PLOS ONE

RESEARCH ARTICLE

Reclassification of *Botryococcus braunii* chemical races into separate species based on a comparative genomics analysis

Devon J. Boland[1,2], **Ivette Cornejo-Corona**[1], **Daniel R. Browne**[1,3], **Rebecca L. Murphy**[1,4], **John Mullet**[1], **Shigeru Okada**[5], **Timothy P. Devarenne**[1]*

1 Department of Biochemistry and Biophysics, Texas A & M University, College Station, Texas, United States of America, **2** Texas A&M Institute for Genome Sciences & Society (TIGSS), College Station, Texas, United States of America, **3** AI & Computational Biology, LanzaTech Inc., Skokie, Illinois, United States of America, **4** Biology Department, Centenary College of Louisiana, Shreveport, Louisiana, United States of America, **5** Laboratory of Aquatic Natural Products Chemistry, Graduate School of Agricultural and Life Sciences, The University of Tokyo, Yayoi, Bunkyo, Tokyo, Japan

* tpd8@tamu.edu

Citation: Boland DJ, Cornejo-Corona I, Browne DR, Murphy RL, Mullet J, Okada S, et al. (2024) Reclassification of *Botryococcus braunii* chemical races into separate species based on a comparative genomics analysis. PLoS ONE 19(7): e0304144. https://doi.org/10.1371/journal.pone.0304144

Editor: Rishiram Ramanan, Central University of Kerala, INDIA

Received: September 12, 2023

Accepted: May 7, 2024

Published: July 29, 2024

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files. Data availability: B. alkenealis and B. lycopadienor sequence data related to this study including the final assemblies can be found under NCBI BioProject: PRJNA1079018.

Funding: This work was supported by funds from the Texas A&M Department of Biochemistry & Biophysics to T.P.D., grant NSF-EFRI-PSBR 1240478 to T.P.D., by contract DE-EE0003046

Abstract

The colonial green microalga *Botryococcus braunii* is well known for producing liquid hydrocarbons that can be utilized as biofuel feedstocks. *B. braunii* is taxonomically classified as a single species made up of three chemical races, A, B, and L, that are mainly distinguished by the hydrocarbons produced. We previously reported a B race draft nuclear genome, and here we report the draft nuclear genomes for the A and L races. A comparative genomic study of the three *B. braunii* races and 14 other algal species within *Chlorophyta* revealed significant differences in the genomes of each race of *B. braunii*. Phylogenomically, there was a clear divergence of the three races with the A race diverging earlier than both the B and L races, and the B and L races diverging from a later common ancestor not shared by the A race. DNA repeat content analysis suggested the B race had more repeat content than the A or L races. Orthogroup analysis revealed the *B. braunii* races displayed more gene orthogroup diversity than three closely related *Chlamydomonas* species, with nearly 24-36% of all genes in each *B. braunii* race being specific to each race. This analysis suggests the three races are distinct species based on sufficient differences in their respective genomes. We propose reclassification of the three chemical races to the following species names: *Botryococcus alkenealis* (A race), *Botryococcus braunii* (B race), and *Botryococcus lycopadienor* (L race).

Introduction

Botryococcus braunii (Trebouxiophyceae) is a colonial green microalga that has long been studied due to its ability to biosynthesize large quantities of liquid hydrocarbons with potential use as biofuel feedstocks [1, 2]. Purified hydrocarbons from *B. braunii* have been refined into petroleum-fuel equivalents to provide so-called "drop-in" fuels that can be used with the

awarded to the National Alliance for Advanced Biofuels and Bioproducts (NAABB) from the U.S. Department of Energy, of which T.P.D. was a member, and JSPS KAKENHI (Grant-in-Aid for Scientific Research B) Grant Number JP22H02432 to S.O.. D.J.B. was supported by funding from the G. Rollie White Trust, Texas A&M University College of Agriculture & Life Sciences Endowed Graduate Fellowship for the Borlaug Scholars Program, and R.L.M was supported by funding form the Texas A&M University Perry L. Adkisson Chair in Agricultural Biology to J.E.M. There was no additional external funding received for this study.

Competing interests: The authors have declared that no competing interests exist.

current combustive fuel infrastructure [3–5]. *B. braunii* is further split into three distinct chemical races based on the hydrocarbons produced in each race: The odd carbon numbered C_{23-33} alkadiene/alkatriene producing A race, the C_{30-37} botryococcene and C_{31-34} methylsqualene producing B race, and the C_{40} lycopadiene producing L race [6–8]. More recently, a potential fourth S chemical race was isolated with the major hydrocarbons identified as shorter chain saturated C_{18} epoxy-*n*-alkanes or a C_{20} *n*-alkane [9]. This study focuses on the A, B, and L races of *B. braunii*.

Historically, studies on *B. braunii* have focused on hydrocarbon biosynthesis, cell morphology, extracellular matrix (ECM) composition, and phylogenetics [1, 10–21]. For the hydrocarbons, the A race alkenes are derived from fatty acids utilizing C_{18} oleic acid as a precursor [10, 11, 13], while the B and L races utilize the isoprenoid-derived molecules C_{15} farnesyl diphosphate and C_{20} geranylgeranyl diphosphate, respectively, as precursors for hydrocarbon production [22]. Morphologically, the cell sizes differ between the races with the L race displaying smaller sizes (8-9μm x 5μm) than the A or B races (13μm x 7-9 μm) [12]. As a consequence of the different hydrocarbons, each race has a distinct ECM composition with the hydrocarbons in each race linked by epoxide bridges to long-chain polyacetals spanning the ECM. Phylogeny studies focused on the 18S rRNA gene revealed distinct clade-level groupings for each of the three races [9, 16, 17, 20, 23]. Despite these many differences, *B. braunii* is taxonomically classified as a single species made up of distinct races consisting of various geographic strains within each race.

One traditional definition for species classification of organisms utilizes the biological species concept, in which organisms are defined as separate species if they cannot cross via sexual reproduction [24]. Chlorophyta contains species that reproduce sexually or asexually [25, 26]. For example, *Chlamydomonas reinhardtii* and its close relatives reproduce sexually through the regulation of two key genes that determine gamete mating type [27, 28]. These sexual *Chlamydomonas* reproduction-relevant genes have homologs in closely related species such as those within the *Volvox* genus [29, 30]. *B. braunii* has not been observed to reproduce sexually and is assumed to reproduce only asexually [20]. Since the conventional method for determining species classification cannot be applied, analysis of genome-wide characteristics may offer insights into the question of speciation among the *B. braunii* races.

Comparative genomic studies have been used as evidence for the classification of species. Recently, the genomes of three redpoll bird species in the genus *Acanthus* were used to reclassify these individual species as a single species using a linked supergene that controls phenotypic color differences [31]. Additionally, it is common practice to utilize comparative genomic analysis when classifying newly discovered bacteria [32, 33]. With the increased accuracy of long-read sequencing and improvement of assembly algorithms for organisms with complex, repeat-rich genomes such as plants, a comparative genomic study of *B. braunii* would aid in the current understanding of the organism and the relationships between the three races. Recently, we published a draft nuclear genome assembly for the B race, Berkeley (or Showa) strain of *B. braunii* [34]. For the current study, we sequenced and assembled draft nuclear genomes for the A and L races and performed a comparative genomic analysis of *B. braunii* at the race and larger taxonomic levels. This analysis revealed significant differences at the genomic level between the three *B. braunii* races that we propose are significant enough to reclassify the races as separate species.

Results

Botryococcus braunii genome assemblies

High quality, high molecular weight genomic DNA [34, 35] was isolated from the A, B, and L races, and this DNA was used for sequencing on the Oxford Nanopore Technologies. This long-read genomic sequencing data was combined with previously obtained Illumina short-read sequencing for all three chemical races to combine the benefits of both sequencing technologies for downstream genome assembly. This produced 18-24x coverage (three samples multiplexed and sequenced on two flow cells) from the Nanopore data and 140-316x coverage from the Illumina data. Coverage was calculated based on total bases from sequencing divided by the estimated genome size for each race [16, 17]. For each race, *de novo*, reference-guided, and hybrid (combing Oxford Nanopore and Illumina sequences) assemblies were tested to determine which method produced an optimal draft genome assembly based on assembly metrics and BUSCO scores. For the L race, a *de novo* assembly approach using only the Oxford Nanopore sequence data produced the best assembly. For the A race, a hybrid assembly method, combining both Oxford Nanopore and Illumina sequence data, produced the most contiguous genome assembly. All attempts to produce an improved genome assembly for the B race resulted in sub-par final assemblies and were less contiguous compared to the current B race genome assembly [34].

