For GigaScience (Technical Note):
https://academic.oup.com/gigascience/pages/technical_note

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

(250 words maximum, separated into separate sections)

Background

De novo transcriptome assemblies are required prior to analyzing RNAseq data from a species without an existing reference. Despite its prevalence, there is a lack of consensus about the effects of using different pipelines on the assemblies. To test this, an automated pipeline was used to assemble and annotate raw data collected by the Marine Microbial Eukaryotic Transcriptome Sequencing Project (MMETSP). Assemblies were evaluated and compared with transcriptomes that were previously assembled with a different pipeline.

Findings

New assemblies contained 70% of the previous contigs as well as new content, with 7.8% of the annotated contigs identified by novel gene names compared to the previous assemblies. A higher number of unique gene names in the new assemblies suggests an increase in genic content. Assembly metrics varied by taxonomic group being assembled, with Dinophyta and Ciliophora groups demonstrating a higher percentage of open reading frames and number of contigs.

Conclusions

Automated pipelines are useful for processing large sets of samples, making it convenient to add additional samples and test different software tools. In addition, analyzing diverse sets of data using a common workflow pipeline provides opportunity for identifying taxon-specific trends. Streamlining workflows to re-assemble existing data in centralized and de-centralized repositories with new tools can potentially yield novel and useful results for the community using reference transcriptomes in downstream analyses.

the prevalence of transcriptom assembly

additional ?

showing?

we don't do this, right?

1st conclusion: diff software → diff, " better results.

Introduction

Reference transcriptomes are required for quantifying and profiling gene expression in biological samples. When organisms lack a reference transcriptome or genome, raw RNA sequence data (RNAseq) must be assembled into a *de novo* transcriptome. This type of analysis is ubiquitous in many fields, including evolutionary developmental biology, cancer biology, agriculture, ecological physiology, and biological oceanography. 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 [1–6].

Methods for *de novo* RNAseq assembly continue to evolve rapidly, especially for non-model species [7]. At this time, there are several major *de novo* transcriptome assembly software tools available to choose from, including Trinity [8], SOAPdenovo-Trans [9], Trans-ABySS [10], Oases [11], SPAdes [12], IDBA-tran [13], and Shannon [14]. The availability of these options stems from continued research into the unique computational challenges associated with transcriptome assembly, including large memory requirements, alternative splicing and allelic variants [8,15].

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. This workflow incorporates Trimmomatic, digital normalization with khmer software, and Trinity [16].

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 facilitate the investigation of diverse marine organisms that influence marine ecosystems and participate in biogeochemical cycling [17]. 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 them unusually comparable.

Reference transcriptomes for the MMETSP were originally assembled using a pipeline developed by the National Center for Genome Research (NCGR) [10]. These transcriptomes have already facilitated investigations in phylogenetic analyses [18], differential gene expression [2], and inter-group comparisons [19]. The meta-omic approaches [20] developed have led to interesting discoveries of gene diversity and evolutionary features [21–24].

In re-assembling the MMETSP data, we sought to compare and improve the original MMETSP reference transcriptomes 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, a lightweight protocol for assembling RNA-seq reads that uses the Trinity *de novo* transcriptome assembler [16]. This protocol generates *de novo* transcriptome assemblies of acceptable quality [1]. The pipeline was used to assemble all of the data from the MMETSP (Figure 1). The code and instructions for 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 (SraRunInfo.csv; Supplemental) 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) [25]. 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.

Initial assemblies were run by the National Center for Genome Resources (NCGR) using methods and data described in the original publication [10]. These transcriptomes were downloaded from the iMicrobe repository to compare with our re-assemblies: https://imicrobe.us/project/view/104

2. Perform quality control

Reads were analyzed with FastQC (version 0.11.5) and multiqc (version 1.2) [26] to confirm overall qualities before and after trimming. A conservative trimming approach was used [27] with Trimmomatic (version 0.33) [28] to remove residual Illumina adapters and cut bases off the start (LEADING) and end (TRAILING) of reads if they are 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 [32], Rfam [33], OrthoDB [34]. 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 [35]. 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 [35]. 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 [36]. A forward comparison was made with the NCGR assembly used as the reference and each DIB assembly as the query. Reverse comparative metrics were calculated with each DIB assembly as the reference and the NCGR assembly as the query.

