Note

Algae 2021, 36(4): 333-340
https://doi.org/10.4490/algae.2021.36.12.7
Open Access

A guide to phylotranscriptomic analysis for phycologists

Seongmin Cheon[1,2], Sung-Gwon Lee[1], Hyun-Hee Hong[1], Hyun-Gwan Lee[3], Kwang Young Kim[3] and Chungoo Park[1,*]

[1]*School of Biological Science and Technology, Chonnam National University, Gwangju 61186, Korea*
[2]*Department of Transdisciplinary Research and Collaboration, Biomedical Research Institute, Seoul National University Hospital, Seoul 03080, Korea*
[3]*Department of Oceanography, Chonnam National University, Gwangju 61186, Korea*

Phylotranscriptomics is the study of phylogenetic relationships among taxa based on their DNA sequences derived from transcriptomes. Because of the relatively low cost of transcriptome sequencing compared with genome sequencing and the fact that phylotranscriptomics is almost as reliable as phylogenomics, the phylotranscriptomic analysis has recently emerged as the preferred method for studying evolutionary biology. However, it is challenging to perform transcriptomic and phylogenetic analyses together without programming expertise. This study presents a protocol for phylotranscriptomic analysis to aid marine biologists unfamiliar with UNIX command-line interface and bioinformatics tools. Here, we used transcriptomes to reconstruct a molecular phylogeny of dinoflagellate protists, a diverse and globally abundant group of marine plankton organisms whose large and complex genomic sequences have impeded conventional phylogenic analysis based on genomic data. We hope that our proposed protocol may serve as practical and helpful information for the training and education of novice phycologists.

Key Words: dinoflagellate; phylotranscriptomics; protocol

INTRODUCTION

Understanding and refining phylogenetic relationships among species is a long-standing prerequisite for studying evolutionary biology. With access to many DNA sequences, researchers now routinely construct molecular phylogenetic trees using DNA or protein sequences. This approach is generally considered more reliable than morphological phylogenetic trees using morphological characters of species, given the low homoplasy, which confuses phylogenetic inference at the DNA or protein sequence level (Zou and Zhang 2016). The DNA sequences of a single or a few genes with considerable degrees of conservation across all species (e.g., small subunit ribosomal RNA) have been extensively used as molecular markers in phylogenetic studies. However, use of different genes often results in different trees (Rokas et al. 2003), partly due to sampling error or discordance between gene trees and species trees. Recent advances in genome sequencing technologies and substantial decreases in cost have made genomics readily available for phylogenetic studies. Phylogenomics, which involves reconstructing the phylogenetic and evolutionary history of a species by analyzing a large number of loci across the genome (Delsuc et al. 2005), can offer marked reductions in stochastic or sampling errors compared to those of the conventional approaches and has indeed led to several well-resolved phylogenies that likely represent the spe-

Received October 11, 2021, **Accepted** December 7, 2021
***Corresponding Author**
E-mail: chungoo@jnu.ac.kr
Tel: +82-62-530-1913, Fax: +82-62-530-2199

cies trees. However, although DNA sequencing costs have decreased considerably in recent years, obtaining high-quality genome assemblies with annotation is still quite expensive and labor-intensive. Because of these challenges, there is still considerable difficulty in detecting orthologous genes, which is a crucial step in phylogenetic inference.

Transcriptome sequencing, also known as RNA sequencing (RNA-seq), was originally developed to measure the transcript abundance of all expressed genes in a sample using direct sequencing of transcripts by high-throughput sequencing technologies (Wang et al. 2009, Martin and Wang 2011). RNA-seq is one of the most cost-effective and versatile methods for efficiently obtaining the DNA sequences of orthologous genes; with increased taxon sampling, a phylogeny can be reliably inferred. Indeed, many authors in recent years have employed phylotranscriptomics to resolve the evolutionary relationships of diverse lineages of organisms (Hittinger et al. 2010, Kocot et al. 2011, Struck et al. 2011, von Reumont et al. 2012, Riesgo et al. 2014, Wickett et al. 2014, Zeng et al. 2014, Irisarri et al. 2017, Janouskovec et al. 2017, Murat et al. 2017, Price and Bhattacharya 2017). Phylotranscriptomics is especially useful where there is a lack of genomic data, as is the case with marine organisms in general. In our recent study (Cheon et al. 2020), we demonstrated that phylotranscriptomic trees are virtually identical to phylogenomic trees, regardless of the tissue of origin of the transcriptome data, and showed that the success of phylotranscriptomics relies on rigorous orthologous gene identification. Consequently, given the relatively low cost of transcriptome sequencing compared with genome sequencing, we foresee wider adoption of phylotranscriptomics in evolutionary studies.