The genome assembly statistics for three green microalgae with relatively complete assemblies, *Chlamydomonas reinhardtii* [36], *Coccomyxa subellipsoidea* [37], and *Chlorella variabilis* [38], are included in Table 1 as references to evaluate the contiguousness of the new *B. braunii* A and L race genome assemblies. Of the three *B. braunii* race genome assemblies, the L race was the most contiguous and whole with the largest N50 value and the smallest number of contigs/scaffolds, meaning it is less fragmented overall (Table 1). Even though the A race had a similar number of contigs/scaffolds as the B race genome assembly, its L50 was nearly half that of the B race assembly (Table 1), meaning the longest scaffolds assembled for the A race are longer and more contiguous than those of the B race. All three of the *B. braunii* race genome assemblies had a low percentage of fragmented BUSCOs while having varying amounts of single-copy, complete BUSCOs (Table 1). Overall, the new *B. braunii* A and L race genome

Table 1. Assembly statistics for all three chemical races of *B. braunii*. Assembly statistics for all three chemical races of *B. braunii*. A and L chemical races are novel in this study. The *B. braunii* B race, *C. reinhardtii*, *C. variabilis*, and *C. subellipsoidea* serve as a comparison to assess the contiguity of the A and L race genomes. Both A and L race assemblies are at scaffold resolution containing a mix of scaffolds and contigs. BUSCO scores were generated using the Chlorophyta ODBv10.

Species	*B. braunii*, race A	*B. braunii*, race B	*B. Braunii*, race L	*Chlamydomonas reinhardtii*	*Coccomyxa subellipsoidea*	*Chlorella variabilis*
Assembly Size (Mb)	188.3	170.2	135.6	111.1	48.9	46.2
Number of unitigs (scaffolds/contigs)	[a]903 (897/28)	983	[a]485 (8/477)	54	45	414
N50 (Mb)	[b]0.752 (0.375/2.93)	0.565	[b]2.91 (3.94/2.64)	7.78	1.96	1.47
L50	[b]47 (85/9)	99	[b]15 (3/13)	7	9	12
GC Content (%)	49.98	50.82	52.51	64.08	52.93	67.14
[c]BUSCO Score (% Complete/ % Fragmented)	83.9/0.6	89.6/1.8	90.7/0.8	96.5/1.7	98.5/0.4	95.7/2.0
[d]BUSCO Score (% Complete/ % Fragmented)	71.7/1.11	70.4/5.33	83.1/0.65	97.9/0.32	95.8/0.85	91.9/2.56

[a]: The top number indicates number of unitigs (scaffolds plus contigs), first number in parentheses indicates number of scaffolds, second number indicates number of contigs.

[b]: For N50 and L50 metrics, the top number indicates the metrics calculated using combined scaffolds and contigs, first number in parentheses indicates metrics calculated using only scaffolds, second number only contigs.

[c]- BUSCO run in "genome" mode.

[d]- BUSCO run in "protein" mode.

https://doi.org/10.1371/journal.pone.0304144.t001

assemblies are more contiguous than the current B race genome assembly. Interestingly the L race had the smallest assembly size of all three races at 135.6 Mb, and the A race had the largest at 188.3 Mb (Table 1). This contradicts earlier genome size estimates that showed the L race had the largest genome size at 211 Mb followed by the B race at 166 Mb and the A race at 160 Mb [16, 17]. Due to the difference in assembled genome sizes versus earlier estimations, all three race assemblies would benefit from additional sequencing including long-read and Hi-C sequencing in hopes of anchoring scaffolds to achieve chromosome resolution genome assemblies. This marks the first time that all three *B. braunii* races have had genomes assembled.

Genome sizes vary among Trebouxiophyceae

Our early estimations of *B. braunii* genome sizes had placed the A, B, and L races among the largest known genomes for green microalgae at the time [16, 17]. Since these studies, more Chlorophyta species have been sequenced and genomes assembled, annotated, and deposited in public databases such as PhycoCosm [39]. This has allowed for a reexamination of *B. braunii* genome sizes and how they compare to other green microalgae.

For this comparison, we obtained genome size information from PhycoCosm for 14 species of green microalgae from the Trebouxiophyceae and Chlorophyceae clades as shown in S1 Table in S1 File. These species were chosen because of their close relation to *B. braunii* to serve as an in-clade comparison (Trebouxiophyceae) or as an outlier group for external clade comparison (Chlorophyceae). Additional criteria for comparison selection were datasets that included masked genomes, assembled transcriptomes, and annotated gene sets (gtf/gff3 formatted files). The comparison of the *B. braunii* genomes to these 14 species of green microalgae showed that the *B. braunii* genome assemblies are the largest among the Trebouxiophyceae species analyzed, but not the largest when including the Chlorophyceae species, with the *Dunaliella salina* genome [40] being the largest of those analyzed (S1 Table in S1 File). The closest known relative to *B. braunii*, *C. subellipsoidea*, had a substantially smaller (48.9Mb) genome assembly size [37] compared to any of the three *B. braunii* genome assemblies (Table 1 and S1 Table in S1 File). This large difference in genome assembly size would suggest that after divergence from a common ancestor either *B. braunii* experienced a genome size expansion event(s) or *C. subellipsoidea* experienced a genome size reduction event. The genome assembly sizes for the 9 species of Trebouxiophyceae (not including *B. braunii)* analyzed, are in the range of 14-60Mb (S1 Table in S1 File), suggesting *B. braunii* underwent a genome size expansion after delineation from a common ancestor shared with *C. subellipsoidea*.

It has been hypothesized that the cause of the large *B. braunii* genome sizes was a large number of genes encoded in the genomes of each race [16, 17]. This seemed to be confirmed when we published the first *B. braunii* genome from the B race [34], which had 20,765 genes. Now with the A and L race genomes assemblies, this finding can be reanalyzed by comparing genome size and gene count in several Trebouxiophyceae species and *C. reinhardtii* from Chlorophyceae. In general, an increase in genome size correlates with an increase in gene count for most Trebouxiophyceae species analyzed (Fig 1 and S1 Table in S1 File). It should be clarified that in this work the number of genes reported as encoded on the B race genome, 23,685, differs from the published value of 20,765 [34]. This is because the annotation software used to perform gene annotation in the A and L races was also applied to the B race genome (see Methods section for detailed description), due to this software's increased capabilities over what was used when the B race genome was published. Genome size and gene count are not positively correlated among the three races of *B. braunii*. The *B. braunii* L race has a ~90Mb larger genome assembly size compared to *C. variabilis* and has ~6,000 more genes. (Fig

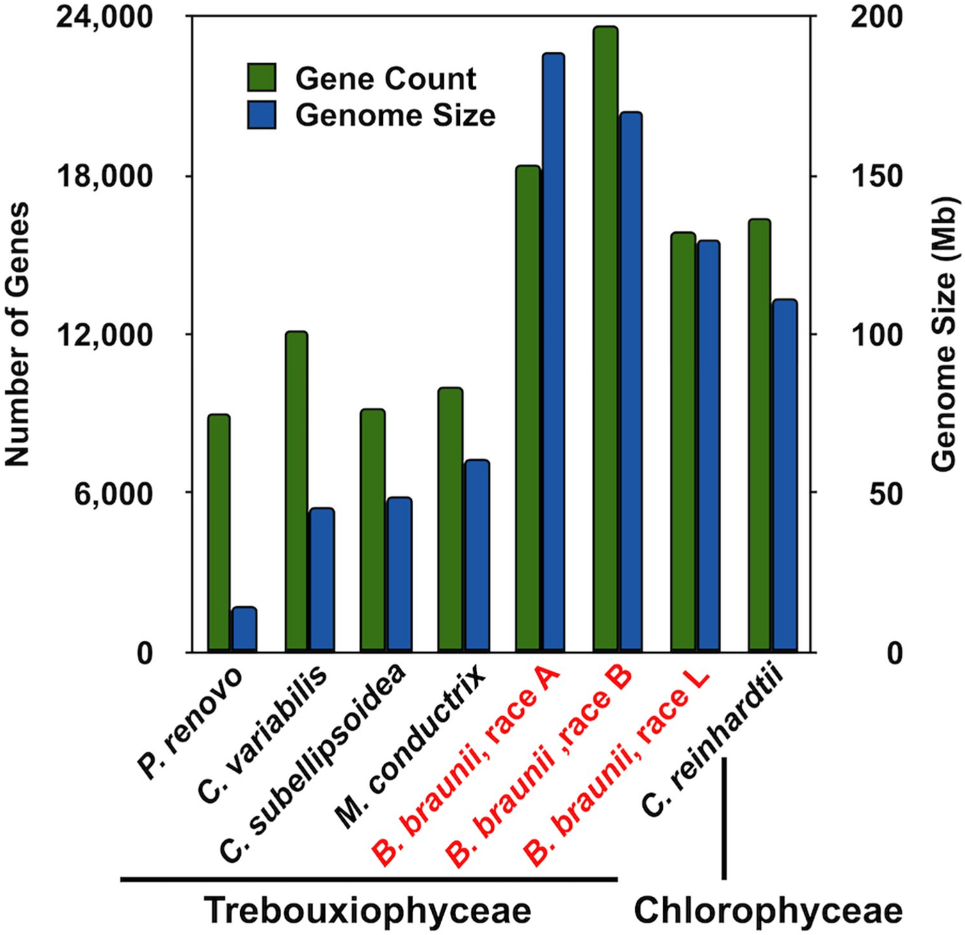

Fig 1. Comparison of gene number and genome size for select Chlorophyta species. Grouped bar chart showing genome assembly size in blue and predicted gene count for each assembly in green. All algae except *C. reinhardtii* are Trebouxiophyceae. *C. reinhardtii* is shown as a reference for a similar size and gene count compared to the *B. braunii* races.

