Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes

Lisa K. Johnson[1,2], Harriet Alexander[1], C. Titus Brown[1,2,3]*

[1] Department of Population Health and Reproduction, School of Veterinary Medicine, University of California, Davis
[2] Molecular, Cellular, and Integrative Physiology Graduate Group, University of California, Davis
[3] Genome Center, University of California, Davis
* Correspondence: ctbrown@ucdavis.edu

Abstract

Background

De novo transcriptome assemblies are required prior to analyzing RNAseq data from a species without an existing reference genome or transcriptome. Despite the prevalence of transcriptomic studies, the effects of using a particular workflow, or "pipeline", on the resulting assemblies are poorly understood. Here, a pipeline was automated and used to assemble and annotate raw trascriptomic data collected by the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP). Assemblies generated through this pipeline were evaluated and compared against assemblies that were previously generated with a different pipeline.

Findings

New assemblies contained 70% of the previous contigs as well as new content. An average of 7.8% ± 0.19 of the annotated contigs in the new assemblies contained novel gene names not found in the previous assemblies. Taxonomic trends were observed in the assembly metrics, with transcriptomes assembled from the phyla Dinoflagellata and Ciliophora showing a higher percentage of open reading frames and number of contigs than transcriptomes from other phyla.

Conclusions

The best possible reference transcriptomes is a moving target. Automated pipelines are useful for processing and performing computationally-intensive tasks on large sets of samples. In addition, analyzing diverse sets of data using a common workflow pipeline provides the opportunity to more truthfully identify taxon-specific trends. Streamlining workflows to re-assemble existing data in centralized and de-centralized repositories with new tools may yield novel and useful results for the community using reference transcriptomes in downstream analyses.

Introduction

The analysis of gene expression from high-throughput nucleic acid sequence data hinges on the presence of a high quality reference genome or transcriptome. When there is no reference genome or transcriptome for an organism of interest, raw RNA sequence data (RNAseq) must be assembled *de novo* into a transcriptome [1]. This type of analysis is ubiquitous across many fields, including evolutionary developmental biology [2], cancer biology [3], agriculture [4,5], ecological physiology [6,7], and biological oceanography [8]. In recent years, substantial investments have been made in data generation, primary data analysis, and development of downstream applications, such as biomarkers and diagnostic tools [9–16].

Methods for *de novo* RNAseq assembly of the most common short read Illumina sequencing data continue to evolve rapidly, especially for non-model species [17]. At this time, there are several major *de novo* transcriptome assembly software tools available to choose from, including Trinity [18], SOAPdenovo-Trans [19], Trans-ABySS [20], Oases [21], SPAdes [22], IDBA-tran [23], and Shannon [24]. The availability of these options stems from continued research into the unique computational challenges associated with transcriptome assembly of short read Illumina RNAseq data, including large memory requirements, alternative splicing and allelic variants [18,25].

With the continuous development of new tools, workflows, and increasing data generation capacity, there is often the opportunity to re-analyze old data with new tools. However, this is rarely done systematically. To evaluate the performance impact of new tools on old data, we developed and applied an automated, modularized and extensible *de novo* transcriptome assembly workflow based on the Eel Pond Protocol, which incorporates Trimmomatic, digital normalization with khmer software, and Trinity.

To evaluate this pipeline, we reanalyzed RNAseq data from 678 samples generated as part of the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP). The MMETSP RNAseq data set was originally generated to broaden the diversity of sequenced marine protists to facilitate our understanding of their evolution and their roles in marine ecosystems and biogeochemical cycles [26]. With data from species spanning more than 40 eukaryotic phyla, the MMETSP provides one of the largest publicly available RNAseq data sets. Moreover, the MMETSP used a standardized library preparation procedure and all of the samples were sequenced at the same facility, making the data sets unusually comparable.

Reference transcriptomes for the MMETSP were originally assembled by the National Center for Genome Research (NCGR) using a pipeline which used the Trans-ABySS software program [26]. These transcriptomes have already facilitated discoveries in the evolutionary history of ecologically significant genes [27,28], differential gene expression under shifting environmental

conditions [8,29], inter-group transcriptome comparisons [30], unique transcriptional features [31–33], and meta-transcriptomic studies of phytoplankton bloom formation [34–36].

In re-assembling the MMETSP data, we sought to compare and improve the original MMETSP reference transcriptome and to create a platform which facilitates automated re-assembly and evaluation. Below, we show that our re-assemblies had higher evaluation metrics, contained most of the NCGR contigs as well as new content. Also, assemblies varied by taxonomic group being assembled.

Methods

Automated Pipeline

An automated pipeline was developed to execute the steps of the Eel Pond mRNAseq Protocol [37], a lightweight protocol for assembling short Illumina RNA-seq reads that uses the Trinity *de novo* transcriptome assembler. This protocol generates *de novo* transcriptome assemblies of acceptable quality [38]. The pipeline was used to assemble all of the data from the MMETSP (Figure 1). The code and instructions for running the pipeline are available at https://doi.org/10.5281/zenodo.249982.

