Monodopsis and Vischeria Genomes Shed New Light on the Biology of Eustigmatophyte Algae

Hsiao-Pei Yang1, Marius Wenzel2, Duncan A. Hauser1, Jessica M. Nelson1, Xia Xu1, Marek Eliáš3, and Fay-Wei Li1,4,*

1Boyce Thompson Institute, Ithaca, New York, USA
2School of Biological Sciences, University of Aberdeen, Aberdeen, United Kingdom
3Department of Biology and Ecology, Faculty of Science, University of Ostrava, Ostrava, Czech Republic
4Plant Biology Section, Cornell University, USA

*Corresponding author: E-mail: fl329@cornell.edu.
Accepted: 9 October 2021

Abstract

Members of eustigmatophyte algae, especially Nannochloropsis and Microchloropsis, have been tapped for biofuel production owing to their exceptionally high lipid content. Although extensive genomic, transcriptomic, and synthetic biology toolkits have been made available for Nannochloropsis and Microchloropsis, very little is known about other eustigmatophytes. Here we present three near-chromosomal and gapless genome assemblies of Monodopsis strains C73 and C141 (60 Mb) and Vischeria strain C74 (106 Mb), which are the sister groups to Nannochloropsis and Microchloropsis in the order Eustigmatales. These genomes contain unusually high percentages of simple repeats, ranging from 12% to 21% of the total assembly size. Unlike Nannochloropsis and Microchloropsis, long interspersed nuclear element repeats are abundant in Monodopsis and Vischeria and might constitute the centromeric regions. We found that both mevalonate and nonmevalonate pathways for terpenoid biosynthesis are present in Monodopsis and Vischeria, which is different from Nannochloropsis and Microchloropsis that have only the latter. Our analysis further revealed extensive spliced leader trans-splicing in Monodopsis and Vischeria at 36–61% of genes. Altogether, the high-quality genomes of Monodopsis and Vischeria not only serve as the much-needed outgroups to advance Nannochloropsis and Microchloropsis research, but also shed new light on the biology and evolution of eustigmatophyte algae.

Key words: Nannochloropsis, spliced leader trans-splicing, simple sequence repeats, LINE, Stramenopiles.

Significance

Our current knowledge of eustigmatophytes mostly comes from the biofuel algae Nannochloropsis and Microchloropsis. Here we generated three high-quality genomes of Monodopsis and Vischeria that are sister to Nannochloropsis + Microchloropsis. We uncovered an extremely high prevalence of simple repeats in these genomes and found evidence of spliced leader trans-splicing. These new genomic resources will greatly facilitate future research to better understand the biology of eustigmatophytes, and to better capitalize on their translational potential.

Introduction

The diversity of algae is vast but largely unexplored. Despite their often inconspicuous nature, algae have played pivotal roles in Earth’s biogeochemical cycles (de Vargas et al. 2015), and some might hold the key to sustainable bioenergy production (Radakovits et al. 2010; Jagadevan et al. 2018). Eustigmatophytes (Class Eustigmatophyceae), a lineage in Ochrophyta (Stramenopiles), are single-celled cocoid algae that can be found in freshwater, soil, and marine environments (Eliáš et al. 2017). The phylogeny and taxonomy of
this group have only been recently clarified (Fawley et al. 2014, 2015; Eliš et al. 2017; Ševčíková et al. 2019; Amaral et al. 2020). To date, there are around 20 genera and 189 species described according to AlgaeBase (Guiry MD and Guiry GM 2021), although this classification substantially underestimates the actual diversity of the class (Fawley et al. 2021).

The eustigmatophytes that have garnered the most attention are undoubtedly *Nannochloropsis* and the recently segregated *Microchloropsis* (Fawley et al. 2015). Many *Nannochloropsis* and *Microchloropsis* species are capable of producing a staggering amount of lipids, up to 60% of the total dry weight (Eliš et al. 2017). Because of this, as well as their fast growth rate, much research effort has been devoted to developing *Nannochloropsis* and *Microchloropsis* as an industrial biofuel alga (Eliš et al. 2017; Jagadevan et al. 2018). The genomes of most *Nannochloropsis* and *Microchloropsis* species, and in some cases multiple strains of species, have been sequenced (Pan et al. 2011; Radakovits et al. 2012; Vieler et al. 2012; Corteggiani Carpinelli et al. 2014; Wang et al. 2014; Schwartz et al. 2018; Brown et al. 2019; Guo et al. 2019; Ohan et al. 2019; Gong et al. 2020). However, only a few assemblies have reached high contig continuity and completeness (fig. 1A). In addition, tools for genetic transformation, gene editing, and marker-less trait-stacking have also been developed (Radakovits et al. 2012; Vieler et al. 2012; Wei et al. 2017; Poliner et al. 2018, 2020; Verruto et al. 2018; Naduthodi et al. 2019; Osorio et al. 2019). The applications of these tools and resources have resulted in substantial improvements of lipid production in *Microchloropsis* (previously *Nannochloropsis*) *gaditana* (Ajawi et al. 2017).

Relatively little is known about the genome structure of eustigmatophytes beyond *Nannochloropsis/Microchloropsis*. 

