Assembly on Cloud 9: Assembly metric assessment using Transcriptomes of closely related Ascidian species

Elijah K. Lowe[1,2], **Billie J. Swalla**[3], **and C. Titus Brown**[4,2,1]

[1]**Michigan State University, Computer Science and Engineering,East Lansing, Michigan, 48823, USA**
[2]**Michigan State University, Quantitative Biology, East Lansing, Michigan, 48823, USA**
[3]**University of Washington, Biology Department and Friday Harbor Laboratories, Seattle, Washington, 98195, USA**
[4]**Michigan State University, Molecular Genetics, East Lansing, Michigan, 48823, USA**

ABSTRACT

De novo transcriptome sequencing and assembly for non-model organisms has become prevalent in the past decade. However, most assembly approaches are computationally expensive, and little in-depth evaluation has been done to compare *de novo* approaches. We sequenced several developmental stages of two free-spawning marine species—*Molgula occulta* and *Molgula oculata*—assembled their transcriptomes using four different combinations of preprocessing and assembly approaches, and evaluated the quality of the assembly. We present a straightforward and reproducible mRNAseq assembly protocol that combines quality filtering, digital normalization, and assembly, together with several metrics to evaluate our *de novo* assemblies. The use of digital normalization in the protocol reduces the time and memory needed to complete the assembly and makes this pipeline available to labs without large computing infrastructure. Despite varying widely in basic assembly statistics, all of the assembled transcriptomes evaluate well in metrics such as gene recovery and estimated completeness.

Keywords: Assembly, Cloud computing, Ascidians, Next-gen sequencing

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 (Wang et al., 2009), allowing us to study organisms that are ecologically or evolutionarily interesting, amongst other biological applications.

There are now several sequencing technologies, Illumina being one of the most versatile (Glenn, 2011), that can produce millions of short reads ranging from 75 to 150 bp in length at a low cost (Zhang et al., 2011). 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 (Pop, 2009; Vinson et al., 2005; Stapley et al., 2010). 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 (Vijay et al., 2012). 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 (Grabherr et al., 2011) and Velvet/Oases (Zerbino and Birney, 2008; Schulz et al., 2012) use De Bruijn-graph based *de novo* approach which build graphs connecting the reads based on k-mer connectivity, 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 (Pop, 2009).

Here we have sequenced the transcriptomes of several developmental stages of *Molgula occult* 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 (Haas et al., 2013), and more broadly enabling transcriptome assembly by researchers without access to large computer resources.

2 METHODS

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 (Swalla and Jeffery, 1990) and culturing (Swalla et al., 1999) 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.

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

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 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 (Lohse et al., 2012; Macmanes, 2014). 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 (Brown et al., 2012). 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 1).

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 (Vijay et al., 2012; Grabherr et al., 2011; Schulz et al., 2012). 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.

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 1. Digital 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.

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>" (Li and Godzik, 2006).

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.

2.5 Gene identification

We used standalone BLAST to find reciprocal best hits (RBH) between the eight assembled transcriptomes and the *C. intestinalis* proteome from the 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.

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 (Langmead and Salzberg, 2012). 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 (Li et al., 2009) 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 RESULTS

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 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 compare to ~26 hours for *M. oculata* and ~24 hours compared to ~39 hours for *M. occulta*. *M. occulta* assembled with 23.17 GB of RAM for assembling the normalized reads versus 34.14 GB of RAM for the unnormalized reads (Figure 1).

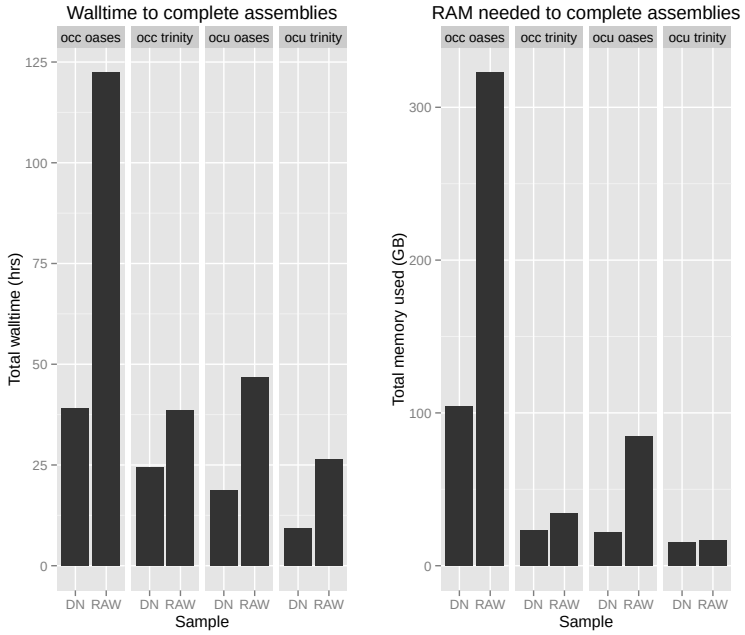


Figure 1. Wall time and memory requirements for assemblies. Wall time (left) in hours to complete the diginorm (DN) and raw read (RAW) assemblies for both species and assemblers. Oases assembled multiple k's, $21 \leq k \leq 35$ opposed to Trinity that assembles a single k. This is one reason the assembly 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.