https://doi.org/10.1371/journal.pone.0304144.g001

1 and S1 Table in S1 File). Similarly, *B. braunii* A race has a ~140Mb larger genome assembly size compared to *C. variabilis* and ~9,000 more genes. Within *B. braunii*, the A race, with the largest *B. braunii* genome assembly size, has ~5,200 fewer and ~2,600 more genes than the B and L races, respectively (Fig 1). Additionally, gene coding sequences in each race were analyzed for overlap with repeat content (see below for description of repeat content). This analysis revealed that only the A race contained gene/repeat overlaps with 3,189 genes having a significant overlap (>70%) to repeat elements (S1 File). Interestingly, *Picochlorum renovo* displays the smallest known Trebouxiophyceae genome size, yet has a remarkably large number of encoded genes given the small genome size (Fig 1) [41]. This compact genome may be a consequence of the fitness this species displays in thermo, salinity, and intense light tolerance [42–45]. Since gene count did not appear to be the sole factor accounting for the large *B. braunii* assembled genome sizes observed, a more detailed investigation into the genomes of *B. braunii* was carried out.

Gene feature distribution of *B. braunii*

Having genome assemblies for the first time for each race of *B. braunii* allowed for direct comparison of gene feature distributions such as the length of 5'-UTRs, 3'-UTRs, and introns.

These *B*. *braunii* gene features were compared to two Trebouxiophyceae species, *C*. *variabilis* and *C*. *subellipsoidea*, and the Chlorophyceae species *C*. *reinhardtii*. These species were chosen as a representative sample of different levels of relation to *B*. *braunii* given our phylogenomic analysis (see Fig 4): *C*. *subellipsoidea* as the closest relative, *C*. *variabilis* as a more distant relative, and *C*. *reinhardtii* as a distant relative form a different clade. The distribution of 5'- and 3'-UTRs was similar for each species analyzed with the largest number of UTR regions in each species being between ~100 and ~500 bp (Fig 2A and 2B). However, differences between species can be seen in the small number of large UTRs seen in each species. For example, among the three *B*. *braunii* races the B race has the longest 5'-UTRs and 3'-UTRs (Fig 2A and 2B). Interestingly, the closely related *C*. *subellipsoidea* had the smallest number of and shortest length of both types of UTRs (Fig 2A and 2B). For intron length distribution, the B race of *B*. *braunii* had a smaller number of and shorter length of introns compared to the A and L races (Fig 2C). This phenomenon could explain the large number of genes encoded in the B race genome assembly, but may also be an artifact of its current highly fragmented state.

DNA repeat content of *B*. *braunii*

The amount of DNA repeat content was examined to explain the relatively large genome sizes of the *B*. *braunii* races since a correlation between the number of putative genes and genome size was not seen (Fig 1). Transposable element (TE) and tandem repeat (TR; a.k.a. satellites and low-complexity repeats) content within the genomic space of all three *B*. *braunii* races were analyzed using RepeatModeler [46] and RepeatMasker [47] for TEs and the Tandem Repeats Finder program [48] for TRs. The B race exhibited the largest amount of total repeat content at 43% of the total genome size (Fig 3B), while the A and L races had less total repeat content at 39%, and 24%, respectively, of the total genome size (Fig 3A and 3C). The makeup of TEs in each race was also analyzed. All three races had more class I (retrotransposons) than class II (DNA transposons) TEs (Fig 3A–3C). Within the class I TEs, each race had a different major category with long interspersed nuclear elements (LINEs) dominating in the A race, long terminal repeats (LTRs) predominant in the B race, and LINEs and LTRs roughly equal in the L race (Fig 3A–3C). Short interspersed nuclear elements (SINEs) were only found in the B race as a minor component of class I TEs (Fig 3A–3C). Each race had different distributions and amounts of each subclass of class I and class II TEs (S1 File). The B race exhibited the most TR content at 22%, with the A and L chemical races at 16% and 11%, respectively (Fig 3A–3C).

To see how *B*. *braunii* total repeat content compared to other Chlorophyta species, the same DNA repeat analysis was performed on the closely related *C*. *subellipsoidea* and the three *Chlamydomonas* species. All of these species, except *C*. *subellipsoidea*, contained similar total repeat content compared to the three *B*. *braunii* races (Fig 3A–3G). While *C*. *subellipsoidea* contained little TE content, those TEs were made up of only class I LINEs and class II TEs (Fig 3G and S1 File). The three *Chlamydomonas* species contained class I LINEs as the main TE with varying amounts of LTRs and class II TEs (Fig 3D–3F). Like the observation that genome assembly size did not linearly correlate to the number genes annotated within each race of *B*. *braunii* (Fig 1), repeat content is also not linear with genome assembly size. For example, the A race has the largest genome assembly size of the three *B*. *braunii* races (Table 1), yet displayed 4% less total repeat content than the B race (Fig 3A and 3B). In contrast, *C*. *schloesseri* and *C*. *Incerta* have the largest genome size of the *Chlamydomonas* species (S1 Table in S1 File) and the most TE content (Fig 3D and 3E) analyzed here. Taken together, these data suggest TEs and TRs contribute to the observed large genome assembly sizes of the three races of *B*. *braunii*.

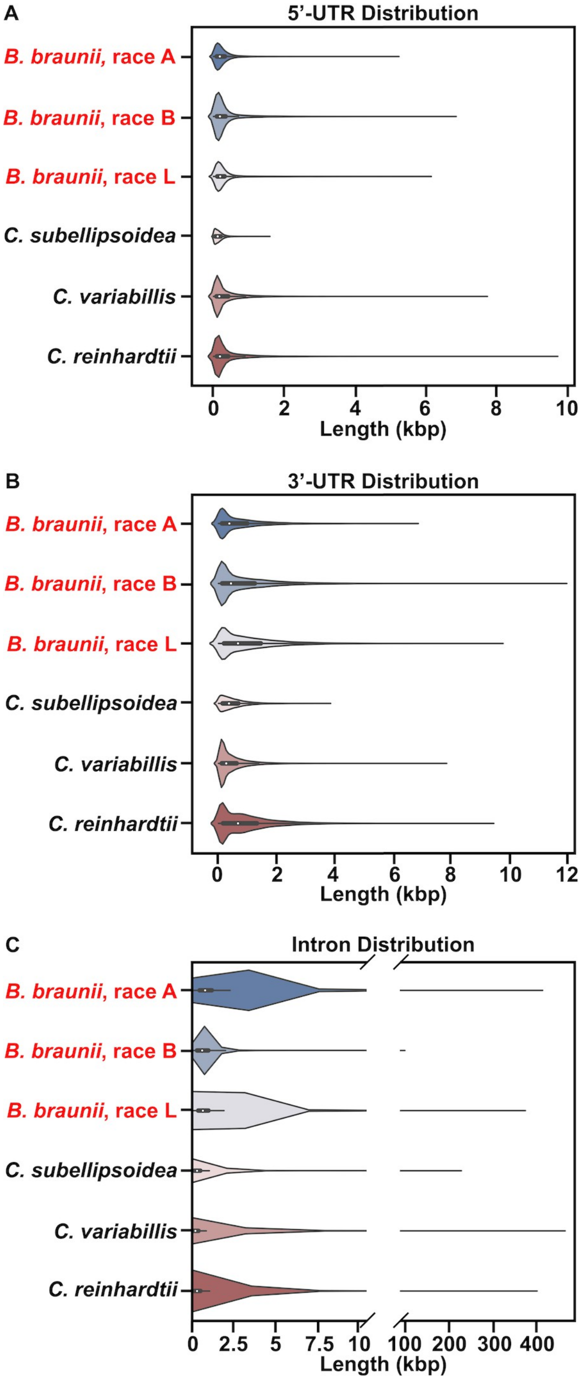

Fig 2. Non-coding gene feature distribution analysis. Box plots inside violin plots for (A) 5'-untranslated regions (UTRs), (B) 3'-untranslated regions (UTR)s, and (C) Intron lengths in the three races of *B. braunii*, *C. subellipsoidea*, *C. variabilis*, and *C. reinhardtii*. Box plots indicate length distribution with the median gene feature length indicated by a white dot inside the box plot. Violin plots show the frequency (width of the curve) of gene features at a given length (x-axis).