The steps of the pipeline applied to the MMETSP are as follows:

1. Download the raw data

Raw RNA-seq data sets were obtained from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) from BioProject PRJNA231566. Data were paired-end (PE) Illumina reads with lengths of 50 bases for each read. The metadata file obtained from the SRA web interface was used to provide a list of samples to the *get_data.py* pipeline script to download and extract fastq files from 719 records. The script uses the fastq-dump program from the SRA Toolkit to extract the SRA-formatted fastq files (version 2.5.4) [39]. There were 18 MMETSP samples with more than one SRA record (MMETSP0693, MMETSP1019, MMETSP0923, MMETSP0008, MMETSP1002, MMETSP1325, MMETSP1018, MMETSP1346, MMETSP0088, MMETSP0092, MMETSP0717, MMETSP0223, MMETSP0115, MMETSP0196, MMETSP0197, MMETSP0398, MMETSP0399, MMETSP0922). In these cases, reads from multiple SRA records were concatenated together per sample. Taking these redundancies into consideration, there were a total of 678 re-assemblies generated from the 719 records in PRJNA231566. Assembly evaluation metrics were not calculated for MMETSP samples with more than one SRA.

Initial transcriptomes, both 'cds' and 'nt' versions, assembled by the National Center for Genome Resources (NCGR) using methods and data described in the original publication [26] were downloaded from the iMicrobe repository to compare with our re-assemblies (ftp://ftp.imicrobe.us/projects/104/). The version used for comparison is noted below in each evaluation step. To our knowledge, the NCGR took extra post-processing steps to filter content leaving only coding sequences in the 'cds' versions of each assembly [26].

2. Perform quality control

Reads were analyzed with FastQC (version 0.11.5) and multiqc (version 1.2) [40] to confirm overall qualities before and after trimming. A conservative trimming approach was used with Trimmomatic (version 0.33) [41] to remove residual Illumina adapters and cut bases off the start (LEADING) and end (TRAILING) of reads if they were below a threshold Phred quality score (Q<2).

3. Apply digital normalization

To decrease the memory requirements for each assembly, reads were interleaved, normalized to a *k*-mer coverage of 20 and a memory size of 4e9, then low-abundance *k*-mers were trimmed. Orphaned reads, where the mated pair was removed during normalization, were included in the assembly.

4. Assemble

Transcriptomes were assembled from normalized reads with Trinity 2.2.0 using default parameters.

The resulting assemblies are referred to below as the "Lab for Data Intensive Biology" assemblies, or DIB. The original assemblies are referred to as the NCGR assemblies.

5. Post-assembly assessment

Transcriptomes were annotated using the dammit pipeline (Scott 2016), which relies on the following databases as evidence: Pfam-A [42], Rfam [43], OrthoDB [44]. In the case where there were multiple database hits, one gene name was selected for each contig by selecting the name of the lowest e-value match (<1e-05).

All assemblies were evaluated using metrics generated by the Transrate program [45]. Trimmed reads were used to calculate a Transrate score for each assembly, which represents the geometric mean of all contig scores multiplied by the proportion of input reads providing positive support

for the assembly [45]. Comparative metrics were calculated using Transrate for each MMETSP sample between DIB and the NCGR assemblies using the Conditional Reciprocal Best BLAST hits (CRBB) algorithm [46]. A forward comparison was made with the NCGR assembly used as the reference and each DIB re-assembly as the query. Reverse comparative metrics were calculated with each DIB re-assembly as the reference and the NCGR assembly as the query. Transrate scores were calculated for each assembly using the Trimmomatic quality-trimmed reads, prior to digital normalization.

Benchmarking Universal Single-Copy Orthologs (BUSCO) software (version 3) was used with a database of 234 orthologous genes specific to protistans and 306 genes specific to eukaryota with open reading frames in the assemblies. BUSCO scores are frequently used as one measure of assembly completeness [47].

To assess the occurrences of fixed-length words in the assemblies, unique 25-mers were measured in each assembly using the HyperLogLog estimator of cardinality built into the khmer software package [48]. Unique gene names were compared from a random subset of 296 samples using the dammit annotation pipeline [49]. If a gene name was annotated in NCGR but not in DIB, this was considered a gene uniquely annotated in NCGR. Unique gene names were normalized to the total number of annotated genes in each assembly.

A Tukey's honest significant different (HSD) range test of multiple pairwise comparisons was used to measure differences between distributions of data from the top seven most-represented phyla using the 'agricolae' package version 1.2-8 in R version 3.4.2 (2017-09-28). Margins sharing a letter in the group label are not significantly different at the 5% level (Figure 8). Averages are reported ± standard deviation.