---

**Fig. 1.** Comparisons of eustigmatophyte genomes. (A) The three genomes reported here (in bold) have the highest BUSCO proteome completeness scores compared with the currently available *Nannochloropsis/Microchloropsis* genomes. The “Stramenopile” data set (n = 100) was used in the BUSCO analyses. The phylogeny on the left was based on 1,302 single-copy loci, and all branches receive bootstrap support of 100. The rooting was determined by OrthoFinder, which is consistent with the published phylogenies (Ševčíková et al. 2019). (B) Overall genome size (left panel) correlates well with repeat content (right panel). Significant expansions of simple repeats and LINEs are evident in *Vischeria* and *Monodopsis* genomes.
To date, most of the research on other eustigmatophytes has focused on the organellar genomes (Ševčiková et al. 2016, 2019; Yurchenko et al. 2016; Huang et al. 2019) and the association with a novel endosymbiont Candidatus Phycorickettsia (Yurchenko et al. 2018). Despite many interesting findings that have emerged from these studies, the lack of sequenced genomes throughout eustigmatophytes is limiting further research. Recently, a draft genome of Eustigmatos sp. was published as a part of large-scale survey of algal genomic diversity (Nelson et al. 2021). This assembly, however, was fragmented (contig N50 = 102 kb) and was not annotated.

Here we report three near-chromosomal genome assemblies of Monodopsis spp. (C73, C141) and Vischeria sp. (C74). Monodopsis is sister to Nannochloropsis + Microchloropsis in the family Monodopsidaceae (Eustigmatales), and Vischeria is a member of the sister family Eustigmataceae, also in the order Eustigmatales (fig. 1A; supplementary fig. S1, Supplementary Material online). We carried out comparative studies of repeats and gene space and found evidence of spliced leader trans-splicing (SLTS) in these eustigmatophytes. Our results here help to gain a more holistic view on the biology and genomic diversity of eustigmatophytes within the Eustigmatales, expanding beyond what was only known from Nannochloropsis and Microchloropsis.

Results and Discussion

Eustigmatophytes Isolated from Bryophytes

In our ongoing effort to isolate symbiotic cyanobacteria from surface-sterilized bryophyte thalli (Nelson et al. 2019), we have occasionally obtained eustigmatophyte algae instead. DNA barcoding using the 18S rDNA marker indicates all our eustigmatophyte isolates belong to either Monodopsis or Vischeria (see supplementary fig. S1, Supplementary Material online for the 18S rDNA phylogeny). So far, we have isolates from multiple species of hornworts, liverworts, and mosses, and from diverse geographic locations spread across North America (supplementary table S1, Supplementary Material online). The nature of interaction between eustigmatophytes and bryophytes (if there is any) is unclear. A symbiotic relationship is a possibility, given that similar algal strains have been repeatedly isolated from bryophytes from different locations (supplementary table S1, Supplementary Material online). The recent finding that Nannochloropsis oceanica could enter an endosymbiotic relationship with the fungus Mortierella (Du et al. 2019) further speaks to the symbiotic competency of eustigmatophytes. On the other hand, both Monodopsis and Vischeria are common soil algae, and it is possible that they are resistant to our sterilization method and came out as “contaminants.” Future experiments are needed to examine the possible eustigmatophyte–bryophyte interaction.

Near-Chromosomal Level Assemblies of Monodopsis and Vischeria

To obtain high quality reference genomes, we generated Illumina short reads and Oxford Nanopore long reads for one Vischeria (C74) and two Monodopsis strains (C73, C141). The K-mer-based genome size estimates were around 60 and 100 Mb for Monodopsis and Vischeria, respectively. After filtering, the Nanopore data represented 45–67x coverage with a read length N50 between 13 and 25 kb (supplementary table S2, Supplementary Material online). The assemblies based on Flye (Kolmogorov et al. 2019) are near chromosomal, with the majority of the contigs containing at least one telomeric end (table 1). The telomeric motif is “TTAGGG,” which was also found in Microchloropsis (=Nannochloropsis) gadinana B-31 (Cortegiani Carpinelli et al. 2014). A total of 13,969, 13,933, and 18,346 protein-coding genes were annotated from Monodopsis C73, Monodopsis C141, and Vischeria C74, respectively, all with a 100% Benchmarking Universal Single-Copy Orthologs (BUSCO) completeness score against the “Stamenpilze” data set. Compared with the published Nannochloropsis and Microchloropsis genomes, the assemblies we present here are by far the most complete (fig. 1A). Interestingly, none of the three genomes contain Ca. Phycorickettsia contigs that were previously reported in other eustigmatophytes (Yurchenko et al. 2018).

To gain a better picture of the genetic diversity, we generated Illumina data for two additional strains: Monodopsis C143 and Vischeria C101. SNP densities between the Monodopsis strains (C73, C141, and C143) ranged from 34 to 44/kb, and 10/kb between the Vischeria strains (C74 and C101) (supplementary table S3, Supplementary Material online). After filtering, the Nanopore data represented 45–67x coverage with a read length N50 between 13 and 25 kb (supplementary table S2, Supplementary Material online). The nature of interaction between eustigmatophytes and bryophytes (if there is any) is unclear. A symbiotic relationship is a possibility, given that similar algal strains have been repeatedly isolated from bryophytes from different locations (supplementary table S1, Supplementary Material online). The recent finding that Nannochloropsis oceanica could enter an endosymbiotic relationship with the fungus Mortierella (Du et al. 2019) further speaks to the symbiotic competency of eustigmatophytes. On the other hand, both Monodopsis and Vischeria are common soil algae, and it is possible that they are resistant to our sterilization method and came out as “contaminants.” Future experiments are needed to examine the possible eustigmatophyte–bryophyte interaction.

