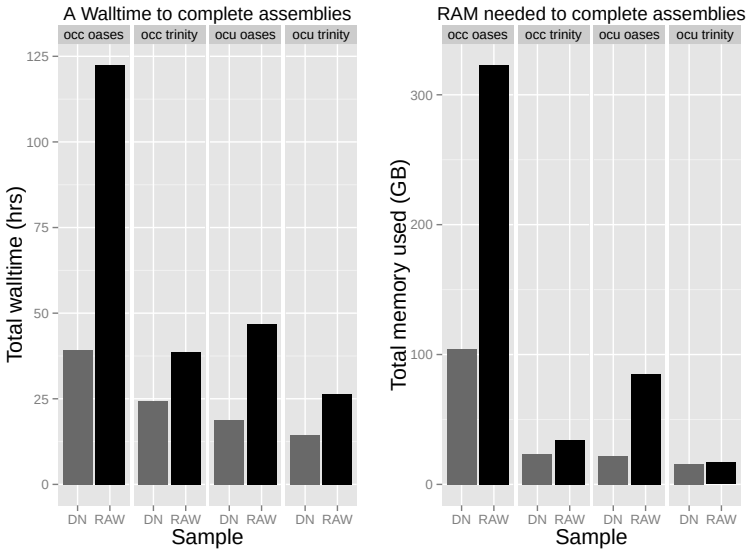


Figure 3.1: **Wall time and memory requirements for assemblies.** Wall time (left) in hours to complete the diginorm (DN) and raw read (RAW) assemblies for both species and assemblers. Oases assembled multiple k's, $21 \leq k \leq 35$ opposed to Trinity that uses only a single k. This is one reason the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

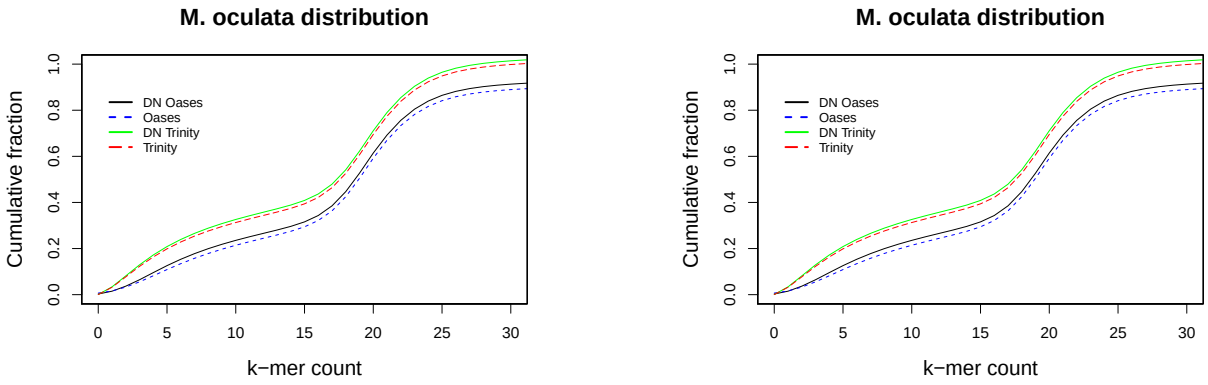
Oases run with the diginormed reads yielded fewer total transcripts than Oases run with the unnormalized reads. The *M. oculata* diginorm assembly produced 300 fewer transcripts, and the *M. occulta* diginorm assembly produced 227 fewer transcripts (Table 3.2). Digital normalization had the opposite 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 [38]. 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