Results

After assemblies and annotations were completed, files were uploaded to Figshare and are available for download [50]. URL?

Differences in available evaluation metrics between NCGR and DIB were variable.

The majority of transcriptome evaluation metrics collected for each sample were higher in Trinity-based DIB re-assemblies than for the Trans-ABySS-based NCGR assemblies (Table 1), with the exception being the Transrate score from the "nt" version of the assembly. The Transrate score with this 'cds' version was higher in DIB compared to NCGR but lower in DIB compared to the NCGR 'nt' version (Supplemental Figure 1).

DIB re-assemblies had more contigs than the NCGR assemblies in 83.5% of the samples (Table 1). The mean number of contigs in the DIB re-assemblies was 48,361 ± 35,703 while the mean

number of contigs in the NCGR 'nt' assemblies was 30,532 ± 21,353 (Figure 2). A two-sample Kolmogorov-Smirnov test comparing distributions indicated that the number of contigs were significantly different between DIB and NCGR assemblies ($p < 0.001$, $D = 0.35715$). Transrate scores [35], which calculate the overall quality of the assembly based on the original reads, were significantly higher in the DIB re-assemblies (0.31 ± 0.1) compared to the 'cds' versions of the NCGR assemblies (0.22 ± 0.09) ($p < 0.001$, $D = 0.49899$). Whereas the Transrate scores in the NCGR 'nt' assemblies (0.35 ± 0.09) were significantly higher than the DIB assemblies (0.22 ± 0.09) ($p < 0.001$, $D = 0.22475$) (Supplemental Figure 1). The frequency of the differences between Transrate scores in the NCGR 'nt' assemblies and the DIB re-assemblies appears to be normally distributed (Figure 2C). Transrate scores from the DIB assemblies relative to the NCGR 'nt' assemblies did not appear to have taxonomic trends (Supplemental Figure 2).

The DIB re-assemblies contained most of the NCGR contigs as well as new content.

A conditional reciprocal best BLAST (CRBB) hit is indicative of sequence containment between assemblies. A positive CRBB result indicates that one assembly contains the same contig information as the other. Thus, the proportion of positive CRBB hits can be used as a scoring metric to compare the relative similarity of content between two assemblies. For example, MMETSP0949 (*Chattonella subsalsa*) had 39,051 contigs and a CRBB score of 0.70968 in the DIB re-assembly whereas in the NCGR assembly of the same sample had 18,873 contigs and a CRBB score of 0.33933. This indicated that 70.968% of the reference of DIB was covered by the NCGR assembly, whereas in the reverse alignment, the NCGR reference assembly was only covered by 33.933% of the DIB re-assembly. The mean CRBB score in DIB when queried against NCGR 'nt' as a reference was 0.70 ± 0.22, while the mean proportion for NCGR 'nt' assemblies queried against DIB re-assemblies was 0.49 ± 0.10 ($p < 0.001$, $D = 0.71121$) (Figure 3). This indicates that more contigs from the NCGR assemblies were included in the DIB re-assemblies than vice versa and also suggests that the DIB re-assemblies overall have additional content. This finding is reinforced by higher unique *k*-mer content found in the DIB re-assemblies compared to NCGR, where 85% of the samples had more unique *k*-mers in the DIB re-assemblies compared to NCGR assemblies (Figure 4).

To investigate whether the new sequence content was genuine, we examined two different metrics that take into account the biological quality of the assemblies. First, the estimated content of open reading frames (ORFs), or coding regions, across contigs was quantified. Though DIB re-assemblies had more contigs, the ORF content is similar to the original assemblies, with a mean of 81.8% ± 9.94 ORF content in DIB re-assemblies and 76.7% ± 10.1 ORF content in the NCGR assemblies. Nonetheless, ORF content in DIB re-assemblies was slightly higher than NCGR assemblies for 95% of the samples (Figure 5A), although DIB re-assemblies had significantly significantly higher ORF content ($p < 0.001$, $D = 2681$). Secondly, when the assemblies were queried against the eukaryotic BUSCO database [37], the percentages of

BUSCO matches in the DIB re-assemblies (63.03% ± 18.6) were less significantly different compared to the original NCGR assemblies (64.90% ± 19.1) ($p = 0.001873$, $D = 0.10291$) (Figure 5B). Thus, although the number of contigs and amount of content was increased in the DIB re-assemblies compared to the NCGR assemblies, the ORF content and contigs matching with the BUSCO database did not decrease, suggesting that the extra content contained similar proportions of ORFs and BUSCO annotations and, therefore, might be biologically meaningful.

Following annotation by the dammit pipeline (Scott 2016), 91% ± 1.58 of the contigs in the DIB re-assemblies had positive matches with sequence content in the databases queried (Pfam, Rfam, and OrthoDB), with 48% ± 0.87 of those containing unique gene names (the remaining are fragments of the same gene). Of those annotations, 7.8% ± 0.19 were identified as novel compared to the NCGR 'nt' assemblies, determined by a "false" CRBB result (Figure 6). Additionally, the number of unique gene names in DIB re-assemblies were higher than in NCGR, suggesting an increase in genic content (Figure 7).