In the present study, we developed a protocol, intended as a guide to aid phycologists with little or no background in bioinformatics programming, using an integrative analysis of high-throughput sequencing data. We hope that our proposed protocol may serve as practical and valuable information for the training and education of novice marine biologists.

Application in marine organisms

A greater understanding of marine organisms will provide essential insights into the evolutionary history and diversity of eukaryotes (Caron et al. 2017, Burki et al. 2020, Strassert et al. 2021). With the advancement of next-generation sequencing technologies, large-scale phylogenetics using genome or transcriptome sequences has become the standard tool for many previously unresolvable lineages (Delsuc et al. 2005, Meusemann et al. 2010, Kocot et al. 2011, Struck et al. 2011, Wickett et al. 2014, Zeng et al. 2014, Janouskovec et al. 2017, Cheon et al. 2020, Song et al. 2020). Because of the relatively small number of marine species for which complete genome sequences are available, transcriptome-based phylogenetic analysis is becoming increasingly popular in the phylogeny of marine organisms (Kocot et al. 2011, von Reumont et al. 2012, Irisarri et al. 2017). Here, we used transcriptomes to reconstruct a molecular phylogeny of dinoflagellate protists, a diverse and globally abundant group of marine plankton organisms whose large and complex genomic sequences have impeded conventional phylogenic analysis based on genomic data.

MATERIALS AND METHODS

Overview of the protocol

There are four steps to inferring a phylotranscriptomic tree from RNA-seq–based transcriptomes. First, we collect RNA-seq data from publicly available RNA-seq data (or one can use one's own generated RNA-seq data). Second, we predict putative protein-coding sequences from *de novo* transcriptome assembly without a reference genome. Third, we identify orthologous genes that can be used for phylogenetic inference. Finally, a phylotranscriptomic tree is constructed based on a concatenated alignment of a set of inferred single copy (one-to-one) orthologs (Fig. 1). The details of our model protocol (using selected marine species) are given below.

Data preparation

We selected 20 dinoflagellate and one alveolate species as an outgroup, with publicly available high-throughput RNA-seq data. The datasets used are summarized in Table 1.

Transcriptome data processing

All raw RNA-seq data sets were initially processed using Trimmomatic (v0.39) to obtain high-quality clean reads with no adapter sequences, poly-N sequences, or low-quality bases (Bolger et al. 2014). The remaining clean reads from each sample were then used for *de novo* transcriptome assembly using Trinity (version 2.2.0) (Haas et al. 2013) with default parameters. After transcriptome

