Kristina Terpis

BES539

Spring 2020

Microorganisms are found all around us, are important to biogeochemical cycles, are human pathogens, and are the base of the food web. The microbial eukaryotes, or protists, are incredibly diverse. Early on, researchers grouped protists based on the few morphological features they could find. However, roughly 2 billion years of evolution produced morphologies that are impossible to use to reconstruct the relationships between major eukaryotic lineages (Keeling & Burki, 2019). Across the tree of life, many taxa have been recognized as important for their roles in human health, oxygen production, and global ecology. Genome-scale research has justifiably focused on taxa with significant human impact, but lesser known taxonomic groups may play important roles in ecosystems as well.

The stramenopiles, alveolates, and Rhizaria (SAR) form one of the most diverse major eukaryotic clades on earth, which contains well known photosynthetic lineages such as diatoms and kelp (stramenopiles) and the dinoflagellates (alveolates). Some members of SAR have been intensely studied such as the parasites *Blastocystis* (stramenopile) and *Plasmodium* (Alveolata). However, many members of this super group are understudied resulting in much to be learned about their role within their environments (Lin et al., 2012; Massana et al., 2004).

Stramenopiles are a diverse phylum including heterotrophic lineages and a photosynthetic (Ochrophyta) clade (Cavalier-Smith et al., 1995). The heterotrophic lineages contain the oomycetes or water molds, which were the causative agent of the Irish Potato Famine of the 1800s. The ochrophytes include members ranging from large multicellular kelp to microscopic unicellular protists. They are found in marine, freshwater, and terrestrial habitats and exhibit a wide range of nutritional strategies. There are classes within the stramenopiles that are parasitic, obtain their food through by grazing on bacteria or other pico-eukaryotes, photosynthesize, or are mixotrophic meaning that there are able to feed by either photosynthesis or phagocytosis. There is a great deal of undescribed diversity within both of these groups. Surveys of marine environments have revealed have found stramenopiles make up a large fraction of the poorly studied marine picoeukaryotes (Massana et al., 2014; Pernice et al., 2016). Many groups have distinctive morphology, whereas others lack identifying characteristics and are only distinguishable based on molecular data. Additionally, losses of certain accessory pigments have led to some species incorrectly assigned to other stramenopile classes or even other phyla. The lack of morphology for most groups has consequences for inferring their evolutionary relationships (Derelle et al., 2016).

I am looking to identify single copy shared orthologs between stramenopiles that represent the 4 different nutritional strategies. I have selected *Aplanochytrium* as the parasitic representative, *Cafeteria roenbergensis* which is a bacterivorous heterotrophic flagellate, *Cylindrotheca*, a diatom that is purely autotrophic, and *Dinobryon*, a mixotroph. All four of the selected species are from a marine environment. The goal is to see how many genes are assigned to an orthogroup, and how many single copy orthologs are shared between all four taxa.The data used in this analysis were generated as part of the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP). This project sequenced over 650 novel transcriptomes from across the tree of life (Keeling et al., 2014). Of the data generated, 269 transcriptomes were stramenopiles. There was a strong focus on generating data for the stramenopiles due to the low abundance of available data relative to the enormity of the group. This script allows the user with one script, to start with raw reads, and finish with orthogroups. This script could be easily expanded to include more data. However, the trimming parameters would need to be set based on the quality of transcriptome being added, and the new names included in the renaming steps in both the shell and R scripts.

Once the raw reads have been downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) using the SRA ToolKit, they are analyzed for quality using FastQC. The trimming parameters are set based on the outputs from FastQC and the trimming is performed using Trimmomatic. The trimmed, paired reads are then *de novo* assembled using RNASPAdes v3.11 (Bushmanova et al., 2018) and the coding regions predicted using TransDecoder (Haas et al., 2013). The headers are then renamed and shortened using a combination of sed and awk scripts. The purpose of renaming the headers is to add identity information to each sequence and once the files have been renamed, they are run through OrthoFinder (Emms & Kelly, 2015). One of the outputs from OrthoFinder is a directory containing files on comparative genome statistics. The file called Statistics per Species is contained in this directory and used in the R script to visualize the outputs. The R script begins with reading in the tab separated values (tsv) file. Since the first ten rows are what I am looking to see if any relationships exist, the slice function is used and the dataframe called first\_10\_rows is created. However, the when tsv was imported, the data contained within it was imported as factors instead of numeric. A function is created to convert the factor values to numeric, then plots are then created using ggplot.