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 less transcripts (Table 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.

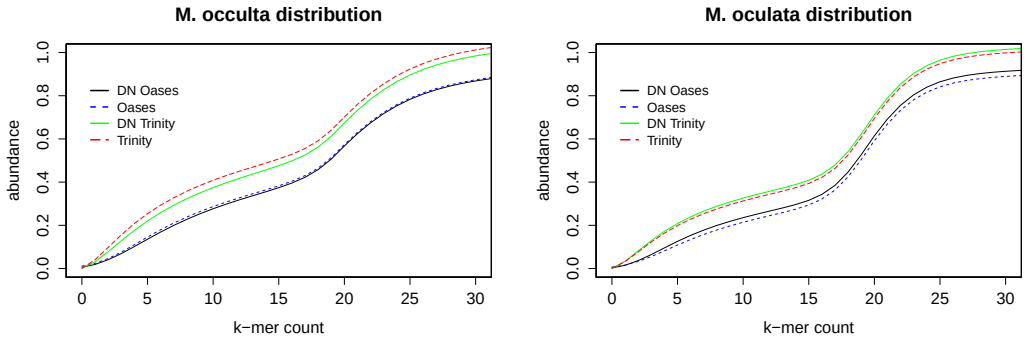
3.3 Trinity assemblies include more low-abundance k-mers than Oases assemblies

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 (Pop, 2009). We first used digital normalization to reduce the reads

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 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.



(a) (b)

Figure 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 (a) *M. occulta* k-mer distribution and (b) *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: Kmer 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. 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.

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)

3.4 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(a)). 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.

3.5 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 (Vijay et al., 2012; Martin and Wang, 2011). 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 (Li et al.,

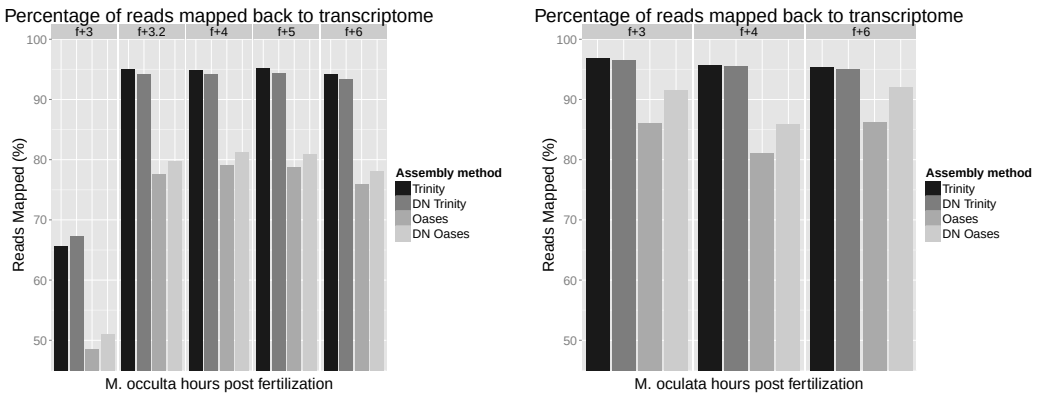


Figure 3. Read mapping. Unnormalized reads were mapped back to each of the assemblies to determine the inclusion of reads in the assembly. (a)*M. occulta* first round of gastrulation reads (f+3), showed the lowest mapping quality for all assemblies, the lowest being Raw Oases at 48.57%. *M. occulta* 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. (b)*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, having 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.

2009). 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.)