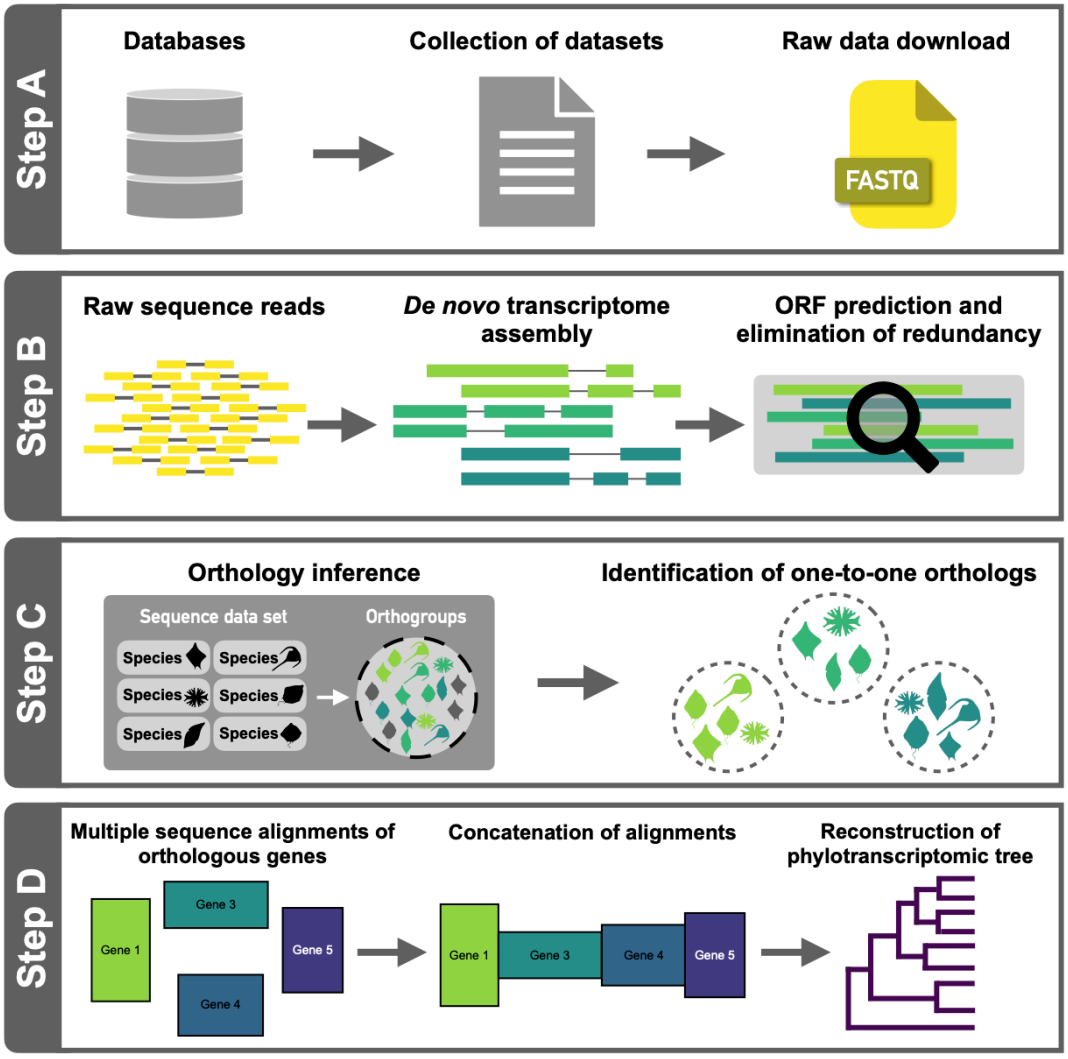

Fig. 1. Phylotranscriptomic analysis pipeline.

assembly, open reading frames (ORFs) were predicted using TransDecoder (version 3.0.0) (https://github.com/TransDecoder/TransDecoder/wiki) with BLASTP (Camacho et al. 2009) searches in the UniProtKB/Swiss-Prot database. ORFs of length less than 100 amino acids were discarded to avoid maintaining transcripts with insufficient evidence for protein-coding regions. To remove the redundant peptide sequences, protein sequences with >99% identity were clustered using the CD-HIT program (version 4.6.6) (Fu et al. 2012), and the longest sequence in each cluster was selected.

Inference of orthologous genes

From *de novo* assembled non-redundant protein-coding sequences in 21 transcriptomes (Table 1), putative orthologous genes were identified using the OrthoFinder tool (Emms and Kelly 2019). Pairwise sequence similarities between protein sequences were calculated using Diamond (Buchfink et al. 2015), and Markov clustering was applied with an inflation parameter of 1.5 for clustering orthogroups. To avoid complications introduced by paralogous genes in phylotranscriptomic inference, we excluded orthologous gene groups containing more than one gene from any given species and exclusively selected orthologous genes with genes present with one-to-one orthologs in at least 50% of the species in the group.

Construction of phylotranscriptomic tree

Amino acid sequences of one-to-one orthologous genes were aligned with Prank (http://wasabiapp.org/

software/prank/) using default options. The aligned sequences with more than 70% of gaps were trimmed using Phyutility *clean* (option: -clean 0.3) (Smith and Dunn 2008). Any trimmed alignments <150 amino acids in length were discarded, and the remaining trimmed sequences were concatenated with Phyutility *concat* to construct the supermatrix. A maximum-likelihood tree was inferred using IQ-TREE (Minh et al. 2020) with the LG + C60 + R + F model and 1,000 ultrafast bootstrap replication.

Materials

We have provided all source codes and other technical details on our GitHub website: https://github.com/CSB-SeongminCheon/Phyco-phylo.

Equipment and setup

Critical. All commands in this protocol are run in a Bash shell on the GNU/Linux operating system. It is important to ensure that each software tool mentioned in this protocol is available within user's your Bash PATH. The simplest way to do this is to run the following command:

$ ***echo*** 'export PATH="/home/user/local/tool_name: $PATH" ' >> ~/.bash_profile

'/home/user/local/tool_name' is the directory path of the application that is being installed.