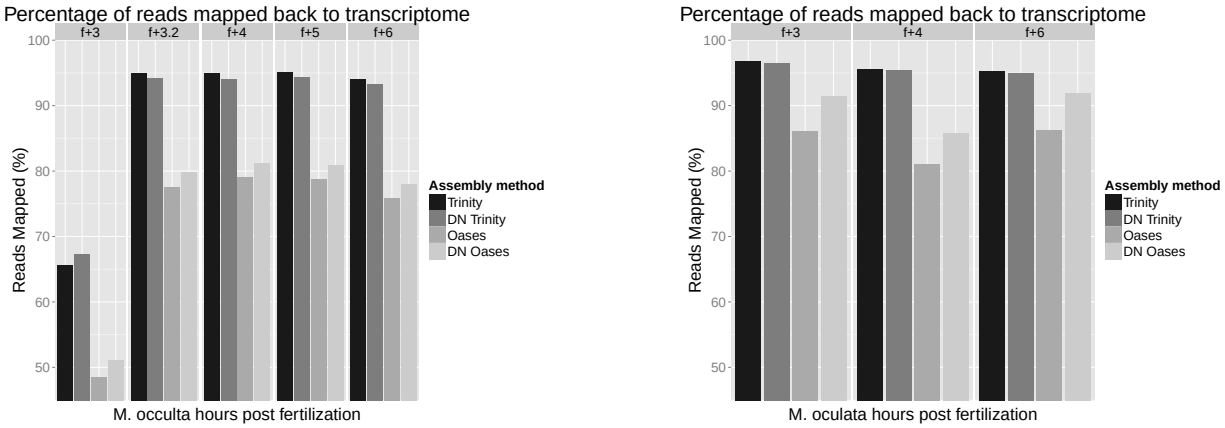
Table 3.3: **Multiplicity.** The k-mer multiplicity shows uniqueness of each assembly. All k-mers with a multiplicity of one are unique. Trinity has a higher percentage of unique k-mers when comparing assemblers. The unnormalized Trinity had the highest number of unique k-mers overall.

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.

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 [56, 28]. We use Molgulid nucleotide sequences from NCBI

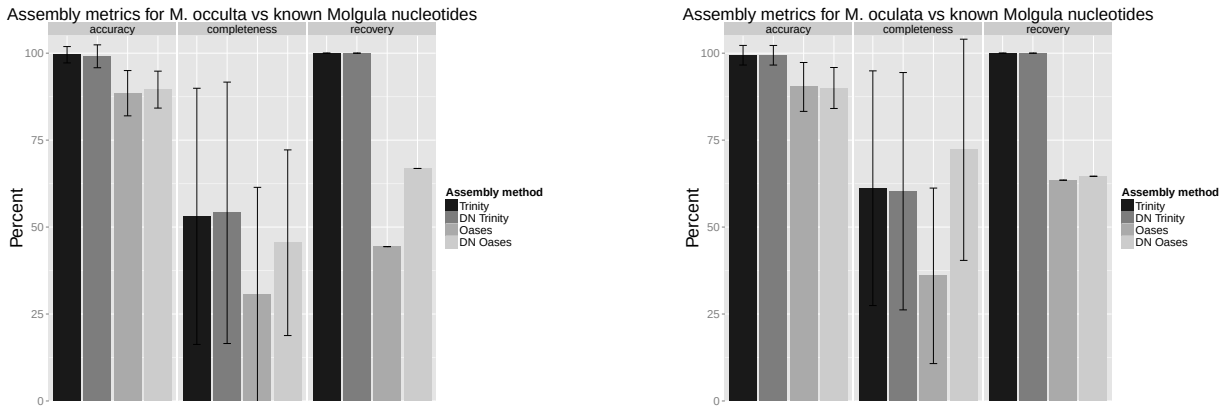


(a) *M. occulta* (b) *M. oculata*

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%.

to measure accuracy, and we define accuracy as the average BLAST identity score for the best match for each gene recovered [23]. 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.

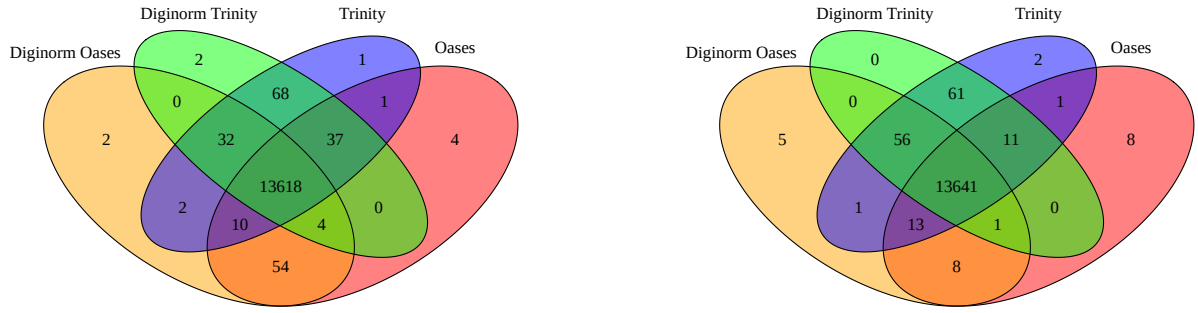
3.3.6 Both unnormalized and normalized assemblies recovered many of the same transcripts

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 [41]. 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 [37]. 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 [12, 44]—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 [37].