A New Annotation of Microchloropsis gaditana Genome

Although three Microchloropsis gaditana genome assemblies have been published to date, two of them (B-31 and CCMP526) were based on short-read technologies and therefore had low contig N50 length (40.5 kb for B-31 and 15.3 kb for CCMP526) as well as low BUSCO completeness scores (fig. 1A) (Radakovits et al. 2012; Cortegiani Carpinelli et al. 2014). Only the M. gaditana CCMP1894 genome was assembled using long reads (Schwartz et al. 2018), but unfortunately its annotation has not been published. Here we used publicly available RNA-seq data and protein evidence to

Downloaded from https://academic.oup.com/genome/article/13/11/evab233/6402010 by University of Aberdeen user on 11 November 2021
annotate the *M. gaditana* CCMP1894 assembly. This new annotation has a much-improved BUSCO score (94% complete) compared with the previous *M. gaditana* annotations (40% and 85%) (fig. 1A).

### Unusually High Percentages of Simple Sequence Repeats

*Monodopsis* and *Vischeria* have considerably larger genomes than those of *Nannochloropsis/Microchloropsis*, which can be partly attributed to their higher percentages of repetitive elements (fig. 1B). The simple sequence repeats (SSRs) and LINEs are particularly noteworthy. Although LINEs are absent in *Nannochloropsis/Microchloropsis*, they cover around 2.9–3.6% of the *Monodopsis* and *Vischeria* genomes (fig. 1B). SSRs have similarly expanded representations, accounting for 11.7–12.2% of the genomic content in *Monodopsis* and 20.8% in *Vischeria* (fig. 1B). Although these SSRs can be found throughout the chromosomes, they are particularly enriched toward the chromosome ends (figs. 2 and 3). The frequencies of SSRs observed here are in fact among the highest of all genomes sequenced to date. For example, the human body louse genome (*Pediculus humanus corporis*) had the highest SSR density according to Srivastava et al. (2019). When reanalyzed with the same repeat annotation pipeline used here, we found SSRs account for 16.9% of the *P. humanus corporis* genome, making *Vischeria* C74 (at 20.8%) the most SSR-dense genome known to date. Future comparative studies incorporating additional genomes across eustigmatophytes are needed to clarify the impact of such high abundance of SSRs on genome structure and evolution.

### Putative Centromeric Regions That Are Enriched in LINEs

Only a few centromere structures have been experimentally characterized in *Stramenopiles*. In the oomycete *Phytophthora sojae*, the centromeric regions are particularly rich in the Copia-like retroelements (Fang et al. 2020), whereas in the diatom *Phaeodactylum tricornutum*, the centromeres are AT-rich but devoid of repetitive elements (Diner et al. 2017). No putative centromeric region has been identified in *Nannochloropsis/Microchloropsis* to date nor in any other eustigmatophyte. Our analysis of *Monodopsis* and *Vischeria* genomes suggest that their centromeres might be characterized by islands of LINE clusters. The distributions of LINEs in *Monodopsis* and *Vischeria* are highly heterogeneous, usually with a sharp peak toward the middle of a chromosome (figs. 2 and 3). It is likely that such LINE-dense (and gene-poor) regions function as centromeres, but further immunolabeling studies are needed. If confirmed, it would also suggest that *Nannochloropsis/Microchloropsis* might have a substantially different centromere organization given their absence of LINE.

### A Haploid-Dominant Life Cycle

The complete life cycle of eustigmatophytes has not been characterized, and no sexual reproduction has been observed. We found that several meiosis-specific genes are present in *Monodopsis* and *Vischeria*, which is consistent with what was found in *Microchloropsis* (supplementary table S4, Supplementary Material online) (Radakovits et al. 2012; Corteggiani Carpinelli et al. 2014) and suggests eustigmatophytes do have cryptic sexual stages. In addition, we were able to identify homologs encoding flagella-related proteins in both *Monodopsis* assemblies (examples provided in supplementary table S4, Supplementary Material online), despite zoospores never having been documented in *Monodopsis* (but known in *Vischeria*) (Hibberd 1981; Eliás et al. 2017). Another missing piece of information about the life cycle of eustigmatophytes is the dominant ploidy level. Although earlier genomic studies on *Nannochloropsis* suggested they are monoploid (Pan et al. 2011), no information is available for other members of eustigmatophytes. In order to assess if there is any heterozygosity present in our *Monodopsis* and *Vischeria* strains, we mapped Illuma reads to the respective genomes. We found very few SNPs could be called, and the vast majority of the alternative alleles were supported by low percentages of reads (supplementary fig. S2, Supplementary Material online), suggesting these SNPs were artifacts of residual sequencing and/or assembly errors. Therefore, we infer
both Monodopsis and Vischeria have a haploid-dominant life cycle similar to Nannochloropsis/Microchloropsis.