https://doi.org/10.1371/journal.pone.0304144.g002

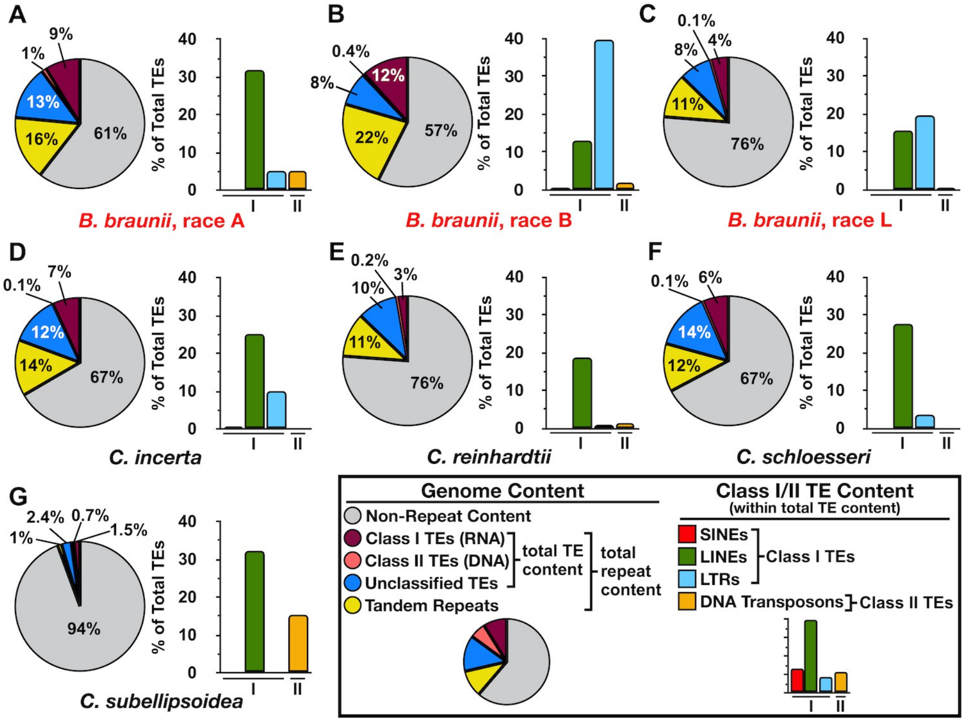

Fig 3. DNA repeat content of select Chlorophyta species. Total repeat content, transposable elements (TE) plus tandem repeats (TR), and the breakdown of TE type was calculated for (A) *B*. *braunii*, race A, (B) *B*. *braunii*, race B, (C) *B*. *braunii*, race L, (D) *C*. *incerta*, (E) *C*. *reinhardtii*, (F) *C*. *schloesseri*, and (G) *C*. *subellipsoidea*. The pie chart represents the amount of the genome assembly for each microalga that is made up of repeat elements by showing the total class I TEs (magenta), class II TEs (salmon), unclassified TEs (blue), total TR (yellow), and non-repeat content (grey) for each genome. The bar graph shows the breakdown of class I TEs, SINEs (red), LINEs (green), and LTRs (light blue), and class II TEs, DNA transposons (orange), as a percent of the total TE content for each race. The distribution of all repeat elements, including class I and class II TE subclasses, for each species can be found in S1 File.

https://doi.org/10.1371/journal.pone.0304144.g003

Comparative genomics analysis of the *B*. *braunii* races with other algal species

Phylogenomic comparison of multiple chlorophyta species. Having genome assemblies of all three *B*. *braunii* races allowed for a comparative genomic analysis with other algal species. First, a phylogenetic comparison was carried out using a large dataset of genes conserved among 17 chosen algae species, including the three *B*. *braunii* races. By using a large dataset of conserved genes, more accurate phylogenetic relationships can be inferred through this type of phylogenomic analysis [49]. The species used for this analysis are listed in S1 Table in S1 File and were chosen based on having genomes, transcriptomes, and predicted gene datasets deposited and publicly available from PhycoCosm. Of the 17 species chosen, 12 were from Trebouxiophyceae (including the three *B*. *braunii* races) and five were from Chlorophyceae. To obtain the gene data set to be used in the phylogenomic analysis we utilized ChlorophytaODB v10 [50], which is a set of single-copy gene orthologs manually curated from 16 genomes of various *Chlorophyta* taxa included in the OrthoDB v10 database. At the time of this study, the ChlorophytaODB v10 had a total of 1,569 orthologous genes. A genome-wide evaluation of the 17 species revealed 1,189 of these genes were shared as single-copy orthologs while allowing any ortholog to be absent in up to three organisms.

Using these conserved, single-copy orthologs, a phylogenomic tree was constructed for the 17 species analyzed (Fig 4). This tree was constructed by first generating a maximum-

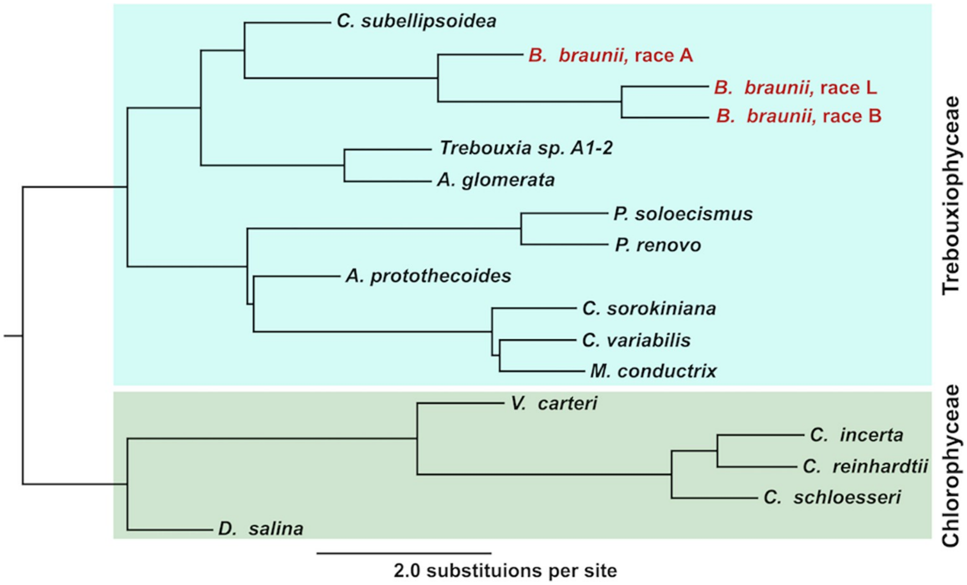

Fig 4. Phylogenomics of chlorophyta. Phylogenomic tree constructed using 1,189 conserved single-copy orthologs among 17 species from the Chlorophyta division. The Trebouxiophyceae class species including all three races of *B. braunii* are outlined in turquoise. The Chlorophyceae class species, used to root the tree, are outlined in olive green. Legend bar is in substitutions per site.

https://doi.org/10.1371/journal.pone.0304144.g004

likelihood (ML) tree for each of the 1,189 orthologous genes. Then all 1,189 ML trees were reconciled into a single phylogenomic tree that was rooted in the Chlorophyceae sub-clade containing *Volvox carteri*, the three *Chlamydomonas* species, and *D. salina* (Fig 4). Phylogenetic inference placed *C. subellipsoidea*, as the most common species branching from the same common ancestor as *B. braunii* (Fig 4). The *B. braunii* A race was the first to diverge from a common *B. braunii* ancestor with the B and L races diverging from each other at a later time point (Fig 4). Additionally, the A race of *B. braunii* exhibits more genetic divergence from the B and L races as compared to the genetic divergence among the three species of *Chlamydomonas* (Fig 4). This phylogenomic analysis led us to conclude at this point that, at minimum, the *B. braunii* A race is likely a separate species from the *B. braunii* B and L races.