Once the script has run to completion, we can see that there are 4479 genes for *Aplanochytrium*, 3271 for *Cafeteria*, 27835 for *Cylindrotheca*, and 6065 for *Dinobryon*. One of the factors that could be leading to *Cylindrotheca* having more genes is because when it was sequenced, it was sequenced using 2x100bp, while the other 3 were sequenced using 2x50bp. Even so, there are interesting trends with this data. Both *Aplanochytrium* and *Cafeteria* contain 2 species specific orthogroups, while *Cylindrotheca* and *Dinobryon* contain 8 and 7 respectively. This could be due to more genes being required to perform photosynthesis. This same trend holds for the number of genes in species-specific orthogroups. Additionally, there are a total of 37 single copy shared orthologs between all four species. Some additional analyses that could be performed would be to figure out which genes make up the 37 single copy shared genes and to see how many and which genes are shared between the mixotroph and the autotroph, and the mixotroph with the two heterotrophic species. This could be accomplished by BLAST-ing the genes to NCBI’s protein database, or the InterProScan to determine protein domains.

References:

Bushmanova, E., Antipov, D., Lapidus, A., & Przhibelskiy, A. D. (2018). rnaSPAdes: A de novo transcriptome assembler and its application to RNA-Seq data. *BioRxiv*, 420208. https://doi.org/10.1101/420208

Cavalier-Smith, T., Chao, E. E., & Allsopp, M. T. E. P. (1995). Ribosomal RNA Evidence for Chloroplast Loss within Heterokonta: Pedinellid Relationships and a Revised Classification of Ochristan Algae. *Archiv Für Protistenkunde*, *145*(3–4), 209–220. https://doi.org/10.1016/S0003-9365(11)80316-7

Derelle, R., López-García, P., Timpano, H., & Moreira, D. (2016). A Phylogenomic Framework to Study the Diversity and Evolution of Stramenopiles (=Heterokonts). *Molecular Biology and Evolution*, *33*(11), 2890–2898. https://doi.org/10.1093/molbev/msw168

Emms, D. M., & Kelly, S. (2015). OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, *16*(1), 157. https://doi.org/10.1186/s13059-015-0721-2

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., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-Seq: Reference generation and analysis with Trinity. *Nature Protocols*, *8*(8). https://doi.org/10.1038/nprot.2013.084

Keeling, P. J., & Burki, F. (2019). Progress towards the Tree of Eukaryotes. *Current Biology*, *29*(16), R808–R817. https://doi.org/10.1016/j.cub.2019.07.031

Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., Armbrust, E. V., Archibald, J. M., Bharti, A. K., Bell, C. J., Beszteri, B., Bidle, K. D., Cameron, C. T., Campbell, L., Caron, D. A., Cattolico, R. A., Collier, J. L., Coyne, K., Davy, S. K., … Worden, A. Z. (2014). The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLOS Biology*, *12*(6), e1001889. https://doi.org/10.1371/journal.pbio.1001889

Lin, Y.-C., Campbell, T., Chung, C.-C., Gong, G.-C., Chiang, K.-P., & Worden, A. Z. (2012). Distribution Patterns and Phylogeny of Marine Stramenopiles in the North Pacific Ocean. *Applied and Environmental Microbiology*, *78*(9), 3387–3399. https://doi.org/10.1128/AEM.06952-11

Massana, R., Castresana, J., Balagué, V., Guillou, L., Romari, K., Groisillier, A., Valentin, K., & Pedrós-Alió, C. (2004). Phylogenetic and Ecological Analysis of Novel Marine Stramenopiles. *Applied and Environmental Microbiology*, *70*(6), 3528–3534. https://doi.org/10.1128/AEM.70.6.3528-3534.2004

Massana, R., del Campo, J., Sieracki, M. E., Audic, S., & Logares, R. (2014). Exploring the uncultured microeukaryote majority in the oceans: Reevaluation of ribogroups within stramenopiles. *The ISME Journal*, *8*(4), 854–866. https://doi.org/10.1038/ismej.2013.204

Pernice, M. C., Giner, C. R., Logares, R., Perera-Bel, J., Acinas, S. G., Duarte, C. M., Gasol, J. M., & Massana, R. (2016). Large variability of bathypelagic microbial eukaryotic communities across the world’s oceans. *The ISME Journal*, *10*(4), 945–958. https://doi.org/10.1038/ismej.2015.170