Benchmarking Universal Single-Copy Orthologs (BUSCO) software (version 2) was used with a database of 234 orthologous genes specific to protistans with open reading frames in the assemblies. BUSCO scores are frequently used as one measure of assembly completeness [37].

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 [38]. Unique gene names were compared from a random subset of 296 samples using the dammit annotation pipeline (Scott 2016). 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

The DIB assemblies consistently ranked higher in evaluation metrics.

The majority of transcriptome evaluation metrics collected for each sample were higher in Trinity-based DIB assemblies than for the ABySS-based NCGR assemblies (Table 1 and Supplement 1, Table 1).

DIB 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 43,882 ± 26,116 while the mean number of contigs in the NCGR assemblies was 30,179 ± 21,341 (Figure 2). A two-sample Kolmogorov-Smirnov test comparing distributions indicated that the number of contigs were significantly different between DIB and NCGR assemblies ($p < 0.001$, $D = 0.29793$). Moreover, the Transrate scores [35], which represents of the DIB assemblies were higher. The mean Transrate score of all the DIB re-assemblies, 0.31 ± 0.1, was significantly higher than the mean score of the NCGR assemblies, 0.22 ± 0.09 ($p < 0.001$, $D = 0.48827$) (Figure 2).

The DIB 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 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 assembly. Extra content was in the DIB assembly that was not in the NCGR. The mean CRBB score in DIB when queried against NCGR as a reference was 0.70 ± 0.28, while the mean proportion for NCGR assemblies queried against DIB re-assemblies was 0.49 ± 0.09 ($p < 0.001$, $D = 0.7616$) (Figure 3). This indicates that more contigs from the NCGR assemblies were included in the DIB assemblies than vice versa, and also suggests that the DIB assemblies overall have additional content. This finding is reinforced by higher unique *k*-mer content found in the DIB assemblies compared to NCGR, where 84.4% of the samples fall above the 1:1 expectation indicating 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 81.9% ± 9.68 ORF content in DIB assemblies and 76.79% ± 10.2 ORF content in the NCGR assemblies. Nonetheless, ORF content in DIB assemblies was slightly higher than NCGR assemblies for 95% of the samples (Figure 5 - left). Secondly, when the assemblies were queried against the BUSCO database [37], the percentages of BUSCO matches in the DIB re-assemblies (61.4% ± 0.20) were not significantly different compared to the original NCGR assemblies (60% ± 0.19) ($p = 0.2096$, $D = 0.058348$) (Figure 5 - right). Thus, although the number of contigs and amount of content was increased in the DIB assemblies compared to the NCGR assemblies, the ORF content and contigs matching with the BUSCO database did not decrease, suggesting that the extra content might be biologically meaningful.

Following annotation by the dammit pipeline (Scott 2016), 91% ± 1.58 of the contigs in the DIB 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 assemblies, determined by a “false” CRBB result (Figure 6). Additionally, the number of unique gene names in DIB assemblies were higher than in NCGR, suggesting an increase in genic content (Figure 7).

Novel contigs in the DIB 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 annotated in the DIB assembly of the sample MMETSP1473 (*Stichococcus* sp.) that are absent in the NCGR assembly of this same sample. While other gene names, for example "Pkinase_Tyr","Bromodomain", and

"DnaJ", have positive annotation matches in the NCGR assembly and in the contigs identified as novel in the DIB assembly of sample MMETSP1473.

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). Assemblies from Ciliates (Ciliophora) had lower unique *k*-mers (Figure 8 D).

Discussion

Transcriptomics has been embraced across many fields. Though widely used, assembly of transcriptomes is typically performed on a small scale for one or a few species at a time. Taking a more holistic approach with a taxonomically-diverse dataset with automated tools such as the pipeline presented here, the reference transcriptome assemblies were improved for these species and broad scale phylogenetic trends were identified.

DIB 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 had no *a priori* expectation that the results would be similar, yet we found that in the majority of cases the new DIB 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.

Reassembly with new tools can yield new results

Evaluation with several different quality metrics suggested that the DIB assemblies were somewhat more inclusive than the NCGR assemblies. In addition to containing more contigs and being more inclusive of the NCGR assemblies than vice versa, the DIB assemblies had significantly higher Transrate scores, indicating better overall read inclusion in the assembled contigs. The DIB assemblies typically contained more *k*-mers, more annotated transcripts, and more unique gene names than the NCGR assemblies. These points all suggest that the additional content assembled with the DIB pipeline might be biologically meaningful. ~~Further~~

investigations into this content might be biologically meaningful, given the diversity of eukaryotic lineages that were sequenced in this project (Caron et al. 2017).