Gene family analysis

The number of genes and gene families shared among the three races of *B. braunii* in comparison to separate but closely related species can be used to support defining the A race as a separate species and provide data to determine if the B and L races are separate species. Thus, gene relationships were inferred using OrthoFinder [51], a program package that reconstructs gene orthologs and paralogs based on sequence identity shared at both the inter- and intraspecies levels into groups that are called orthogroups (hereafter referred to as gene families). This analysis was carried out on the 17 species analyzed in the phylogenomic tree and found a combined total of 226,181 genes and 22,024 gene families (S2 Table in S1 File). Next, an analysis of gene family relationships between organisms of the same genus was performed. Since the phylogenomic analysis (Fig 4) suggests the A race is likely a separate species from the B and L races, we investigated the amount of A race unique gene families not shared with the B and L races, or if the B and L races shared more gene families. An UpSet plot [52] was generated to visualize gene family intersections between 9 species: 6 Trebouxiophyceae algae: *C. subellipsoidea*, *C. variabilis*, *C. sorokinia*, and the three *B. braunii* races, and three Chlorophyceae species:

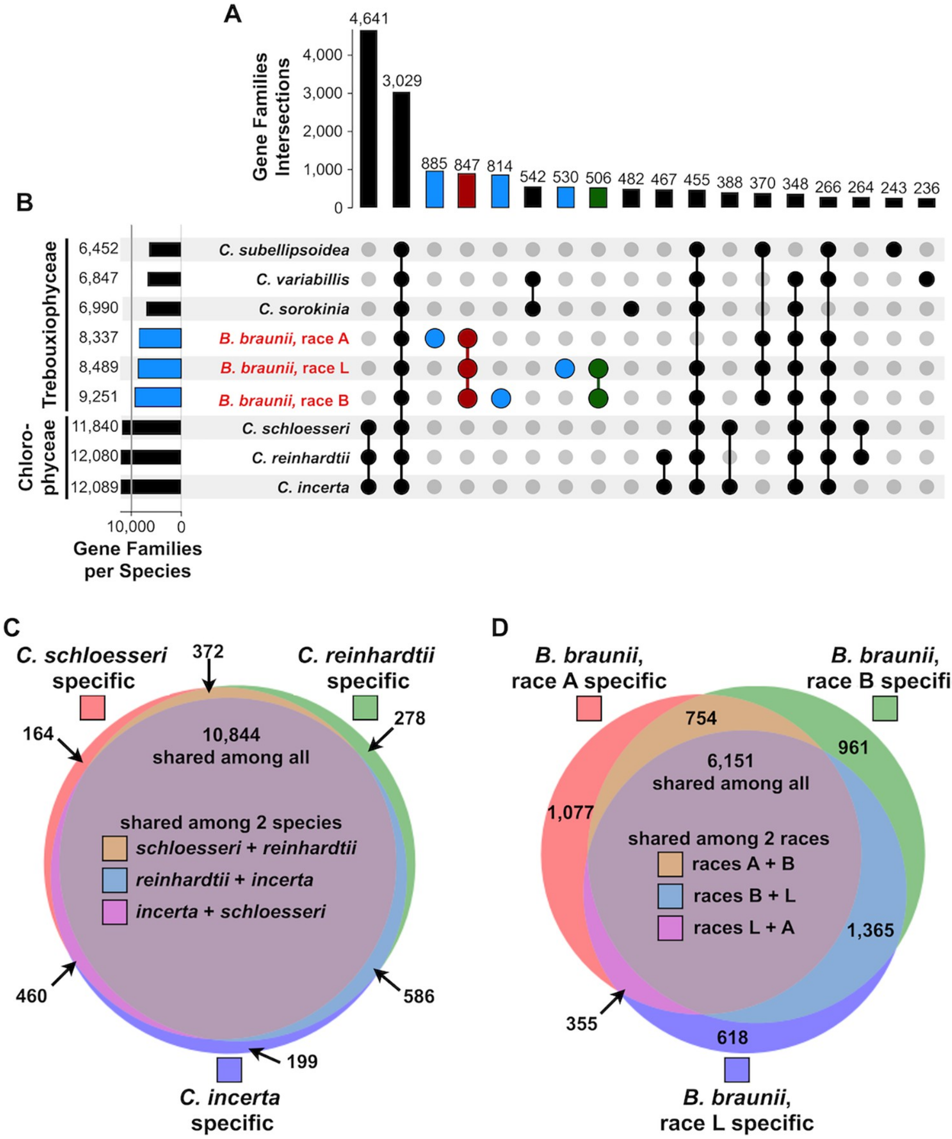

Fig 5. Comparative gene family analysis. UpSet plot showing the orthogroups or gene families identified by OrthoFinder grouped by species. (A) Bar chart showing the number of gene families that intersect with the given species in (B). (B) Shared gene families at different species intersections. Dots connect the species that contain the number of gene families shown in (A). Horizontal bar char in (B) corresponds to the total number of gene families with a given species. (C) and (D) Venn diagrams of shared gene families between the three *Chlamydomonas* species (C) and the three *B. braunii* races (D).

https://doi.org/10.1371/journal.pone.0304144.g005

the three *Chlamydomonas* species *C. reinhardtii*, *C. schloesseri*, and *C. incerta* (Fig 5A and 5B). The top portion of the UpSet plot (Fig 5A) groups intersections of the reconstructed gene families represented by the accompanying connected dot matrix (Fig 5B). The three *Chlamydomonas* species together had the most genus-specific gene families at 4,641, and the next highest intersection of gene families was 3,029 families shared by all 9 algae (Fig 5A and 5B). While each race of *B. braunii* shared 847 gene families (red dots, Fig 5B), each race had a large number of gene families specific to each race (blue dots, Fig 5B). This was not seen for the three *Chlamydomonas* species. Interestingly, the B and L races shared 506 gene families (green dots, Fig 5B), while neither the B nor L shared a large (>200) number of gene families with the A race. This supports the relationships seen between the *B. braunii* races in the phylogenomic analysis (Fig 4) and suggests there are more genes, and as a consequence biosynthetic/metabolic pathways, that are similar or shared between the B and L races that the A race lacks.

It would be reasonable to expect a set of gene families shared by a common ancestor of all Trebouxiophyceae and have some portion of them retained in the divergent organisms. However, there was not a large number of gene families (>200) specific only to the Trebouxiophyceae algae (Fig 5A and 5B). Interestingly, the three races of *B. braunii* exhibited the largest amount of genus-specific gene families of the Trebouxiophyceae algae analyzed (red dots, Fig 5B).

The large difference in unique gene families identified in each *B. braunii* race led us to further investigate the gene families shared between the three races in comparison to the three closely related *Chlamydomonas* species using OrthoFinder [51] and weighted Venn diagrams (Fig 5C and 5D). There were a total of 12,903 gene families identified in the three *Chlamydomonas* species, jointly they shared 10,844, or ~83%, of these gene families, and the gene families shared between any two of the *Chlamydomonas* species ranged from 372 to 586 (Fig 5C). The three races of *B. braunii* contained 11,281 gene families, 6,151 of which, or ~55%, were shared among the three races, and the gene families shared between any two races ranging from 355 to 1,365 (Fig 5D). Additionally, there was a greater number of gene families specific to each race of *B. braunii* than specific to each *Chlamydomonas* species (Fig 5C and 5D). Taken together these data indicate there is a greater degree of diversity in the gene families between the three races of *B. braunii* than between the three *Chlamydomonas* species. These significant differences in the genic space between the three *B. braunii* races offer genome-wide evidence supporting each race of *B. braunii* is an individual species.

Gene duplications. Commonly, some genes cannot be placed in an orthogroup by OrthoFinder when analyzing gene families. These genes are marked as unassigned and referred to as lineage-specific genes (LSGs) unique to a given species [53]. Over time an LSG may undergo gene duplication producing lineage-specific paralogs [54, 55], and are identified by OrthoFinder as a lineage-specific orthogroup (LSO). LSG and LSO analysis was carried out on the 17 algal species used in this study. In general, the three *B. braunii* races have more LSGs and LSOs than the other algae analyzed except *Trebouxia sp*. A1-2 and *D. salina* (Fig 6A and S2 Table in S1 File). Among the *B. braunii* races, the B race contained the largest number of LSGs with 5,160, while the A and L races had fewer at 2,160 and 1,660, respectively (Fig 6A). For the number of genes in LSOs, the A race had the most among the *B. braunii* races at 3,741 with the B and L races having 3,298 and 2,095, respectively (Fig 6A). On a percentage basis, each race of *B. braunii* had a high percentage of LSGs and LSOs compared to most of the other algae (Fig 6A). The genes comprising the LSOs and LSGs identified in each race of *B. braunii* were analyzed for the presence of annotated domains using the Pfam database. Only 17.4%, 15.3%, and 18.3% of the genes in the A, B, and L races, respectively, contained an annotated Pfam domain (data not shown).