Terpenoid Biosynthesis Pathways Differ between Monodopsis/Vischeria and Nannochloropsis

Terpenoids are an important class of natural products and have high bioenergy potentials. There are two pathways for terpenoid biosynthesis: the mevalonate pathway (MVA) and the nonmevalonate pathway (MEP). Many Stramenopiles, such as diatoms, have both pathways, whereas all the Nannochloropsis/Microchloropsis genomes sequenced to date have only the MEP pathway. Interestingly, in the Monodopsis and Vischeria genomes, we were able to find intact MVA and MEP pathways present (supplementary fig. S3, Supplementary Material online). The top BLASTP hits of
these MVA pathway genes are from other stramenopile species, indicating vertical inheritance of the genes from a stramenopile ancestor instead of horizontal gene transfer into the eustigmatophyte lineage. Because *Nannochloropsis/Microchloropsis* is nested within *Monodopsis* + *Vischeria*, the most likely scenario is that *Nannochloropsis/Microchloropsis* secondarily lost the MVA pathway. This finding highlights the importance of having biodiverse genomes to infer the biology of eustigmatophytes.

**Presence of SLTS and Operons**

Our initial analysis of the RNA-seq data revealed a low read mapping rate (~85%), which is surprising given the high genome completeness and continuity. One possible explanation is the presence of SLTS, which was reported in *M. gaditana* in a patent application (Seshadri et al. 2018). SLTS is a special mRNA maturation process, in which the 5’ end of a premRNA is capped by a spliced leader (SL) sequence that is
transcribed from a separate SL locus. The main function of SLTs is to add the necessary 5' cap to each cistron in a eu-karyotic operon (Lasda and Blumenthal 2011). A diverse group of organisms have been shown to have SLTs, including nematodes, cnidarians, and several unrelated protist lineages (Bitar et al. 2013; Krchnáková et al. 2017).

Upon closer inspection with SL detection pipelines, we found evidence of a single SL type in Monodopsis and Vischeria, and also confirmed the SL previously reported in M. gaditana (table 2). The main variants of these SLs were supported by at least 155,671 reads, ensuring confidence in their accuracy (supplementary table S5, Supplementary Material online). All species also possess several minor SL sequence variants at much lower read coverage (supplementary table S5, Supplementary Material online). This may suggest a complex genome-wide landscape of alternative SLs in all species, similar to kinetoplastids (Nilsson et al. 2010). The main SL variants were encoded by 24–239 candidate SL RNA genes. Except for Monodopsis C141, all species possess at least two dissimilar SL RNA gene variants, which may indicate the presence of pseudogenes (supplementary table S5, Supplementary Material online). Functional SL RNA copies are expected to possess a T-rich region (Sm binding motif) that is required for interaction with the splicing machinery (Stover et al. 2006). We found the canonical Sm binding motif ATTTTG (Bitar et al. 2013) in six out of 170 SL RNA genes in Vischeria, but not in Monodopsis and Microchloropsis (supplementary table S5, Supplementary Material online). This may indicate that the more recently diverged species Monodopsis and Microchloropsis have an altered SLTS machinery with different Sm motifs, which will require functional molecular studies to elucidate. The secondary structures of the SL RNA genes of all species display at least one major stem loop (supplementary table S5, Supplementary Material online), consistent with SL RNAs in dinoflagellates (Zhang et al. 2007) and tunicates (Ganot et al. 2004), but divergent from the typical three-loop structure in most other organism groups (Krchnáková et al. 2017).

Having established the presence of SLTS in all species, we then tested whether the physical locations of genes that receive SLs may imply the presence of operons. We first reconstructed the 5' UTRs of gene annotations aided by the identified SLs, which yielded improved annotations for 40–80% of genes (supplementary table S6, Supplementary Material online). Using these improved annotations, we then detected SLs at 36% of genes in Vischeria, 58–61% in Monodopsis, and 89% in Microchloropsis. Requiring downstream genes in operons to receive the SL and intercistronic distances to be no greater than 1,000 bp predicted 682–1,253 operons per species, containing 8–30% of all genes (table 3). Only 21–44 of these operons had intercistronic distances of at most 100 bp (supplementary table S6, Supplementary Material online). Consistent with the much higher SLTs rate, 90% of the putative Microchloropsis operons receive the SL at both upstream and downstream genes, whereas Vischeria and Monodopsis show upstream SLTs at only 44–64% of the putative operons. We found no significant (FDR ≤ 0.1) GO or KEGG enrichment in operonic genes compared with the full genomic background, contrary to expectations from other organisms (e.g., Zeller 2010). This may suggest that operon evolution in these species was not necessarily driven by functional coordination of gene expression.

Although these predictions are likely not exhaustive and will require functional validation, they are entirely consistent with other organisms where a single SL is added to both monocistronic and operonic genes, for example, tunicates (Ganot et al. 2004) and platyhelminths (Boroni et al. 2018). Although SLTS has been reported in some algal lineages (Kuo et al. 2017; Roy 2017), our results provide the first insight into the genome-wide landscape of SLTS and putative operons in Systems
several eustigmatophyte algae in the order Eustigmatales. Future long-read RNA or cDNA sequencing will help to better define these operons and clarify the functional significance.