Note. Consistent text formatting helps readers to interpret information. We use *italic* typeface for the Linux command line, and ***bold-italic*** typeface indicates the program name. The '$' character at the beginning of a line of command indicates Linux shell.

Hardware

- 32-core processor (recommend: >8-core processor)
- 128 Gb (Gigabytes) of RAM (recommend: >32 Gb of RAM)
- At least 1 Tb (Terabyte) of memory space required for raw and processed sequence data (note: the amount of memory required depends on the number of tax and the size of the RNA-seq datasets)

Software

- Ubuntu 20.04 LTS (recommend LTS version)
- Python v.3.6 with Biopython package (https://biopython.org/)
- SRA Toolkit v.2.11.0 (https://www.ncbi.nlm.nih.gov/home/tools/)
- Trimmomatic v.0.39 (http://www.usadellab.org/cms/?page=trimmomatic)

Table 1. Summary statistics of dinoflagellate transcriptomes used in this study

Species	SRA accession	Raw data		*De novo* transcriptome assembly	Open reading frame prediction
		Total bases (bp)	No. of reads	No. of assembled transcripts	No. of non-redundant unigenes
Alexandrium affine	SRR10426049	8,339,693,458	55,229,758	151,908	98,575
Alexandrium minutum	ERR1558595	6,829,726,800	68,297,268	159,701	91,556
Alexandrium pacificum	SRR10426048	8,609,710,450	57,017,950	153,976	96,747
Crypthecodinium cohnii	SRR5277468	4,473,194,700	29,821,298	180,463	84,768
Gonyaulax spinifera	SRR1300518	2,631,976,200	52,639,524	75,677	38,623
Lingulodinium polyedra	SRX090641	45,999,584,404	518,980,453	212,804	122,798
Protoceratium reticulatum	SRR1296738	2,343,479,600	46,869,592	108,095	65,535
Amphidinium carterae	SRR1610335	3,042,776,298	30,126,498	82,274	50,133
Gymnodinium catenatum	SRR1296705	3,673,570,000	73,471,400	121,935	75,074
Noctiluca scintillans	SRR1296929	2,498,412,800	49,968,256	65,313	44,709
Oxyrrhis marina	SRR1296901	6,209,966,400	62,099,664	136,039	63,723
Brandtodinium nutricula	SRR1300537	2,043,267,100	40,865,342	100,792	61,282
Durinskia baltica	SRR1296839	2,349,621,600	46,992,432	124,574	70,379
Heterocapsa arctica	SRR1300520	2,733,720,500	54,674,410	75,927	40,382
Heterocapsa rotundata	SRR1296810	2,162,826,300	43,256,526	70,471	41,409
Heterocapsa triquetra	SRR1296978	2,267,966,600	45,359,332	100,147	53,038
Kryptoperidinium foliaceum	SRR1296840	3,674,222,600	73,484,452	118,442	57,931
Polarella glacialis	SRR401178	5,335,148,052	68,399,334	195,466	80,746
Symbiodinium minutum	DRR003869	5,964,317,250	79,524,230	91,950	52,830
Scrippsiella trochoidea	SRX1032815	13,980,008,700	77,666,715	683,846	252,310
Perkinsus marinus	SRR1154655	3,073,904,064	32,019,834	40,037	18,098

- Samtools v.1.12 (http://www.htslib.org)
- Bowtie v.2.3 (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml)
- Trinity v.2.2.0 (https://github.com/trinityrnaseq/trinityrnaseq)
- Transdecoder v.3.0.0 (https://github.com/TransDecoder/TransDecoder)
- BLAST+ v.2.9 (https://www.ncbi.nlm.nih.gov/home/download/)
- CD-hit v.4.6.6 (http://weizhongli-lab.org/cd-hit/)
- OrthoFinder v.2.5.2 (https://github.com/davidemms/OrthoFinder)
- DIAMOND v.0.9.24 (https://github.com/bbuchfink/diamond/releases)
- MCL v.14.137 (https://micans.org/mcl/)
- Prank v.17 (http://wasabiapp.org/software/prank/)
- Phyutility v.2.7.1 (https://github.com/blackrim/phyutility)
- IQ-Tree v.2.1.2 (http://www.iqtree.org/)

Executable files of all the above programs should be created and saved in your working directory, such as '/home/user/bin'. Please see the following GitHub website for details of the installation or compile process on Linux/UbuntuOS operation system: https://github.com/CSB-SeongminCheon/Phyco-phylo.

Procedure

Data preparation.