In comparison, the three closely related *Chlamydomonas* species had low and relatively equal LSGs and LSOs (Fig 6A), again indicating a large amount of genetic diversity between the three *B. braunii* races. Since no apparent whole genome duplication (WGD) events have been observed or inferred in *B. braunii* [56] it is likely the large number of LSGs and LSOs are the result of gene duplication events and subsequent functionalization [57–59]. This analysis also indicates that anywhere from 1 in 5 to 1 in 3 genes are unique to a given *B. braunii* race. In the *Chlamydomonas* only 1 in ~33 genes are unique to a given species between *C. reinhardtii*, *C. incerta*, and *C. schloesseri*. These findings further highlight the differences in the genomic space between the three *B. braunii* races that are not seen in the *Chlamydomonas* species.

Gene gain/loss analysis. To investigate if gene duplication events could be a contributing factor to the large number of unique genes identified in each race of *B. braunii*, the evolution

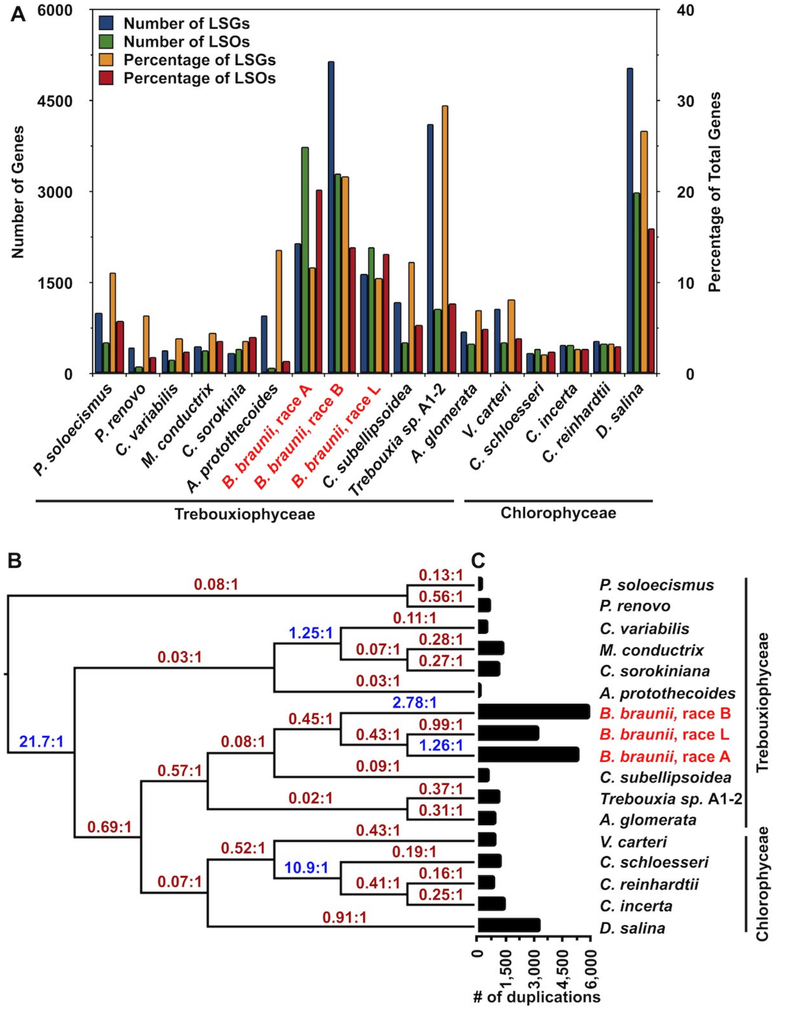

Fig 6. Gene duplication event analysis in chlorophyta species. (A) Distribution of the number and percentage of lineage-specific genes (LSGs) and lineage-specific orthogroups (LSOs) for each species listed. (B) Phylogenetic trees output for each single-copy ortholog from OrthoFinder were reconciled into a single tree with branch lengths mapped to gene gain/losses by Notung. At each evolutionary node break, the ratio of gene gain to loss from duplication events is labeled (e.g. 21.7:1). If the ratio indicates a net increase in gene gain it is colored in blue text. If the ratio indicates a net decrease in gene gain it is in red text. Ratios were calculated by dividing the total number of gains by losses for each branch. Tree branches represent the number of genes gained and loss from the total number of duplications calculated at each tree node. (C) Total gene duplication events per species inferred by OrthoFinder mapped as a horizontal bar chart.

https://doi.org/10.1371/journal.pone.0304144.g006

of gene duplication events were tracked for all 17 algal species used in this study. 22,024 individual orthogroup phylogenetic trees from the OrthoFinder analysis were used as input to Notung [60] to determine whether a gene duplicate was retained (gain) or lost (loss). The output of Notung is a single phylogenetic tree (Fig 6B) reconciled from the OrthoFinder inputs. The total number of gains or losses for each branch in the reconciled phylogenetic tree was extracted and counted and the gain-to-loss ratios calculated (Fig 6B). Lastly, gene duplication events with high bootstrap evidence (50%) for each species were counted from the individual gene family trees from OrthoFinder (Fig 6C and S1 Fig in S1 File). *P. renovo* and *Picochlorum soloecismus* separated into a single clade distinct from all other species and experienced more losses than gains with a high ratio of loss to gain over any evolutionary period (Fig 6B). This may contribute to the small genome sizes of these algae compared to other *Chlorophyta* microalgae (S1 Table in S1 File). There were a few instances of net gene gain from duplications in the analysis. For example, the split delineating the other 15 species from *P. renovo* and *P. soloecismus*, the delineation of *Auxenochlorella protothecoides* from the clade containing *C. variabilis*, *Micractinium conductrix*, and *C. sorokiniana*, and the delineation of the *Chlamydomonas* species from *V. carteri* all had more gene gains than losses (Fig 6B and S1 Fig in S1 File). Importantly, the resulting tree topology is different between Figs 4 and 6B and arises from the different analyses that were performed. While Fig 4 is comprised of single-copy orthologs identified in at least 14 of the 17 species analyzed and infers tree topology from sequence similarity/identity, the phylogenetic inference depicted in Fig 6B focuses only on gene duplication events, and not on sequence identity itself.

The three *B. braunii* races experienced varying levels of gene gain/loss from duplication events. The B race experienced the highest gain-to-loss ratio of 2.78 gains per loss from duplication events, the A race experienced 1.26 gains per loss, and the L race was the only race to experience a higher ratio of loss to gain from gene duplication events, at 0.99 gains per loss (Fig 6B). In terms of absolute numbers of gene duplications, *B. braunii* experienced the largest amount of gene duplication events among the 17 species analyzed (Fig 6C and S1 Fig in S1 File). Specifically, the B race had the largest amount of gene duplication events, followed by the A race, while the L race had the fourth-largest number of genome duplication events falling below *D. salina* (Fig 6C and S1 Fig in S1 File). This is interesting considering the *D. salina* genome size is nearly triple that of the largest *B. braunii* A race genome assembly (S1 Table in S1 File). A large amount of *B. braunii* gene duplication events (Fig 6C and S1 Fig in S1 File) may have given rise to their significantly larger genomes when compared to other Trebouxiophyceae algae (S1 Table in S1 File). Additionally, the varying ratios of gain to loss from gene duplication events between the three races suggest that over an evolutionary period, the races of *B. braunii* have experienced significantly different changes in their genic space. Most *Chlorophyta* species, such as the closely related *Chlamydomonas* species, appear to have lower levels of gene duplication events, and never experienced retention of genes through duplication events at the time of species delineation [61]. The finding that the A and B races of *B. braunii* had more retention of genes over losses from duplication events is the first report of this finding for a C*hlorophyta* species. Since the B and L races have different gene gain-to-loss ratios (Fig 6B), it suggests the B and L races underwent divergent genome expansion or reduction events after delineation from a common ancestor. These differences mark another separation between the B and L races.