**Conclusion**

Here we present three high-quality genome assemblies of *Monodopsis* and *Vischeria*. We found that in many aspects, *Monodopsis* and *Vischeria* genomes are substantially different from those of *Nannochloropsis/Microchloropsis*. For instance, *Monodopsis* and *Vischeria* genomes are two to three times larger, and boast one of the highest proportions of simple repeats among sequenced eukaryotic genomes. The centromeric regions in *Monodopsis* and *Vischeria* might be made up by LINE repeats, which are notably absent in *Nannochloropsis/Microchloropsis*. In addition, although *Nannochloropsis/Microchloropsis* lacks the MVA pathway for terpenoid biosynthesis, both MVA and MEP are present in *Monodopsis* and *Vischeria* and likely represent the ancestral state.

We also identified important features that are shared among these eustigmatophyte genomes in the order Eustigmatales. Notably, our finding and the initial characterizations of SLTS unraveled a new aspect of eustigmatophyte biology. We anticipate our new genomic data and associated analyses will greatly facilitate future research to better understand the biology of eustigmatophytes, and to better capitalize on their translational potential.

**Materials and Methods**

**Strain Isolation**

The three *Monodopsis* (C73, C141, and C143) and two *Vischeria* (C74 and C101) strains sequenced here were isolated from surface-sterilized bryophytes. The localities can be found in supplementary table S1, Supplementary Material online. We followed the methods outlined in Nelson et al. (2019) for cleaning and sterilizing the bryophyte thalli, as well as for establishing unialgal cultures that grew out from the plants. These new algal cultures are available through UTEX Culture Collection of Algae (accession numbers UTEX 3167–3171).

**Genome Sequencing**

We sequenced the genomic DNA on both Oxford Nanopore MiniON device as well as Illumina NextSeq500 platform. Nanopore libraries were prepared using the Ligation Sequencing kit (SQK-LSK109), and sequenced on MiniON R9 flowcells (FL0-MIN106D) for 60 h or until the flowcells died. We carried out basecalling using Guppy v3.0.3 (https://nanoporetech.com/, last accessed July 2021) with the high accuracy flip–flop mode. For *Monodopsis* C73 and C141 strains, reads shorter than 15 kb were discarded prior to assembly, and for *Vischeria* C74, a threshold of 5 kb was used. For Illumina libraries, we followed the general protocol of Nelson et al. (2019) using the SparQ DNA Frag & Library Prep kit and Adapter Barcode Set A. The libraries were pooled with nine other samples and sequenced on one Illumina NexSeq500 mid-output flowcell (150 bp paired-end) at Cornell Institute of Biotechnology. Reads were trimmed and quality-filtered by fastp v0.20.1 (Chen et al. 2018).

**RNA Sequencing**

Cells grown on BG11 solution under 12/12 dark/light cycle and 22 °C were harvested by centrifugation and disrupted by an SPEX SamplePrep 1600 MiniG tissue homogenizer. RNA was extracted using Sigma Spectrum Plant Total RNA kit, and strand-specific RNA-seq libraries were made by YourSeq Duet RNAseq Library Kits from Amaryllis Nucleics. The RNA libraries were pooled with 16 other samples and sequenced on one Illumina NovaSeq6000 S-Prime flowcell (150 bp paired-end). Reads were trimmed and quality-filtered by Trimmomatic v0.39 (Bolger et al. 2014).

**Genome Assembly**

We first estimated the genome size based on the K-mer frequency of Illumina reads using MaSuRCA v3.3.2 (Zimin et al. 2013, 2017). To assemble the Nanopore reads, we used Flye v2.4.1 (Kolmogorov et al. 2019) with four iterations of built-in polishing, followed by one round of medaka v0.7.1 (https://github.com/nanoporetech/medaka) processing. The

---

**Table 3**

Summary of Operons Predicted in *Monodopsis*, *Vischeria*, and *Microchloropsis* on the basis of SLTS

| Genome             | % Genes SL Trans-Spliced | Predicted Operons | % Predicted Operons with SLTS Upstream Genes | Predicted Operonic Genes | % Total Genes | Median Intercistronic Distance (bp) |
|--------------------|--------------------------|-------------------|---------------------------------------------|--------------------------|---------------|-------------------------------------|
| *Vischeria* C74    | 36%                      | 682               | 44%                                        | 1,408                    | 8%            | 564                                 |
| *Monodopsis* C73   | 61%                      | 1,164             | 64%                                        | 2,442                    | 17%           | 542                                 |
| *Monodopsis* C141  | 58%                      | 1,068             | 60%                                        | 2,216                    | 16%           | 554                                 |
| *Microchloropsis* gaditana CCMP1894 | 89% | 1,253 | 90% | 2,765 | 30% | 655 |