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 [9, 2]. 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 [36]. 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 [56, 8, 54]. 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 now for there invariant cell lineage. Only very few solitary ascidians have deviated from there typical developmental something. Many of the genes have been studied across a number of ascidian, showing that gene function tends to be orthologous within the phyla. However there are cases where gene expression differences led to variation, in addition to genes with important functions . Genomics allows us to identify elements involved in the regulation of genes. To identify elements of such, information for a number of species

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-

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 [2, 43], Howe et al., 2014). Genome assembly was done using velvet v1.2.08 (Zerbino and Birney, 2008) with k-mer overlap length ('k') ranging from 19 to 69 and scaffolding was done by Velvet, by default. Velvet does not pro-

duce 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 (Parra et al., 2007). 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/ (Tassy et al., 2010). 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

Third-nine hox genes were identified and downloaded from the NCBI database. These sequences were then BLAST against each of the three assembled Molgula genomes. The alignments were then extracted BLAST against the NCBI non-redundant database. Alignments were annotated and placed in the following files, $\mathrm{mocc}_h ox_a a.fa, mocu_h ox_a a.fa, and moxi_h ox_a a.fa, which ar$
$13 were located on the same$

4.3 Results

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 1). In addition to N50 lengths, we also used CEGMA (Core Eukaryotic Genes Mapping Approach) scores, in order to evaluate the assemblies' representative completeness (Parra et al., 2007). CEMGA 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 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 (Bradnam et al., 2013). 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, Laird, 1971; Simmen et al., 1998; Satou et al., 2008). Our sequencing efforts revealed extreme genetic divergence not only between Ciona and Molgula, as expected, but even within the Molgulids. For example, 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., 2014). Ciinte.Mesp is the sole ortholog of vertebrate genes coding for MesP and Mesogenin bHLH transcription factor family members (Satou et al., 2004). VISTA alignment shows high sequence similarity between sequences 5? upstream of the Mesp genes from the closely related M. oculata and M. occulta (Figure

1B). 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 (Figure 1?figure supplement 1). 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 (Hadfield et al., 1995; Huber et al., 2000; Tsagkogeorga et al., 2009). Our sequencing results support the view that M. occidentalis is highly diverged from other Molgula spp.

4.3.1 Gene complexes

When studying development it is important to characterize the genome for particular gene cluster/families. There are 4 HOX clusters, in humans totaling in 39 genes. Ciona has be found to have 9 box genes, Hox1-6, Hox10, and Hox12-13. Ciona is known to have to two clusters of hox genes across two chromosomes. Od also has 9 hox genes, hox1-2, hox4, a duplicate hox9, and hox10-13. Eight hox genes have be found in M. occulta and M. oculata, and nine have been found in M. occidentalis. Hox1-5, hox10 and hox12-13, with hox3-4 being found on the same contig in for both species. Additionally hox10, and hox12-13 are found on the same contig in M. oculata. However, it appears that the hox genes have been rearranged and hox10 is downstream of hox12-13. 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. The same set of hox genes were found in M. occidentalis, hox1-5, hox10 and hox12-13, however, there appears to be a duplicate hox10 gene ~12kb apart found on the same contig. M. occidentalis hox genes span across 5 contigs, hox2 has a stop codon located in the 3-4 helix.

Chapter 5

Another chapter

Chapter 6

Conclusions

APPENDIX

BIBLIOGRAPHY

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

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

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

[27] 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.

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

[29] 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.

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

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

[37] 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.

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

[39] 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.

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

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

[42] 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.

[43] 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.

[44] 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.

[45] 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.

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

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

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

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

[59] H. Yasuo and N. Satoh. Function of vertebrate t gene. *Nature*, 364(6438):582–583, Aug. 1993.

[60] 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.

[61] 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.

[62] 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.

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