1. Download 21 RNA-seq data (Table 1) from the NCBI SRA database (https://www.ncbi.nlm.nih.gov/sra) and convert the SRA files into FASTQ files using the fastq-dump program of the SRA Toolkit. The collected dataset and command lines can be found at the GitHub site: https://github.com/CSB-SeongminCheon/Phyco-phylo.

$ ***fastq-dump*** *--defline-seq '@$sn[_$rn]/$ri' --split-files SRA_AccessionID*

Preprocessing of RNA-seq data and *de novo* transcriptome assembly.

2. Obtain high-quality clean reads and assemble them using Trinity software. This tool must work separately on each RNA-seq sample. Please note that the output directory name must include 'trinity.'

$ ***Trinity*** *--seqType fq --trimmomatic --quality_trimming_params "ILLUMINACLIP:/home/your/path/trinity-plugins/Trimmomatic-0.36/adapters/TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36" --max_memory 200G --CPU 32 --full_cleanup --left forward_reads_1.fastq --right reverse_reads_2.fastq --output trinity_output_Name*

Prediction of non-redundant protein-coding sequences.

3. Predict ORFs from each *de novo* assembled transcript using TransDecoder.

$ ***TransDecoder.LongOrfs*** *-t trinity_output_Name/Trinity.fasta -S*

4. Download UniProt/Swiss-Prot protein databases, which are manually reviewed and more reliable.

$ ***wget*** https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/*complete/uniprot_sprot.fasta.gz*

$ ***gzip*** *-d uniport_sprot.fata.gz*

5. Make a BLAST database with the UniProt/Swiss-Prot protein databases using makeblastdb program from BLAST+.

$ ***makeblastdb*** *-i uniport_sprot.fasta -dbtype prot*

6. Execute BLASTP to compare the output of TransDecoder.LongOrfs with the UniProt/Swiss-Prot protein databases.

$ ***blastp*** *-query Trinity.fasta.transdecoder_dir/longest_orfs.pep -db uniprot_sprot.fasta -max_target_seqs 1 -outfmt 6 -evalue 10 -num_threads 32 -out Genus_Species.outfmt6*

7. Identify ORFs using the TransDecoder tool.

$ ***TransDecoder.Predict*** *-t Trinity.fasta –retain_blastp_hits Genus_Species.outfmt6*

8. Cluster sequences with 99% similarity, and select the longest sequence in each cluster as a representative sequence of that cluster by using CD-HIT:

$ ***cd-hit*** *-I Trinity.fasta.transdecoder.pep -o Genus_Species.fa.cdhit -c 0.99 -n 5*

9. Modify a description line of the FASTA formatted CD-HIT outputs to prevent special characters (e.g., "-", and "*") from hindering downstream analyses. The corresponding Python code can be found at the GitHub site: https://github.com/CSB-SeongminCheon/Phyco-phylo/blob/main/fix_names_from_CDhit.py.

$ *Python* ***fix_names_from_CDhit.py*** *Genus_Species.fa.cdhit GenusName SpeciesName*

Inference of orthologous genes.

10. Run OrthoFinder to infer the orthogroups.

$ ***orthofinder*** *-f WorkingDirectory/*

11. Select orthologous genes with genes present with one-to-one orthologs in at least 50% of the species in the group. The corresponding Python code can be found at the GitHub site: https://github.com/CSB-SeongminCheon/Phyco-phylo/blob/main/singlecopy_from_OrthoFinder.py.