Note.—Predictions required intercistronic distances of at most 1,000 bp and did not require SLTS at upstream operonic genes. The table presents the percentage of genes receiving SL reads, the numbers of operons, the percentage of operons where the upstream operonic gene receives SL reads, the numbers and percentages of operonic genes, and the median intercistronic distances among operonic genes. Details for SL read quantification and operon prediction using alternative criteria are provided in supplementary table S6, Supplementary Material online.
nanopore assemblies were further error-corrected by Illumina reads using pilon v1.23 (Walker et al. 2014) with four iterations. To better assemble the telomeric regions, we used teloclip v0.0.3 (https://github.com/Adamtaranto/teloclip) to recover telomeric nanopore reads that can be aligned and appended to the contig ends. Organellar genomes were assembled separately using either GetOrganelle v1.7 (Jin et al. 2020) with Illumina reads, or Flye with a subset of nanopore reads that mapped to organellar genomes of closely related species. The Flye organellar assemblies were polished by pilon until no correction can be made. Finally, the organellar genomes were BLASTN to the nuclear genome assembly to identify and remove any redundant organellar contigs.

Repeat Annotation

Our initial repeat analysis revealed a large percentage of simple microsatellite repeats, which caused RepeatMasker (Smit et al. 2015) to make many spurious matches to other repeat classes. To address this, we first identified and masked the simple repeats from the genome using RepeatMasker, before building the custom repeat database with RepeatModeler2 (Flynn et al. 2020). RepeatMasker was then used again to annotate and mask all the repeat classes from the genomes. Tandem repeats were identified separately using Tandem Repeats Finder (Benson 1999).

Gene Model Prediction

Gene predictions were done by BRAKER2 v2.1.5 (Būna et al. 2021), integrating both protein and transcript evidence with --etpmode and --softmasking flags on. To provide transcript evidence, we mapped RNA-seq reads to the corresponding genome using HiSAT2 v2.1.0 (Kim et al. 2015). To compile the protein evidence, we first used MAKER2 (Holt and Yandell 2021) to train SNAP (Korf 2004) on Monodopsis C73 based on reference-guided transcriptome assembly from Trinity v2.1.1 (Grabherr et al. 2011) and Nannochloropsis/ Microchloropsis protein records from GenBank. The resulting gene models were then annotated with eggNOG v5.0 (Huerta-Cepas et al. 2019), and only genes with annotations were kept as the protein evidence for BRAKER gene prediction. We used the same approach to annotate M. gaditana CCMP1894 genome, with transcript evidence from three publicly available RNA-seq data sets (SRA accession numbers: SRR5152511, SRR5152512, and SRR5152516) and protein sequences from M. gaditana B31 and M. salina CCMP1776. To filter out spurious gene models from BRAKER2, we removed genes that failed to meet all of the following criteria: 1) a TPM expression level at least 0.001, 2) has functional annotation from eggNOG, and 3) was assigned into orthogroups when including all the focal eustigmatophyte genomes in an OrthoFinder v2.3.12 (Emms and Kelly 2019) run. We used BUSCO v4.0.6 (Simão et al. 2015) to assess the completeness of genome assemblies and annotations with the “Stramenopiles” lineage data set. The final gene sets were functionally annotated (including GO and KEGG) by eggNOG v5.0. KEGG pathways were reconstructed using the KEGG Mapper tool (Kanehisa and Sato 2020).

Visualization of Genome Structures

We used circos (Krzywinski et al. 2009) to visualize the distributions of genes, repeats, and GC content along the genome assemblies. All the sliding windows had a window size of 50 kb and a step size of 25 kb. Gene and repeat density values were calculated using BEDTools 2.28.0 (Quinlan and Hall 2010). GC content deviations were calculated based on whole genome average, which is 0.4615, 0.4620, and 0.5313 for Monodopsis C73, Monodopsis C141, and Vischeria C74, respectively.

SNP Calling

For each genome, we used bwa v0.7.17 (Li and Durbin 2009) to map Illumina reads to self as well as to the related genomes. We then use bcftools v1.9 (Li 2011) to call SNPs and keep those with quality over 50 and read depth over 20.

Phylogenetic Relationship of Currently Available Eustigmatophyte Genomes

We compiled a list of the eustigmatophyte genomes that have annotations available (fig. 1), and used Orthofinder v2.3.12 to infer gene orthology. A total of 1,302 single-copy loci were identified, and protein sequence alignments were done by MAFFT (Katoh and Standley 2013). We then carried out phylogenetic reconstruction using IQ-TREE v2.0.3 (Nguyen et al. 2015) on the concatenated alignment matrix with automatic model selection (Kalyaanamoorthy et al. 2017) and 1,000 replicates of ultrafast bootstrapping (Hoang et al. 2018).

Identification of SLTS

We identified SLs in the C73, C74, and C141 strains as well as M. gaditana CCMP1894 (RNA-Seq library SRR10431616 from SRA) using SLDR 1.1.4 with distance-based clustering (Wenzel et al. 2021). We relaxed the SL length limit (−x 1.25), required GT/AG splice sites and disabled the Sm binding motif filter. Identified SL RNA genes were inspected and aligned using MAFFT v7.407. Secondary sequence structures were inferred using RNAfold Web Server (Gruber et al. 2008). Identified SL trans-splice acceptor sites were compared against gene annotations using BEDTools 2.28.0 (Quinlan and Hall 2010).