Novel contigs in the DIB re-assemblies likely represent a combination of unique annotations, allelic variants and alternatively spliced isoforms. For example, "F0XV46_GROCL", "Helicase_C", "ODR4-like","PsaA_PsaB", and "Metazoa_SRP" are novel gene names found annotated in the DIB re-assembly of the sample MMETSP1473 (*Stichococcus* sp.) that were absent in the NCGR assembly of this same sample. Other gene names, for example "Pkinase_Tyr","Bromodomain", and "DnaJ", are found in both the NCGR and DIB assemblies, but are identified as novel contigs based on negative CRBB results in the DIB re-assembly of sample MMETSP1473 compared to the NCGR reference.

Assembly metrics varied by taxonomic group being assembled.

To examine systematic taxonomic differences in the assemblies, several different metrics for content and assembly quality were assessed (Figure 8). Metrics were grouped by the top seven most represented phyla in the MMETSP data set as follows: Bacillariophyta (N=193), Dinophyta (N=128), Ochrophyta (N=78), Haptophyta (N=63), Chlorophyta (N=62), Ciliophora (N=31), Cryptophyta (orange, N=22). While there were no differences between the phyla in the number of input reads (Figure 8 A), the Dinoflagellates (Dinophyta) had higher ORF percentages and more contigs than other groups (Figure 8 B,C), and assemblies from Ciliates (Ciliophora) had lower unique *k*-mers (Figure 8 D).

Discussion

DIB re-assemblies contained the majority of the previously-assembled contigs.

We used a different pipeline than the original one used to create the NCGR assemblies, in part because new software was available [8] and in part because of new trimming guidelines [27]. We

But we would improve assemblies.

had no *a priori* expectation for the similarity of the results, yet we found that in the majority of cases the new DIB re-assemblies included substantial portions of the previous NCGR assemblies. Moreover, both the fraction of contigs with ORFs and the mean percentage of BUSCO matches were similar between the two assemblies, suggesting that both pipelines yielded equally valid contigs, even though the NCGR assemblies ~~were less sensitive~~.

found fewer contigs.

Reassembly with new tools can yield new results

Evaluation with several quality metrics suggested that the DIB re-assemblies were more inclusive than the NCGR assemblies. In addition to containing more contigs and being more inclusive of the NCGR assemblies than vice versa. The Transrate scores in the DIB re-assemblies compared to the NCGR 'nt' assemblies were significantly lower, indicating that the NCGR 'nt' assemblies had better overall read inclusion in the assembled contigs, whereas the DIB assemblies had higher Transrate scores than the NCGR 'cds' version. This indicates that the NCGR 'cds' version, which were post-processed to only include coding sequence content, were missing information originally in the quality-trimmed reads. To our knowledge, the Transrate score [45] is the only single metric available for evaluating the quality of a *de novo* transcriptome. The based on RSEM-eval DETONATE score is similar to the Transrate score in that it returns assembly likelihood given the read data [51]. Metrics evaluating the de Bruijn graph data structure used to produce the assembled contigs may be better evaluators of the assembly quality in the future. We see here in this study that the DIB re-assemblies which used the Trinity *de novo* assembly software typically contained more *k*-mers, more annotated transcripts, and more unique gene names than the NCGR assemblies. These points all suggest that additional content in these re-assemblies might be biologically meaningful and that these re-assemblies provide new content not available in the previous NCGR assemblies. Since contigs are probabilistic predictions made by assembly software for full-length transcripts [52], "final" reference assemblies are approximations of the full set of transcripts in the transcriptome. This study suggests that a reference transcriptome may be a moving target and that these predictions may be improved with updated tools.

are

demonstrates

The evaluation metrics described here serve as a framework for better contextualizing the quality of protistan transcriptomes. For some species and strains in the MMETSP data set, these data represent the first nucleic acid sequence information available [26].

Automated pipelines can be used to process arbitrarily many RNAseq samples

The automated and modularized nature of this pipeline was useful for processing large data sets like the MMETSP as it allowed for batch processing of the entire collection, including re-analysis when new tools become available (see op-ed Alexander et al. 2018). During the course of this project, we ran four entire re-assemblies of the entire MMETSP data set as versions of the component tools were updated. Each re-analysis required only a single command and

approximately half a CPU-year of compute. The value of modularized automation is clear when new data sets become available, tools are updated, or many tools are compared in benchmark studies. Despite this, few assembly efforts completely automate their process, perhaps because the up-front cost of doing so is high compared to the size of the dataset typically being analyzed.