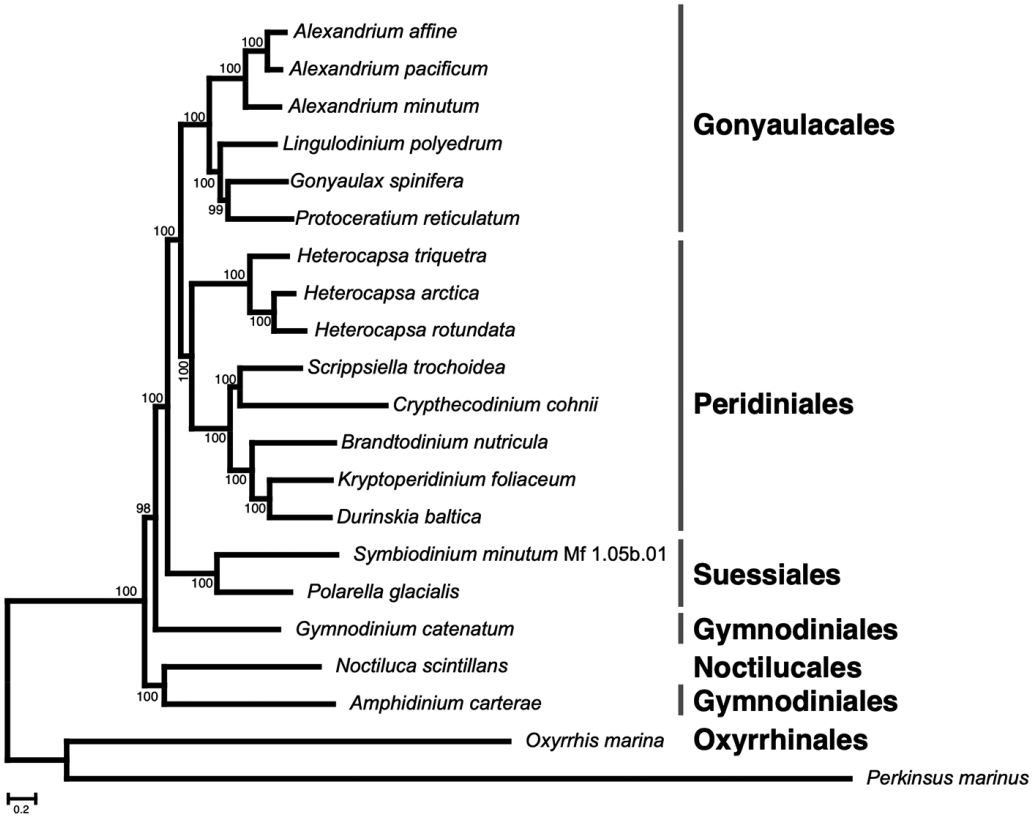

Fig. 2. Dinoflagellate phylotranscriptomic tree.

$ *python* ***singlecopy_from_OrthoFinder.py*** *WorkingDirectory SingleCopyResults*

Generation of phylotranscriptomic tree.

12. Perform multiple sequence alignment for each one-to-one orthogroup with prank_wrapper Python script implementing Prank alignment software. The corresponding Python code can be found at the GitHub site: https://github.com/CSB-SeongminCheon/Phyco-phylo/blob/main/prank_Wrapper.py.

$ *python* ***prank_Wrapper.py*** *SingleCopyResults*

13. Identify and remove poorly aligned sequences with >70% of gaps using Phyutility. The corresponding Python code can be found at the GitHub site: https://github.com/CSB-SeongminCheon/Phyco-phylo/blob/main/phyutility_Wrapper.py.

$ *python* ***phyutility_Wrapper.py*** *SingleCopyResults 0.3*

14. Concatenate multiple alignments of the one-to-one orthogroup, and only select concatenated alignments with >150 amino acids for the downstream analysis of phylotranscriptomic tree. The corresponding Python code can be found at the GitHub site: https://github.com/CSB-SeongminCheon/Phyco-phylo/blob/main/supermatrix_concatenate.py.

$ *python* ***supermatrix_concatenate.py*** *SingleCopyResults 150 11 Dinoflagellate_Supermatrix*

15. Run IQ-Tree tool with the LG+C60+R+F model and 1,000 ultrafast bootstrap replication.

$ ***iqtree*** *-s Dinoflagellate_Supermatrix.phy -spp Dinoflagellate_Supermatrix.model -m LG+C60+R+F -bb 1000 -nt 80*

RESULTS AND DISCUSSION

Anticipated results

Our protocol was adopted for reconstructing a dinoflagellate phylotranscriptomic tree. After preprocessing and assembling procedures of 21 transcriptome data sets, the average number of non-redundant putative protein-coding sequences predicted per sample was 74,316 (Table 1). When gene orthology inference was performed with OrthoFinder, 351,269 orthogroups present in more than one unigene within or between species were identified. This study restricted our investigation to only 306 ortho-

groups containing exactly one gene in at least 11 species (out of 21; greater than 50%). After excluding multiple sequence alignments of less than 150 amino acid residues, the final alignment of 64,147 aligned positions from 250 orthogroups was used for maximum-likelihood analysis using IQ-Tree under the LG + C60 + R + F model with 1,000 replications. The reconstructed dinoflagellate phylotranscriptomic tree is resolved with each interior branch, with >98% bootstrap support (Fig. 2), and its tree topology is mainly congruent with previously published molecular phylogenetic trees (Janouskovec et al. 2017, Price and Bhattacharya 2017, Stephens et al. 2018), with some exceptions. For example, in our tree, the dinoflagellates *Noctiluca scintillans* and *Amphidinium carterae* are sister taxa, although they are not part of the same clade in the tree constructed by Stephens et al. (2018), and the position of Suessiales clade is incongruent between our tree and the one proposed by JanouŠkovec et al. (2017). Such slight discrepancies in phylogenetic topologies may be due to the number of taxa studied and the different ortholog identification algorithms.