We then tested whether genome-wide SL trans-splicing events may indicate the presence of operonic gene organization using SLOPPR 1.1.3 (Wenzel et al. 2021). Because SLOPPR requires accurate gene annotations, particularly at the 5′ end, we first predicted 5′ UTRs guided by identified SLs using UTRme (Radio et al. 2018), relaxing maximum UTR
length to 10,000 bp and maximum UTR ORF length to 400 amino acids. Reads containing at least 8 bp of the SL at the 5’ end were then identified and quantified against transcript annotations using SLOPPR. Operon inference was tested with four intercistronic distance cutoffs (infinity, 1,000 bp, 100 bp, and automatic inference) and did not require up-steam operonic genes to be SL trans-spliced. The functional annotations (GO, KEGG) of candidate operonic genes were tested for overrepresentation against the genome-wide background using hypergeometric tests in ClusterProfiler 3.14.2 (Yu et al. 2012).

Supplementary Material

Supplementary data are available at Genome Biology and Evolution online.

Acknowledgments

This study was supported by the National Science Foundation Dimensions of Biodiversity (Grant No. DEB1831428 to F.W.L) and the Czech Science Foundation (Grant No. 20-27648S to M.E.). We thank the reviewers and editor for their thoughtful comments.

Data Availability

The nanopore and Illumina sequencing reads were deposited in NCBI SRA under the BioProject PRJNA730568. The nuclear genome assemblies and annotations are available through NanDeSyn data portal (Gong et al. 2020) and https://figshare.com/s/c4bf156c2764ba410c30.

Literature Cited

Ajajwi I, et al. 2017. Lipid production in Nannochloropsis gaditana is doubled by decreasing expression of a single transcriptional regulator. Nat Biotechnol. 35(7):647–652.

Amaral R, et al. 2020. Toward modern classification of eustigmatophytes, including the description of Neomonodaceae fam. nov. and three new genera. J Phycol. 56(3):630–648.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27(2):573–580.

Bitar M, Boroni M, Macedo AM, Machado CR, Franco GR. 2013. The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles. Front Genet. 4:199.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27(2):573–580.

Bitar M, Boroni M, Macedo AM, Machado CR, Franco GR. 2013. The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles. Front Genet. 4:199.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27(2):573–580.

Bitar M, Boroni M, Macedo AM, Machado CR, Franco GR. 2013. The spliced leader trans-splicing mechanism in different organisms: molecular details and possible biological roles. Front Genet. 4:199.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120.
Huang L, Gao B, Wang F, Zhao W, Zhang C. 2019. The complete chloroplast genome of an edaphic oleaginous microalga Vischeria stellata SAG 33.83 (Eustigmatophyceae). Mitochondrial DNA Part B 4(1):1041–1043.

Huerta-Cepas J, et al. 2019. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 47(1):D309-D314.

Jagadevan S, et al. 2018. Recent developments in synthetic biology and metabolic engineering in microalgae towards biofuel production. Biotechnol Biofuels 11:185.

Jin J-J, et al. 2020. GetOrganelle: a fast and versatile toolkit for accurate de novo assembly of organelle genomes. Genome Biol. 21(1):241.

Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. Nat Methods 14(6):587–589.

Kanehsa M, Sato Y. 2020. KEGG Mapper for inferring cellular functions from protein sequences. Protein Sci. 29(1):28–35.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30(4):772–780.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12(4):357–360.

Kolmogorov M, Yuan J, Lin Y, Pevzner PA. 2019. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 37(5):545–546.

Korf I. 2004. Gene finding in novel genomes. BMC Bioinformatics 5:59.

Krčičáková Z, Krajčová J, Vesteg M. 2017. On the possibility of an early evolutionary origin for the spliced leader trans-splicing. J Mol Evol. 85(1–2):37–45.

Krzywinski M, et al. 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19(10):1639–1645.

Kuo RC, Zhang H, Zhuang Y, Hannick L, Lin S. 2013. Transcriptional study reveals widespread spliced leader trans-splicing, short 5′-UTRs and potential complex carbon fixation mechanisms in the euglenoid Alga Eutreptiella sp. PLoS One 8(4):e60826.

Lasda EL, Blumenthal T. 2011. Trans-splicing. Wiley Interdiscip Rev RNA 2(3):417–434.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetic estimation from sequencing data. Bioinformatics 27(12):2987–2993.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25(14):1754–1760.

Nadathurk M, et al. 2019. CRISPR–Cas ribonuclease protein mediated homology-directed repair for efficient targeted genome editing in microalgae Nannochloropsis oceanica IMET1. Biotechnol Biofuels 12:66.

Nelson JM, et al. 2019. Complete genomes of symbiotic cyanobacteria clarify the evolution of Vanadium-nitrogenase. Genome Biol Evol. 11(7):1959–1964.

Nelson DR, et al. 2021. Large-scale genome sequencing reveals the driving forces of viruses in microalgal evolution. Cell Host Microbe 29(2):250–266.e8.

Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 32(1):268–274.

Nilsson D, et al. 2010. Spliced leader trapping reveals widespread alternative splicing patterns in the highly dynamic transcriptome of Trypanosoma brucei. PLoS Pathog. 6(1):e1000937.

Ohan JA, et al. 2019. Nuclear genome assembly of the microalga Nannochloropsis salina CCMP1776. Microb Resour Announc. 8(44):e00705–19.