Analyzing many samples using a common pipeline identifies taxon-specific trends

The MMETSP dataset presents an opportunity to examine transcriptome qualities for hundreds of taxonomically diverse species spanning a wide array of protistan lineages. This is among the largest set of diverse RNAseq data to be examined. In comparison, the Assemblathon2 project compared genome assembly pipelines using data from three vertebrate species [53]. The BUSCO paper assessed 70 genomes and 96 transcriptomes representing groups of diverse species (vertebrates, arthropods, other metazoans, fungi) [47]. Other benchmarking studies have examined transcriptome qualities for samples representing dozens of species from different taxonomic groupings [54,51].

Assembly evaluation tools yielded results outside the range of what is normal for some organisms. For example, the case of low ORF predictions in Ciliophora. It has recently been found that ciliates have an alternative triplet codon dictionary, with codons normally encoding STOP serving a different purpose [31–33]. In addition, Dinophyta demonstrated a significantly higher number of unique *k*-mers and total contigs in assemblies. Such a finding supports previous evidence from studies that large gene families are constitutively expressed in Dinophyta [55]. In future development of *de novo* transcriptome assembly software, the incorporation of phylum-specific information may be useful in improving the overall quality of assemblies for different taxa. Phylogenetic trends are important to consider in the assessment of transcriptome quality, given that the assemblies from Dinophyta and Ciliophora are distinguished from other assemblies by some metrics. Applying domain-specific knowledge, such as specialized transcriptional features in Ciliophora and Dinophyta, in combination with other evaluation metrics can help to evaluate whether a transcriptome is of good quality or "finished" enough to serve as a high quality reference to answer the biological questions of interest.

Conclusion

As the rate of sequencing data generation continues to increase, efforts to automate the processing and evaluation of sequence data are becoming increasingly important. Ultimately, the goal in generating *de novo* transcriptomes is to create the best possible reference against which downstream analyses can be accurately based. This study demonstrated that re-analysis of old data with new tools and methods improved the quality of the reference assembly through an expansion of the gene catalogue of the dataset. Notably, these improvements arose without further experimentation or sequencing.

With the growing volume of nucleic acid data in centralized and de-centralized repositories, streamlining methods into pipelines ~~such as this~~ will not only enhance the reproducibility of the analysis, but will help to facilitate inter-group comparisons amongst datasets from diverse taxa from large collections of samples. Automation tools were key in successfully processing and analyzing this large collection of 678 samples.

Acknowledgements

Camille Scott, Luiz Irber and other members of the Data Intensive Biology lab at UC Davis provided helpful assistance with troubleshooting the assembly, annotation and evaluation pipeline. Funding was provided from the Gordon and Betty Moore Foundation under award number GBMF4551 to CTB. Scripts were tested and run on the MSU HPCC and NSF-XSEDE Jetstream with allocation TG-BIO160028.

References

1. Geniza M, Jaiswal P. Tools for building de novo transcriptome assembly. 2018 [cited 2018 Mar 14]; Available from: https://ac.els-cdn.com/S2214662817301032/1-s2.0-S2214662817301032-main.pdf?_tid=dd330b95-f4a0-4c7b-81b9-9c0f9ec61db2&acdnat=1521094046_f477c447e64870799243b677bd77fcb3
2. Tulin S, Aguiar D, Istrail S, Smith J. A quantitative reference transcriptome for Nematostella vectensis early embryonic development: a pipeline for de novo assembly in emerging model systems. 2013 [cited 2018 Mar 15]; Available from: http://www.evodevojournal.com/content/4/1/16
3. Mittal VK, McDonald JF. De novo assembly and characterization of breast cancer transcriptomes identifies large numbers of novel fusion-gene transcripts of potential functional significance. BMC Med. Genomics [Internet]. BioMed Central; 2017 [cited 2018 Mar 14];10:53. Available from: http://www.ncbi.nlm.nih.gov/pubmed/28851357
4. Yang S, Liu H-D, Qiang Z, Zhang H-J, Zhi-Dong Z, Li Y-D, et al. ScienceDirect High-throughput sequencing of highbush blueberry transcriptome and analysis of basic helix-loop-helix transcription factors. J. Integr. Agric. [Internet]. 2017 [cited 2018 Mar 15];16:591–604. Available from: https://ac.els-cdn.com/S2095311916614612/1-s2.0-S2095311916614612-main.pdf?_tid=09e6059a-44db-4e2a-bc90-35745fe2294f&acdnat=1521157724_1465eaf1289f7bc1c251646a81b4dc10
5. Suárez-Vega A, Gutiérrez-Gil B, Klopp C, Tosser-Klopp G, Arranz J-J. Comprehensive RNA-Seq profiling to evaluate lactating sheep mammary gland transcriptome. Sci. Data [Internet]. Nature Publishing Group; 2016 [cited 2018 Mar 15];3:160051. Available from: http://www.nature.com/articles/sdata201651
6. Carruthers M, Yurchenko AA, Augley JJ, Adams CE, Herzyk P, Elmer KR. De novo transcriptome assembly, annotation and comparison of four ecological and evolutionary model salmonid fish species. BMC Genomics [Internet]. 2018 [cited 2018 Mar 15];19. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5759245/pdf/12864_2017_Article_4379.pdf
7. Mansour TA, Rosenthal JJC, Brown CT, Roberson LM. Transcriptome of the Caribbean stony coral Porites astreoides from three developmental stages. [cited 2018 Mar 15]; Available from:

https://gigascience.biomedcentral.com/track/pdf/10.1186/s13742-016-0138-1?site=gigascience.biomedcentral.com
8. Frischkorn KR, Harke MJ, Gobler CJ, Dyhrman ST. De novo assembly of Aureococcus anophagefferens transcriptomes reveals diverse responses to the low nutrient and low light conditions present during blooms. Front. Microbiol. [Internet]. Frontiers; 2014 [cited 2017 Sep 20];5:375. Available from: http://journal.frontiersin.org/article/10.3389/fmicb.2014.00375/abstract
9. Mansour TA, Scott EY, Finno CJ, Bellone RR, Mienaltowski MJ, Penedo MC, et al. Tissue resolved, gene structure refined equine transcriptome. BMC Genomics [Internet]. BMC Genomics; 2017;18:103. Available from: http://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-016-3451-2
10. Gonzalez VL, Andrade SCS, Bieler R, Collins TM, Dunn CW, Mikkelsen PM, et al. A phylogenetic backbone for Bivalvia: an RNA-seq approach. Proc. R. Soc. B Biol. Sci. [Internet]. The Royal Society; 2015 [cited 2018 Mar 15];282:20142332–20142332. Available from: http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2014.2332
11. Müller M, Seifert S, Lübbe T, Leuschner C, Finkeldey R. De novo transcriptome assembly and analysis of differential gene expression in response to drought in European beech. Chen Z-H, editor. PLoS One [Internet]. Public Library of Science; 2017 [cited 2017 Sep 22];12:e0184167. Available from: http://dx.plos.org/10.1371/journal.pone.0184167
12. Heikkinen LK, Kesäniemi JE, Knott KE. De novo transcriptome assembly and developmental mode specific gene expression of *Pygospio elegans*. Evol. Dev. [Internet]. 2017 [cited 2017 Sep 22];19:205–17. Available from: http://doi.wiley.com/10.1111/ede.12230
13. Li F, Wang L, Lan Q, Yang H, Li Y, Liu X, et al. RNA-Seq Analysis and Gene Discovery of Andrias davidianus Using Illumina Short Read Sequencing. Davies WIL, editor. PLoS One [Internet]. Public Library of Science; 2015 [cited 2018 Mar 15];10:e0123730. Available from: http://dx.plos.org/10.1371/journal.pone.0123730
14. Yu J, Lou Y, Zhao A. Transcriptome analysis of follicles reveals the importance of autophagy and hormones in regulating broodiness of Zhedong white goose. Sci. Rep. [Internet]. Nature Publishing Group; 2016 [cited 2018 Mar 15];6:36877. Available from: http://www.nature.com/articles/srep36877
15. Seo M, Kim K, Yoon J, Jeong JY, Lee H-J, Cho S, et al. RNA-seq analysis for detecting quantitative trait-associated genes. Sci. Rep. [Internet]. Nature Publishing Group; 2016 [cited 2018 Mar 15];6:24375. Available from: http://www.nature.com/articles/srep24375
16. Pedrotty DM, Morley MP, Cappola TP. Transcriptomic biomarkers of cardiovascular disease. Prog. Cardiovasc. Dis. [Internet]. NIH Public Access; 2012 [cited 2018 Mar 15];55:64–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22824111
17. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol [Internet]. 2016;17:13. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26813401
18. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. [Internet]. Nature Research; 2011 [cited 2017 Sep 21];29:644–52. Available from: http://www.nature.com/doifinder/10.1038/nbt.1883
19. Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, et al. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics [Internet]. Oxford University Press; 2014 [cited 2017 Sep 20];30:1660–6. Available from:

https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu077
20. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, et al. De novo assembly and analysis of RNA-seq data. Nat. Methods [Internet]. 2010;7:909–12. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20935650
21. Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics [Internet]. Oxford University Press; 2012 [cited 2017 Sep 20];28:1086–92. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts094
22. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J. Comput. Biol. [Internet]. Mary Ann Liebert, Inc. 140 Huguenot Street, 3rd Floor New Rochelle, NY 10801 USA ; 2012 [cited 2017 Sep 20];19:455–77. Available from: http://online.liebertpub.com/doi/abs/10.1089/cmb.2012.0021
23. Peng Y, Leung HCM, Yiu S-M, Lv M-J, Zhu X-G, Chin FYL. IDBA-tran: a more robust de novo de Bruijn graph assembler for transcriptomes with uneven expression levels. Bioinformatics [Internet]. Oxford University Press; 2013 [cited 2017 Sep 20];29:i326–34. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btt219
24. Kannan S, Hui J, Mazooji K. Shannon : An Information-Optimal de Novo RNA-Seq Assembler. 2016;1–14.
25. Chang Z, Wang Z, Li G. The Impacts of Read Length and Transcriptome Complexity for De Novo Assembly: A Simulation Study. Papavasiliou FN, editor. PLoS One [Internet]. Public Library of Science; 2014 [cited 2017 Sep 20];9:e94825. Available from: http://dx.plos.org/10.1371/journal.pone.0094825
26. Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. Roberts RG, editor. PLoS Biol. [Internet]. Public Library of Science; 2014 [cited 2017 Feb 16];12:e1001889. Available from: http://dx.plos.org/10.1371/journal.pbio.1001889
27. Durkin CA, Koester JA, Bender SJ, Armbrust EV. The evolution of silicon transporters in diatoms. Kroth P, editor. J. Phycol. [Internet]. 2016 [cited 2017 Sep 20];52:716–31. Available from: http://doi.wiley.com/10.1111/jpy.12441
28. Groussman RD, Parker MS, Armbrust EV. Diversity and Evolutionary History of Iron Metabolism Genes in Diatoms. Missirlis F, editor. PLoS One [Internet]. Public Library of Science; 2015 [cited 2017 Sep 20];10:e0129081. Available from: http://dx.plos.org/10.1371/journal.pone.0129081
29. Harke MJ, Juhl AR, Haley ST, Alexander H, Dyhrman ST. Conserved Transcriptional Responses to Nutrient Stress in Bloom-Forming Algae. Front. Microbiol. [Internet]. Frontiers; 2017 [cited 2018 Mar 15];8:1279. Available from: http://journal.frontiersin.org/article/10.3389/fmicb.2017.01279/full
30. Koid AE, Liu Z, Terrado R, Jones AC, Caron DA, Heidelberg KB. Comparative Transcriptome Analysis of Four Prymnesiophyte Algae. Xiao J, editor. PLoS One [Internet]. Public Library of Science; 2014 [cited 2017 Sep 20];9:e97801. Available from: http://dx.plos.org/10.1371/journal.pone.0097801
31. Alkalaeva E, Mikhailova T. Reassigning stop codons via translation termination: How a few eukaryotes broke the dogma. BioEssays [Internet]. 2017 [cited 2017 Sep 20];39:1600213.

Available from: http://doi.wiley.com/10.1002/bies.201600213
32. Heaphy SM, Mariotti M, Gladyshev VN, Atkins JF, Baranov P V. Novel Ciliate Genetic Code Variants Including the Reassignment of All Three Stop Codons to Sense Codons in *Condylostoma magnum*. Mol. Biol. Evol. [Internet]. Oxford University Press; 2016 [cited 2017 Sep 20];33:2885–9. Available from: https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msw166
33. Swart EC, Serra V, Petroni G, Nowacki M. Genetic Codes with No Dedicated Stop Codon: Context-Dependent Translation Termination. Cell [Internet]. The Author(s); 2016;166:691–702. Available from: http://dx.doi.org/10.1016/j.cell.2016.06.020
34. Alexander H, Jenkins BD, Rynearson TA, Dyhrman ST. Metatranscriptome analyses indicate resource partitioning between diatoms in the field. Proc. Natl. Acad. Sci. U. S. A. [Internet]. National Academy of Sciences; 2015 [cited 2017 Sep 20];112:E2182-90. Available from: http://www.ncbi.nlm.nih.gov/pubmed/25870299
35. Alexander H, Rouco M, Haley ST, Wilson ST, Karl DM, Dyhrman ST. Functional group-specific traits drive phytoplankton dynamics in the oligotrophic ocean. Proc. Natl. Acad. Sci. U. S. A. [Internet]. National Academy of Sciences; 2015 [cited 2018 Mar 15];112:E5972-9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26460011
36. Gong W, Browne J, Hall N, Schruth D, Paerl H, Marchetti A. Molecular insights into a dinoflagellate bloom. ISME J. [Internet]. Nature Publishing Group; 2017 [cited 2018 Mar 15];11:439–52. Available from: http://www.nature.com/articles/ismej2016129
37. Brown CT, Scott C, Crusoe MR, Sheneman L, Rosenthal J, Howe A. khmer-protocols 0.8.4 documentation. 2013 [cited 2017 Oct 17]; Available from: https://www.mendeley.com/import/?url=https://figshare.com/articles/khmer_protocols_0_8_3_documentation/878460
38. Lowe EK, Swalla BJ, Brown CT. Evaluating a lightweight transcriptome assembly pipeline on two closely related ascidian species. PeerJ Prepr. [Internet]. 2014;2:e505v1. Available from: https://dx.doi.org/10.7287/peerj.preprints.505v1
39. Leinonen R, Sugawara H, Shumway M. The Sequence Read Archive. Nucleic Acids Res. [Internet]. Oxford University Press; 2011 [cited 2017 Oct 17];39:D19–21. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkq1019
40. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics [Internet]. Oxford University Press; 2016 [cited 2017 Oct 17];32:3047–8. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw354
41. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics [Internet]. Oxford University Press; 2014 [cited 2017 Oct 17];30:2114–20. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btu170
42. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. [Internet]. Oxford University Press; 2016 [cited 2017 Oct 17];44:D279–85. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1344
43. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, et al. Rfam: updates to the RNA families database. Nucleic Acids Res. [Internet]. Oxford University Press; 2009 [cited 2017 Oct 17];37:D136–40. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkn766