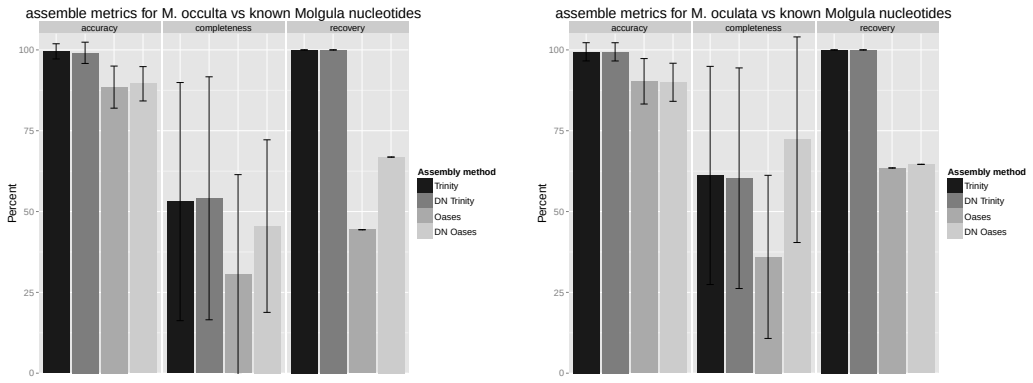
Of these 178 nucleotide sequences, 8 of them are *M. occulta* sequences and 15 of them are *M. oculata* sequences. All *M. occulta* assemblies recovered all 8 of the NCBI *M. occulta* sequences with a 94% or greater accuracy. *M. oculata* assemblies recovered *M. oculata* transcripts at a 93% accuracy 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.

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

We next 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. *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.

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

We next 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 (Satoh and Levine, 2005). 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



(a) (b)

Figure 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 preformed the best, recovering all known sequences. *M. occulta* unnormalized assembled preformed the worst, only recovering 79 (44%) of the transcripts. *M. occulta* tended to recover less of the known transcripts as well.

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. occult* and *M. oculata* assemblies with less than 1% difference in presence between the several assemblies (Figure 4b).

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.

3.8 CEGMA analysis shows high recovery of genes

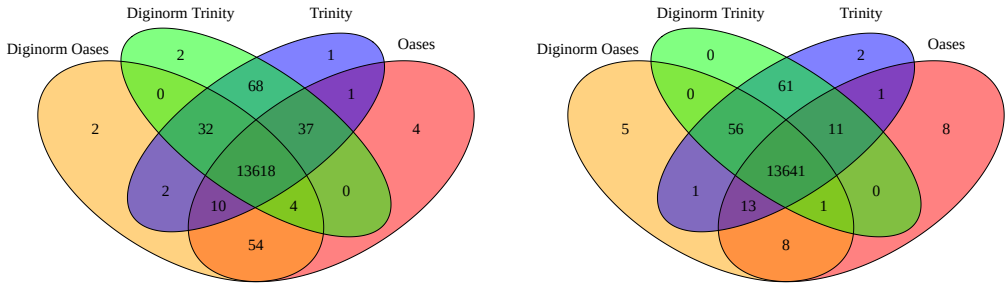
CEGMA uses a list of highly conserved eukaryotic proteins to evaluate genome and transcriptome completeness (Parra et al., 2007). We used CEGMA to analyze the number of protein families that are present in each assembly. The default CEGMA parameters were used for analysis. CEGMA reports recovery as "complete" 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.

4 DISCUSSION

4.1 Transcriptome assembly accurately recovers known transcripts and many genes

Diginorm increased the recovery rate for the known Molgula nucleotides for the Oases assemblies. Diginorm did not have much of an effect on *M. oculata*, however, *M. occulta* which has 10s of millions more reads, had a much more noticeable effect, with a 51% increase in recovered sequences. Diginorm also had a positive effect on the completeness of transcripts when assembling with Oases. Trinity performed better than Oases whether the reads are digitally normalized or unnormalized. Completeness was not over 60% for any of the assemblies, but this is explained by the fact that most of the nucleotide sequences were not mRNA so our transcripts would not align to the intronic regions. This confirms that all assembly techniques yielded good transcriptomes when applied to both data sets.

All of the transcriptome assemblies also 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 (Huber et al., 2000; Stach and Turbeville, 2002)—approximately 84% of *Ciona* proteins were found in all assemblies via BLAST,



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

Figure 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 strong overlap in homology for both assemblers and both assembly conditions.

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 (Parra et al., 2007).

4.2 Digital normalization eases assembly without strongly affecting assembly content

One of our goals in this study was explore the biological implications of digital normalization on 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 (Haas et al., 2013; Brown et al., 2012). 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 (O'Neil and Emrich, 2013). 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.

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 (Vijay et al., 2012; Grabherr et al., 2011; Van Belleghem et al., 2012). Those transcripts were indeed the results of Trinity assembling low coverage reads, having an average coverage of 5x compared to 75x.

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.

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

O'Neil, S. T. and Emrich, S. J. (2013). Assessing de novo transcriptome assembly metrics for consistency and utility. *BMC Genomics*, 14(1):465.

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

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

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

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

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

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

Swalla, B. J. and Jeffery, W. R. (1990). 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.

Swalla, B. J., Just, M. a., Pederson, E. L., and Jeffery, W. R. (1999). 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.

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

Vijay, N., Poelstra, J. W., Künstner, A., and Wolf, J. B. W. (2012). 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.

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

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

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

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