The evaluation metrics described here serve as a framework for better contextualizing the quality of protistan transcriptomes. For some species/strains in the MMETSP data set, these data represent the first nucleic acid sequence information available [17]. More reference data sets are needed to expand the range of known genes and functions available in protistan organisms [45].

Automated pipelines can be used to process arbitrarily many RNAseq samples

The automated and modularized nature of this pipeline is useful for processing large data sets like the MMETSP, and it allows 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 automation is obvious 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 set of species that span 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 [45]. The BUSCO paper assessed 70 genomes and 96 transcriptomes representing groups of diverse species (vertebrates, arthropods, other metazoans, fungi) [37]. Other benchmarking studies have examined transcriptome qualities for samples representing dozens of species from different taxonomic groupings [40,42].

Assembly evaluation tools yielded results outside the range of what is normal for some organisms, e.g. 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 [21–23]. 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 [46]. 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.

Conclusion

As the rate of sequencing data generation continues to increase, efforts to facilitated automated processing and evaluation of such data are increasingly important. This study has demonstrated that re-analyzing old data with new tools and methods improves the quality of reference transcriptome assemblies and expands the gene catalogue of the dataset. Notably, these improvements arose without further experimentation or sequencing. Automation tools were key in successfully processing and analyzing this large collection of 678 samples, allowing taxon-specific features to be identified because the pipelines were processing all samples together. With the growing volume of nucleic acid data in centralized and de-centralized repositories, streamlining methods into pipelines such as this can not only assist with the reproducibility of the analysis, but can help to identify features among diverse taxa from large collections of samples, showing that new and useful information can be discovered from re-analysis of existing data.

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. 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
2. 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
3. Vittoria Roncalli, Matthew C. Cieslaka, Stephanie A. Sommera RRH, Lenz PH. De novo transcriptome assembly of the calanoid copepod Neocalanus flemingeri: A new resource for emergence from diapause. Mar. Genomics [Internet]. Elsevier; 2017 [cited 2017 Sep 22]; Available from: http://www.sciencedirect.com/science/article/pii/S1874778717302155
4. Rana SB, Zadlock FJ, Zhang Z, Murphy WR, Bentivegna CS. Comparison of De Novo Transcriptome Assemblers and k-mer Strategies Using the Killifish, Fundulus heteroclitus.

Davies WIL, editor. PLoS One [Internet]. Public Library of Science; 2016 [cited 2017 Sep 22];11:e0153104. Available from: http://dx.plos.org/10.1371/journal.pone.0153104
5. 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
6. 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
7. 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
8. 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
9. 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
10. 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
11. 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
12. 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
13. 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
14. Kannan S, Hui J, Mazooji K. Shannon : An Information-Optimal de Novo RNA-Seq Assembler. 2016;1–14.
15. 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
16. 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_d

ocumentation/878460
17. 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
18. 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
19. 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
20. 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
21. 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
22. 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
23. 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
24. 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
25. 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
26. 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
27. MacManes MD. On the optimal trimming of high-throughput mRNA sequence data. Front. Genet. [Internet]. Frontiers; 2014 [cited 2017 Oct 17];5:13. Available from: http://journal.frontiersin.org/article/10.3389/fgene.2014.00013/abstract
28. 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