CONCLUSION

Orthology inference is the most crucial step in transcriptome-based phylogenetics, but can be challenging due to the incomplete genomic data and the complexity of transcriptome data. Therefore, in this study, we present a user-friendly protocol for phylotranscriptomic analysis to assist field and experimental biologists unfamiliar with UNIX command-line interface and bioinformatics tools. With a practical case study of reconstructing the dinoflagellate phylotranscriptomic tree, it is suggested that transcriptome sequence data may become the standard for generating large and phylogenetically informative data sets and will offer fascinating insights into the understanding of evolutionary history of organisms and their genomes. Finally, we hope that our proposed protocol may serve as practical and valuable information for the training and education of novice marine biologists.

ACKNOWLEDGEMENTS

We thank the CSB lab members. This research was supported by the Basic Science Research Program of the National Research Foundation of Korea (NRF), funded by the Ministry of Education (NRF-2019R1F1A1062411 to CP, NRF-2016R1A6A1A03012647 to H-G Lee, NRF-2020R1A2C3005053 to KYK) and by the "Research center for fishery resource management based on the information and communication technology," funded by the Ministry of Oceans and Fisheries, Korea (2021, grant number 20180384 to CP).

CONFLICTS OF INTEREST

Kwang Young Kim serves as editor for the Algae, but has no role in the decision to publish this article. All remaining authors have declared no conflicts of interest.

REFERENCES

Bolger, A. M., Lohse, M. & Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Buchfink, B., Xie, C. & Huson, D. H. 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Methods 12:59–60.

Burki, F., Roger, A. J., Brown, M. W. & Simpson, A. G. B. 2020. The new tree of eukaryotes. Trends Ecol. Evol. 35:43–55.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. & Madden, T. L. 2009. BLAST+: architecture and applications. BMC Bioinformatics 10:421.

Caron, D. A., Alexander, H., Allen, A. E., Archibald, J. M., Armbrust, E. V., Bachy, C., Bell, C. J., Bharti, A., Dyhrman, S. T., Guida, S. M., Heidelberg, K. B., Kaye, J. Z., Metzner, J., Smith, S. R. & Worden, A. Z. 2017. Probing the evolution, ecology and physiology of marine protists using transcriptomics. Nat. Rev. Microbiol. 15:6–20.

Cheon, S., Zhang, J. & Park, C. 2020. Is phylotranscriptomics as reliable as phylogenomics? Mol. Biol. Evol. 37:3672–3683.

Delsuc, F., Brinkmann, H. & Philippe, H. 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6:361–375.

Emms, D. M. & Kelly, S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20:238.

Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152.

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. &

Regev, A. 2013. *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. 8:1494–1512.

Hittinger, C. T., Johnston, M., Tossberg, J. T. & Rokas, A. 2010. Leveraging skewed transcript abundance by RNA-Seq to increase the genomic depth of the tree of life. Proc. Natl. Acad. Sci. U. S. A. 107:1476–1481.

Irisarri, I., Baurain, D., Brinkmann, H., Delsuc, F., Sire, J. -Y., Kupfer, A., Petersen, J., Jarek, M., Meyer, A., Vences, M. & Philippe, H. 2017. Phylotranscriptomic consolidation of the jawed vertebrate timetree. Nat. Ecol. Evol. 1:1370–1378.

Janouškovec, J., Gavelis, G. S., Burki, F., Dinh, D., Bachvaroff, T. R., Gornik, S. G., Bright, K. J., Imanian, B., Strom, S. L., Delwiche, C. F., Waller, R. F., Fensome, R. A., Leander, B. S., Rohwer, F. L. & Saldarriaga, J. F. 2017. Major transitions in dinoflagellate evolution unveiled by phylotranscriptomics. Proc. Natl. Acad. Sci. U. S. A. 114:E171–E180.

Kocot, K. M., Cannon, J. T., Todt, C., Citarella, M. R., Kohn, A. B., Meyer, A., Santos, S. R., Schander, C., Moroz, L. L., Lieb, B. & Halanych, K. M. 2011. Phylogenomics reveals deep molluscan relationships. Nature 477:452–456.