544 44. Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, et al.
545 OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant,
546 archaeal, bacterial and viral orthologs. Nucleic Acids Res. [Internet]. Oxford University Press;
547 2017 [cited 2017 Sep 21];45:D744–9. Available from: https://academic.oup.com/nar/article-
548 lookup/doi/10.1093/nar/gkw1119
549 45. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate: reference-free quality
550 assessment of de novo transcriptome assemblies. Genome Res. [Internet]. Cold Spring Harbor
551 Laboratory Press; 2016 [cited 2017 Oct 17];26:1134–44. Available from:
552 http://www.ncbi.nlm.nih.gov/pubmed/27252236
553 46. Aubry S, Kelly S, Kümpers BMC, Smith-Unna RD, Hibberd JM. Deep Evolutionary
554 Comparison of Gene Expression Identifies Parallel Recruitment of Trans-Factors in Two
555 Independent Origins of C4 Photosynthesis. Bomblies K, editor. PLoS Genet. [Internet]. Public
556 Library of Science; 2014 [cited 2017 Oct 17];10:e1004365. Available from:
557 http://dx.plos.org/10.1371/journal.pgen.1004365
558 47. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO:
559 assessing genome assembly and annotation completeness with single-copy orthologs.
560 Bioinformatics [Internet]. Oxford University Press; 2015 [cited 2017 Sep 21];31:3210–2.
561 Available from: https://academic.oup.com/bioinformatics/article-
562 lookup/doi/10.1093/bioinformatics/btv351
563 48. Irber Junior LC, Brown CT. Efficient cardinality estimation for k-mers in large DNA
564 sequencing data sets. doi.org [Internet]. Cold Spring Harbor Laboratory; 2016 [cited 2017 Oct
565 17];56846. Available from: https://www.biorxiv.org/content/early/2016/06/07/056846
566 49. Scott C. dammit: an open and accessible de novo transcriptome annotator. in prep. [Internet].
567 2016; Available from: www.camillescott.org/dammit
568 50. Johnson, L; Alexander, H; Brown CT. Marine Microbial Eukaryotic Transcriptome
569 Sequencing Project, re-assemblies [Internet]. 2017. Available from:
570 https://doi.org/10.6084/m9.figshare.3840153.v6
571 51. Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, et al. Evaluation of de novo
572 transcriptome assemblies from RNA-Seq data. Genome Biol. [Internet]. BioMed Central; 2014
573 [cited 2017 Oct 17];15:553. Available from:
574 http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0553-5
575 52. Li B, Fillmore N, Bai Y, Collins M, Thomson J a., Stewart R, et al. Evaluation of de novo
576 transcriptome assemblies from RNA-Seq data. bioRxiv [Internet]. 2014;1–21. Available from:
577 http://biorxiv.org/content/early/2014/06/13/006338.abstract
578 53. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, et al. Assemblathon 2:
579 evaluating de novo methods of genome assembly in three vertebrate species. Gigascience
580 [Internet]. Oxford University Press; 2013 [cited 2017 Oct 17];2:10. Available from:
581 https://academic.oup.com/gigascience/article-lookup/doi/10.1186/2047-217X-2-10
582 54. MacManes MD. The Oyster River Protocol: A Multi Assembler and Kmer Approach For de
583 novo Transcriptome Assembly. doi.org [Internet]. Cold Spring Harbor Laboratory; 2017 [cited
584 2017 Sep 21];177253. Available from: https://www.biorxiv.org/content/early/2017/08/16/177253
585 55. Aranda M, Li Y, Liew YJ, Baumgarten S, Simakov O, Wilson MC, et al. Genomes of coral
586 dinoflagellate symbionts highlight evolutionary adaptations conducive to a symbiotic lifestyle.
587 Sci. Rep. [Internet]. Nature Publishing Group; 2016 [cited 2017 Feb 28];6:39734. Available
588 from: http://www.nature.com/articles/srep39734
589