456 29. Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, Cartwright R, et al. The khmer
457 software package: enabling efficient nucleotide sequence analysis. F1000Research [Internet].
458 2015;4:900. Available from: http://f1000research.com/articles/4-900/v1
459 30. Zhang Q, Pell J, Canino-Koning R, Howe AC, Brown CT. These are not the K-mers you are
460 looking for: Efficient online K-mer counting using a probabilistic data structure. PLoS One.
461 2014;9.
462 31. Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH. A Reference-Free Algorithm for
463 Computational Normalization of Shotgun Sequencing Data. 2012 [cited 2017 Oct 17]; Available
464 from: http://arxiv.org/abs/1203.4802
465 32. Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein
466 families database: towards a more sustainable future. Nucleic Acids Res. [Internet]. Oxford
467 University Press; 2016 [cited 2017 Oct 17];44:D279–85. Available from:
468 https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv1344
469 33. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, et al. Rfam: updates to
470 the RNA families database. Nucleic Acids Res. [Internet]. Oxford University Press; 2009 [cited
471 2017 Oct 17];37:D136–40. Available from: https://academic.oup.com/nar/article-
472 lookup/doi/10.1093/nar/gkn766
473 34. Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, et al.
474 OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant,
475 archaeal, bacterial and viral orthologs. Nucleic Acids Res. [Internet]. Oxford University Press;
476 2017 [cited 2017 Sep 21];45:D744–9. Available from: https://academic.oup.com/nar/article-
477 lookup/doi/10.1093/nar/gkw1119
478 35. Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. TransRate: reference-free quality
479 assessment of de novo transcriptome assemblies. Genome Res. [Internet]. Cold Spring Harbor
480 Laboratory Press; 2016 [cited 2017 Oct 17];26:1134–44. Available from:
481 http://www.ncbi.nlm.nih.gov/pubmed/27252236
482 36. Aubry S, Kelly S, Kümpers BMC, Smith-Unna RD, Hibberd JM. Deep Evolutionary
483 Comparison of Gene Expression Identifies Parallel Recruitment of Trans-Factors in Two
484 Independent Origins of C4 Photosynthesis. Bomblies K, editor. PLoS Genet. [Internet]. Public
485 Library of Science; 2014 [cited 2017 Oct 17];10:e1004365. Available from:
486 http://dx.plos.org/10.1371/journal.pgen.1004365
487 37. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO:
488 assessing genome assembly and annotation completeness with single-copy orthologs.
489 Bioinformatics [Internet]. Oxford University Press; 2015 [cited 2017 Sep 21];31:3210–2.
490 Available from: https://academic.oup.com/bioinformatics/article-
491 lookup/doi/10.1093/bioinformatics/btv351
492 38. Jr LCI, Brown CT. Efficient cardinality estimation for k-mers in large DNA sequencing data
493 sets. 2016;1–5.
494 39. Titus Brown C, Irber L. sourmash: a library for MinHash sketching of DNA. J. Open Source
495 Softw. [Internet]. 2016 [cited 2017 Oct 17];1. Available from:
496 http://joss.theoj.org/papers/10.21105/joss.00027
497 40. MacManes MD. The Oyster River Protocol: A Multi Assembler and Kmer Approach For de
498 novo Transcriptome Assembly. doi.org [Internet]. Cold Spring Harbor Laboratory; 2017 [cited
499 2017 Sep 21];177253. Available from: https://www.biorxiv.org/content/early/2017/08/16/177253
500 41. O'Neil ST, Emrich SJ. Assessing De Novo transcriptome assembly metrics for consistency
501 and utility. BMC Genomics [Internet]. 2013;14:465. Available from:

http://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-14-465
42. Li B, Fillmore N, Bai Y, Collins M, Thomson JA, Stewart R, et al. Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biol. [Internet]. BioMed Central; 2014 [cited 2017 Oct 17];15:553. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-014-0553-5
43. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C. Salmon provides accurate, fast, and bias-aware transcript expression estimates using dual-phase inference. bioRxiv [Internet]. Nature Publishing Group; 2015;21592. Available from: http://biorxiv.org/content/early/2016/08/30/021592.abstract
44. Sibbald SJ, Archibald JM. More protist genomes needed. Nat. Ecol. Evol. [Internet]. Nature Publishing Group; 2017 [cited 2017 Oct 5];1:145. Available from: http://www.nature.com/articles/s41559-017-0145
45. Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, et al. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Gigascience [Internet]. Oxford University Press; 2013 [cited 2017 Oct 17];2:10. Available from: https://academic.oup.com/gigascience/article-lookup/doi/10.1186/2047-217X-2-10
46. Aranda M, Li Y, Liew YJ, Baumgarten S, Simakov O, Wilson MC, et al. Genomes of coral dinoflagellate symbionts highlight evolutionary adaptations conducive to a symbiotic lifestyle. Sci. Rep. [Internet]. Nature Publishing Group; 2016 [cited 2017 Feb 28];6:39734. Available from: http://www.nature.com/articles/srep39734