Analysis of hydrocarbon biosynthesis gene distribution within the three races of *B. braunii*

Historically, it has been noted that the three races of *B. braunii* differ from each other mainly by the type of hydrocarbon produced by each race [7, 8, 62]. That is to say, the hydrocarbons of one race are not found in the other two races. It has been presumed this is due to each race having only the genes specific to the production of the hydrocarbons in that race. However, it is possible that hydrocarbon biosynthesis genes from one race exist in the other races but are not expressed. With the genome sequence for each race, the existence of these genes in each race of *B. braunii* can be analyzed. Thus, each *B. braunii* race was analyzed for the presence of known hydrocarbon biosynthesis genes from each race using HMMER searches with specific Pfam protein domains and BLAST searches for hydrocarbon genes from each race.

Distribution of potential A race hydrocarbon biosynthesis genes within *B. braunii*: Long-chain fatty acid synthesis genes. Though the hydrocarbon biosynthesis genes responsible for alkadiene/triene biosynthesis in the A race of *B*. *braunii* have yet to be identified, evidence suggests that these hydrocarbons are derived from the elongation of fatty acids to produce very long chain fatty acids (VLCFA) followed by cleavage of the carboxylic acid moiety to generate an alkene with a terminal double bond [10, 11, 13]. Production of VLCFAs through FA elongation has been well studied in many organisms, is highly conserved, the genes/enzymes involved have been identified, and is a four-step process [63]. The first enzyme, 3-ketoacyl-CoA synthase (KCS) adds an acetate moiety from malonyl-CoA to the FA generating a β-carbonyl, the second enzyme, 3-ketoacyl-CoA reductase (KCR), reduces the β-carbonyl to a β-hydroxyl, the third enzyme, 3-hydroxyacyl-CoA dehydratase (HCD), dehydrates the β-hydroxyl to a C=C bond, and the fourth enzyme, *trans*-2,3-enoyl-CoA reductase (ECR), reduces the C=C bond [63]. These 4 reactions repeat until the needed acyl chain length is produced. The KCR, HCD, and ECR genes are highly conserved, are single-copy genes, and accept a broad range of acyl chain lengths [63]. The KCS genes belong to a large gene family [64, 65] and each encoded enzyme can be very selective for substrate acyl chain lengths [66]. KCS genes can be subdivided into two subfamilies: the plant specific fatty acid elongase (FAE)-type and the more universal elongation deficient (ELO)-type [63, 67, 68].

As a starting point and in an attempt to identify A race genes/enzymes specific for alkadiene/triene production we analyzed each race for the presence of FAE- and ELO-type KCS proteins, and the single-copy KCR, HCD, and ECR proteins. The Pfam ketoacyl-synt domain (Pfam: PF08392) was used to search for the FAE-type KCS sequences and a total of 63 targets distributed among the three *B*. *braunii* races were found with 16, 20, and 25 for the A, B, and L races, respectively (S2 Fig in S1 File). A search for ELO sequences in each race using the Pfam ELO domain (Pfam: PF01151) revealed a total of four targets in the A race, eight in the B race, and three in the L race (S3 Fig in S1 File). The KCR and HCD genes were identified by BLAST using the proteins for single-copy genes identified in *Arabidopsis thaliana* (NP_564905.1, and NP_193180.1 respectively) that were downloaded from NCBI and used as queries in a BLAST + protein alignment. For ECR, the Pfam enoyl-CoA reductase domain (Pfam: PF12241) was used to search for homologous sequences in each race. These searches yielded more than one target sequence per race, which were narrowed down to a single protein sequence when possible based on nearest homolog analysis against the NCBI non-redundant protein database. The KCR protein search revealed a single target in each race. However, for the HCD and ECR searches a single target could be identified in only two of the three races. For the HCD search, a single target was found in the A and L races, but not in the B race. For the ECR search, a single target was found in the A and B races, but not in the L race. From this analysis, it is difficult to determine if any of these sequences are specific for the A race and contribute to alkene biosynthesis. A detailed functional analysis of the encoded genes will be required in the future.

Distribution of B and L race hydrocarbon biosynthesis genes within *B. braunii*: Squalene synthase/phytoene synthase family genes. Squalene synthase (SS) condenses two molecules of C_{15} farnesyl diphosphate (FPP) to produce C_{30} squalene as the first committed step in sterol biosynthesis in eukaryotic organisms and is generally found as a single-copy gene in yeast and humans [69, 70], but multiple copies have been isolated in some plant species [71–75]. Phytoene synthase (PYS) catalyzes a reaction similar to SS by utilizing C_{20} geranylgeranyl diphosphate (GGPP) to produce C_{40} phytoene [76], which is structurally similar to squalene. Thus, PYS can be considered an SS-like sequence, and most photosynthetic organisms have a single-copy of *PYS*. Each race of *B*. *braunii* has been shown to contain a single SS gene [18, 77]. In addition to squalene production by the conventional SS, biosynthesis of the triterpene hydrocarbon botryococcene in the B race has been shown to use the SS-like genes SSL-1 and SSL-3 [78], In relation to this, there is an alternative route for squalene production by the combination of SSL-1 and SSL-2 in the B race [78]. Another SS-like enzyme, lycopaoctaene synthase (LOS), is responsible for lycopadiene production in the L race [18]. *PYS* genes have not yet been published for any race of *B. braunii*. All three races were screened for the presence of all SS, SS-like, and PYS protein sequences using the curated SQS-PSY Pfam domain (Pfam: PF00494). The analysis found a single SS sequence for each race, sequences corresponding to the SSL1, 2, and 3 were only found in the B race, and LOS was only found in the L race (S4 Fig in S1 File). Thus, the hydrocarbon biosynthesis enzymes for the B and L races are only found in those races. Additionally, a single PYS was found in each race (S4 Fig in S1 File).

Meiosis genes present in the genome assemblies of *B. braunii*. To date, no evidence has been presented indicating that any race of *B. braunii* displays sexual reproduction behavior [20]. Within Chlorophyta, there are several species that are suspected to display cryptic sex [79]. For example, there are several Trebouxiophyceae species that encode meiosis genes within their genomes including several analyzed in this study: *A. protothecoides*, *C. variabilis*, and *C. subellipsoidea* [79]. To investigate if *B. braunii* contains the ability for sexual reproduction, the genome assemblies of each race of *B. braunii* were searched for nine essential meiotic genes: DMC1, HOP1, HOP2, MER3, MND1, MSH4, MSH5, REC8, and SPO11 [79]. Orthologs for each of these genes, with supporting Pfam domains, were identified in the three races of *B. braunii*, suggesting the presence of a meiosis pathway in the genomes of all three races and cryptic sex in this alga.

Discussion

The current understanding of *B. braunii* has been largely limited to studying morphological and hydrocarbon content/biosynthesis differences between the three races. This limitation is due to a lack of "omics" datasets for each race, and recently only the B race of *B. braunii* had a sequenced genome available [34]. The current study aimed to obtain genomic sequences for the A and L races to perform a comparative genomic analysis of the three *B. braunii* races. This allowed us to probe for differences between the races other than hydrocarbon biosynthesis. Comparative genomics has long been leveraged to investigate relationships between similar organisms. *Arabidopsis thaliana* is a single species made up of many ecotypes (natural variants), classified based on geographical and subtle phenotypic differences [80]. These ecotypes have been investigated at the genomic sequence and TE level [81–83]. Multiple ecotypes of *A. thaliana* had near identical distributions and composition of TEs in all five chromosomes [83], yet the only large difference at the genomic level is a moderate variance in polymorphisms between genomes [81]. One major difference observed among the ecotypes has been a large variation in transcriptional response to environmental stressors, likely a result of different selective pressures from their respective geographical environments [84]. *A. thaliana* ecotypes are somewhat akin to the *B. braunii* chemical races; the races have multiple strains isolated from different waters [20]. However, unlike the *A. thaliana* ecotypes, our comparative genomic analysis revealed substantial differences between the three races in their genomic space indicating the races are separate species.

Our findings suggest the three races divergently evolved from a common ancestor and each race contains substantial variability within the genomic space for each to be considered a separate species (Figs 1, 3A-3C and 6A). By comparison, the three *Chlamydomonas* species included in our analyses were more similar to each other in the genomic space than the three *B. braunii* races are to each other (Fig 5C and 5D). Thus, we suggest a reclassification of the *B. braunii* races into three distinct species of the genus *Botryococcus*.