Oorio H, Jara C, Fuenzalida K, Rey-Jurado E, Vásquez M. 2019. High-efficiency nuclear transformation of the microalgae Nannochloropsis oceanica using Tn5 Transposome for the generation of altered lipid accumulation phenotypes. Biotechnol Biofuels. 12:134.

Pan K, et al. 2011. Nuclear monoploidy and asexual propagation of Nannochloropsis oceanica (Eustigmatophyceae) as revealed by its genome sequence. J Phycol. 47(6):1425–1432.

Poliner E, Clark E, Cummings C, Benning C, Farre EM. 2020. A high-capacity gene stacking toolkit for the oleaginous microalga, Nannochloropsis oceanica CCM1779. Algal Res. 45:101664.

Poliner E, Takeuchi T, Du Z-Y, Benning C, Farre EM. 2018. Nontransgenic marker-free gene disruption by an episomal CRISPR system in the oleaginous microalga, Nannochloropsis oceanica CCM1779. ACS Synth Biol. 7(4):962–968.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26(6):841–842.

Radakovits R, et al. 2012. Draft genome sequence and genetic transformation of the oleaginous alga Nannochloropsis gaditana. Nat Commun. 3:686.

Radakovits R, Jinkerson RE, Darzins A, Posewitz MC. 2010. Genetic engineering of algae for enhanced biofuel production. Eukaryot Cell 9(4):486–501.

Radio S, Fort RS, Garat B, Sotelo-Silveira J, Smirich P. 2018. UTRm: a scoring-based tool to annotate untranslated regions in trypanosomatid genomes. Front Genet. 9:671.

Roy SW. 2017. Genomic and transcriptomic analysis reveals spliced leader trans-splicing in cryptomonads. Genome Biol Evol. 9(3):468–473.

Schwartz AS, et al. 2018. Complete genome sequence of the model oleaginous alga Nannochloropsis gaditana CCMP1984. Genome Announc. 6(7):e01448-17.

Sehadi R, Schwartz AS, Soraga L, Brown RC. 2018. Nannochloropsis spliced leader sequenced and uses therefor. US Patent.

Sévigné T, et al. 2016. A comparative analysis of mitochondrial genomes in eustigmatophyte algae. Genome Biol Evol. 8(3):705–722.

Sévigné T, et al. 2019. Plastid genomes and proteins illuminate the evolution of eustigmatophyte algae and their bacterial endosymbionts. Genome Biol. Evol. 11(2):362–379.

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. 2015. BUSCO: assessing genome annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210–3212.

Smit A, Hubley R, Green P. 2015. RepeatMasker Open-4.0. Available from: http://www.repeatmasker.org.

Srivastava S, Avarau AK, Sowpati DT, Mishra RK. 2019. Patterns of microsatellite distribution across eukaryotic genomes. BMC Genomics 20(1):153.

Stover KA, Haye MS, Cavalanti ARO. 2006. Spliced leader trans-splicing. Curr Biol. 16(1):R8–R9.

Verruto J, et al. 2018. Unrestrained markerless trait stacking in Nannochloropsis gaditana through combined genome editing and marker recycling technologies. Proc Natl Acad Sci U S A. 115(30):E7015–E7022.

Vieler A, et al. 2012. Genome, functional gene annotation, and nuclear transformation of the heterokont oleaginous alga Nannochloropsis oceanica CCM1779. PLoS Genet. 8:e1003064.

Walker BJ, et al. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 9(11):e112963.

Wang D, et al. 2014. Marker-free gene disruption by an episomal CRISPR system in the oleaginous microalga, Nannochloropsis gaditana, as revealed by its genome sequence. J Phycol. 50(4):1041–1042.

Wenzel MA, Müller B, Pettitt J. 2021. SLDR and SLOPPR: flexible identification of spliced leader trans-splicing and prediction of eukaryotic operons from RNA-Seq data. BMC Bioinformatics 22(1):140.

Genome Biol. Evol. 13(11) https://doi.org/10.1093/gbe/evab233 Advance Access publication 19 October 2021
Yu G, Wang L-G, Han Y, He Q-Y. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS 16(5):284–287.

Yurchenko T, et al. 2018. A gene transfer event suggests a long-term partnership between eustigmatophyte algae and a novel lineage of endosymbiotic bacteria. ISME J. 12(9):2163–2113.

Yurchenko T, Ševčiková T, Strnad H, Butenko A, Eliáš M. 2016. The plastid genome of some eustigmatophyte algae harbours a bacteria-derived six-gene cluster for biosynthesis of a novel secondary metabolite. Open Biol. 6(11):160249–160222.

Zeller RW. 2010. Computational analysis of Ciona intestinalis operons. Integr Comp Biol. 50(1):75–85.

Zhang H, et al. 2007. Spliced leader RNA trans-splicing in dinoflagellates. Proc Natl Acad Sci U S A. 104(11):4618–4623.

Zimin AV, et al. 2017. Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. Genome Res. 27(5):787–792.

Zimin AV, et al. 2013. The MaSuRCA genome assembler. Bioinformatics 29(21):2669–2677.

Associate editor: John Archibald