Martin, J. A. & Wang, Z. 2011. Next-generation transcriptome assembly. Nat. Rev. Genet. 12:671–682.

Meusemann, K., von Reumont, B. M., Simon, S., Roeding, F., Strauss, S., Kück, P., Ebersberger, I., Walzl, M., Pass, G., Breuers, S., Achter, V., von Haeseler, A., Burmester, T., Hadrys, H., Wägele, J. W. & Misof, B. 2010. A phylogenomic approach to resolve the arthropod tree of life. Mol. Biol. Evol. 27:2451–2464.

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A. & Lanfear, R. 2020. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol. Biol. Evol. 37:1530–1534.

Murat, F., Armero, A., Pont, C., Klopp, C. & Salse, J. 2017. Reconstructing the genome of the most recent common ancestor of flowering plants. Nat. Genet. 49:490–496.

Price, D. C. & Bhattacharya, D. 2017. Robust Dinoflagellata phylogeny inferred from public transcriptome databases. J. Phycol. 53:725–729.

Riesgo, A., Farrar, N., Windsor, P. J., Giribet, G. & Leys, S. P. 2014. The analysis of eight transcriptomes from all poriferan classes reveals surprising genetic complexity in sponges. Mol. Biol. Evol. 31:1102–1120.

Rokas, A., Williams, B. L., King, N. & Carroll, S. B. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425:798–804.

Smith, S. A. & Dunn, C. W. 2008. Phyutility: a phyloinformatics tool for trees, alignments and molecular data. Bioinformatics 24:715–716.

Song, H., Béthoux, O., Shin, S., Donath, A., Letsch, H., Liu, S., McKenna, D. D., Meng, G., Misof, B., Podsiadlowski, L., Zhou, X., Wipfler, B. & Simon, S. 2020. Phylogenomic analysis sheds light on the evolutionary pathways towards acoustic communication in Orthoptera. Nat. Commun. 11:4939.

Stephens, T. G., Ragan, M. A., Bhattacharya, D. & Chan, C. X. 2018. Core genes in diverse dinoflagellate lineages include a wealth of conserved dark genes with unknown functions. Sci. Rep. 8:17175.

Strassert, J. F. H., Irisarri, I., Williams, T. A. & Burki, F. 2021. A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. Nat. Commun. 12:1879.

Struck, T. H., Paul, C., Hill, N., Hartmann, S., Hösel, C., Kube, M., Lieb, B., Meyer, A., Tiedemann, R., Purschke, G. & Bleidorn, C. 2011. Phylogenomic analyses unravel annelid evolution. Nature 471:95–98.

von Reumont, B. M., Jenner, R. A., Wills, M. A., Dell'ampio, E., Pass, G., Ebersberger, I., Meyer, B., Koenemann, S., Iliffe, T. M., Stamatakis, A., Niehuis, O., Meusemann, K. & Misof, B. 2012. Pancrustacean phylogeny in the light of new phylogenomic data: support for Remipedia as the possible sister group of Hexapoda. Mol. Biol. Evol. 29:1031–1045.

Wang, Z., Gerstein, M. & Snyder, M. 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. 10:57–63.

Wickett, N. J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M. S., Burleigh, J. G., Gitzendanner, M. A., Ruhfel, B. R., Wafula, E., Der, J. P., Graham, S. W., Mathews, S., Melkonian, M., Soltis, D. E., Soltis, P. S., Miles, N. W., Rothfels, C. J., Pokorny, L., Shaw, A. J., DeGironimo, L., Stevenson, D. W., Surek, B., Villarreal, J. C., Roure, B., Philippe, H., dePamphilis, C. W., Che, T., Deyholos, M. K., Baucom, R. S., Kutchan, T. M., Augustin, M. M., Wang, J., Zhang, Y., Tian, Z., Yan, Z., Wu, X., Sun, X., Wong, G. K. -S. & Leebens-Mack, J. 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. Proc. Natl. Acad. Sci. U. S. A. 111:E4859–E4868.

Zeng, L., Zhang, Q., Sun, R., Kong, H., Zhang, N. & Ma, H. 2014. Resolution of deep angiosperm phylogeny using conserved nuclear genes and estimates of early divergence times. Nat. Commun. 5:4956.

Zou, Z. & Zhang, J. 2016. Morphological and molecular convergences in mammalian phylogenetics. Nat. Commun. 7:12758.