New *Botryococcus* species names

The *B. braunii* species was first named in 1849 by Friedrich Traugott Kützing [85] and was named after Alexander Karl Heinrich Braun, a 19th century botanist [86]. The A and B races of

B. *braunii* were classified in 1985 [62] and the L race in 1987 [8] based on the type of hydrocarbons produced in each race. Based on the data presented here we created new species names for the three races of *B*. *braunii* using the general rule for generating species names that the species name should match the gender of the genus name [87]. *Botryococcus* is considered as male [86]. Thus, we propose the following species names. The A race is renamed *Botryococcus alkenealis*. The general term for the A race hydrocarbons, alkene, is used with the Latin suffix *-alis* added, which is male or female and means pertaining to, i.e. pertaining to alkenes. The B race retains the name *Botryococcus braunii* to preserve the history of being the first to have its hydrocarbons identified, i.e. botryococcenes [88]. The L race is renamed *Botryococcus lycopadienor*. The root of the hydrocarbon name for the L race, lycopadiene, is used with the Latin suffix *-or*, which is a male suffix with no descriptive meaning. Hereafter, the three races of *B*. *braunii* will be referred to by these new species names and all researchers in the *Botryococcus* field are encouraged to adopt this new nomenclature.

Evidence in support of the chemical races as three different species

Genome repeat content. It is well established that repeat content is the largest driving force in genome size and is responsible for some of the largest observed plant genomes to date [89, 90]. In the core-*Chlamydomonas* species, extensive TE curation revealed large differences in TE content among closely related species relative to *C*. *reinhardtii* [91]. We observed similar differences between the three *Botryococcus* species (Fig 3A–3C and S1 File), and the large repeat content of the three *Botryococcus* species likely contributes to their large genome sizes. Additionally, the large gene count annotated in the *B*. *braunii* genome likely contributes to its large assembly size. Future studies on the repeat content in *Botryococcus* species should look at multiple representative strains from each species (i.e. each former race) to investigate whether strains of the same species contain similar or varying amounts of repeat content. Variation of TE content is observed in other plant species of the same genus, like those within the genuses of *Oryza*, *Zea*, and *Arabidopsis* [92–94]. The large variation of TE content between *B*. *alkenealis*, *B*. *braunii*, and *B*. *lycopadienor* supports our theory that the three former chemical races are three distinct species. While automated TE annotation is convenient and fast, it can be error prone and cannot replace manual TE curation. Thus, our TE analysis likely represents only a portion of the actual TE content present in the *Botryococcus* genomes. Manual curation of TEs would generate a more robust cataloging of the TE content in each *Botryococcus* species and provide a better picture of the contribution of repeat content to *Botryococcus* genome size.

Genome-wide phylogenomics. Prior 18S rRNA-based phylogenetic inferences have suggested the three races are three different species [9, 16, 17, 20, 95]. These studies showed that *B*. *alkenealis* (A race) diverged earlier and separated from a common ancestor with *B*. *braunii* and *B*. *lycopadienor* (B and L races, respectively). Nearly all these studies concluded that *B*. *alkenealis* was likely a different species but there was no observable separation between the *B*. *braunii* and *B*. *lycopadienor*. However, these observations are limited since the phylogenetic trees generated were based on a single gene limiting the ability to infer divergent evolution [49]. In our phylogenomic analysis (Fig 4), the use of 1,189 highly conserved orthologs indicates a clear separation between the three *Botryococcus* species and supports the theory that *B*. *braunii* and *B*. *lycopadienor* diverged later than *B*. *alkenealis* from a common ancestor. To date, this is the first genome-wide evidence supporting the theory of divergent evolution between the *Botryococcus* species (i.e. the three races). This is further supported by the divergence of specialized secondary metabolite biosynthetic pathways related to hydrocarbon biosynthesis. An analysis using derivatives of the *B*. *braunii* and *B*. *lycopadienor* hydrocarbons as biomarkers in sediments concluded that the genes for these hydrocarbons likely evolved no more than 55 million years ago [96]. This analysis also shows that *B*. *braunii* and *B*. *lycopadienor* diverged separately but at a similar rate from a later shared ancestor that appeared after *B*. *alkenealis* diverged. It is important to note that even though the hydrocarbons biosynthesized by *B*. *braunii* and *B*. *lycopadienor* are different they are both derived from isoprenoid biosynthetic pathways and utilize similar enzymes that may have arose from gene duplication and neo-functionalization [18, 78]; the SS-like SSL and LOS genes in *B*. *braunii* and *B*. *lycopadienor*, respectively. This close similarity in the biosynthesis of their major metabolite reflects their divergence with each other from *B*. *alkenealis* (Fig 4). Generating chromosome resolution genome assemblies for all three *Botryococcus* species could further isolate similarities and differences between *B*. *braunii* and *B*. *lycopadienor*.

Gene content. We add evidence for separate species based on the comparison of gene content between the *Botryococcus* species. The genome-wide gene family comparison of the 17 Chlorophyta species (Figs 5 and 6) revealed large differences in gene family content between the three *Botryococcus* species that are larger than the differences between related taxa that are classified as different species. For example, the three *Chlamydomonas* species share ~83% of their total gene families, while all three *Botryococcus* species share ~55% of total gene families (Fig 5C and 5D). This data combined with previous genetic divergence rates of 18S rRNA genes [9] supports our proposal of three separate *Botryococcus* species.

All three *Botryococcus* species had a large number of gene duplication events leading to the acquisition of new genes (Fig 6C). However, *B*. *alkenealis* and *B*. *braunii* had approximately 2-fold more gene duplication events than *B*. *lycopadienor*, and all three *Botryococcus* species had nearly 24-36% of their total gene pool classified as specific to that species (Fig 6A). Additionally, the three *Botryococcus* species underwent different rates of gene duplication/loss ratios over a recent evolutionary period (Fig 6B). Taken together these observations suggest that after duplication of orthologs and paralogs these genes underwent a high level of neo-functionalization and/or sub-functionalization in the *Botryococcus* genus, giving rise to a large number of species-specific gene families and genes [53, 57, 58]. The three *Chlamydomonas* species analyzed here had similar levels of gene duplication events and nearly identical gene gain/loss ratios at the point of delineation (Fig 6B and 6C). In addition, the three *Chlamydomonas* species all had lineage-specific genes, however, the total amount was <10% of the total gene pool (Fig 6A and 6B). The three *Botryococcus* species all had relatively large levels of lineage-specific genes identified (Fig 6A and 6B). The role these lineage-specific genes/gene families and their expansion play in eukaryotic organism evolution has been well established [97]. The vast amount of lineage-specific genes/gene families among the *Botryococcus* species has likely played a large role in the differences we have observed in the genomic space and supports our proposal of separate *Botryococcus* species.

It has been previously established that TE content and whole genome duplication (WGD) events are linked to the evolvability and complexity of an organism at the genome level in both maize and rice [98, 99]. While no WGDs have been observed, nor is the chromosome numbers known, for any of the three *Botryococcus* species, it is too early to speculate about the level of genome duplication that may have occurred in *Botryococcus*. Our data presented here including the phylogenomic placement, genome sizes, gene duplication events, and TE content of the three *Botryococcus* species suggests that some large level of genomic sequence duplication has occurred. However, this is speculatory and requires further investigation of the *Botryococcus* genomes including obtaining chromosome resolution assemblies and karyotyping each *Botryococcus* species.

Future directions

Additional genomic analyses within the *Botryococcus* genus are required to resolve several issues related to species and races. For example, phylogenetic analysis of the previously mentioned S race [9] using the 18S rRNA gene [9, 20] shows the S race is closely related to and may be a member of *B. lycopadienor* (i.e. the L race). A study published in 1992 suggested the existence of several different species of *Botryococcus* based solely on morphological differences in cell and colony shape and size [100]. However, since *Botryococcus* colony and cell morphology can be easily influenced by environmental factors causing observable differences within the same race/strain [101], these proposed species may not actually be separate species. For example, one of the proposed species, *B. terribilis* [100, 102] may actually be a strain of *B. braunii* (i.e. the B race) based on subsequent studies analyzing hydrocarbon content and phylogenetic placement [103, 104]. Additionally, phylogenetic studies [9, 20] have shown that *B. braunii* can be divided into two distinct subclades, suggesting *B. braunii* may consist of two separate species. Genomic sequencing, assembly, and comparative analysis on the S race, *B. terribilis*, members of each *B. braunii* subclade, and other proposed *Botryococcus* species is needed to resolve their positions within the *Botryococcus* genus.

Materials and methods

Culturing of *B. braunii*

Culture conditions for *B. alkenealis*, *B. braunii*, and *B. lycopadienor* were followed as described previously [18]. Briefly, *B. alkenealis*, (race A, Yamanaka strain [105]), *B. braunii* (race B, Showa strain [106]), and *B. lycopadienor (*race L, Songkla Nakarin strain [8]) were grown in modified Chu 13 medium under continuous aeration with filter sterilized air mixed at 2.5% CO_2. The composition of the modified Chu 13 medium was as follows: KNO_3 (0.4 g/L), $MgSO_4 \cdot 7H_2O$ (0.1 g/L), K_2HPO_4 (0.052 g/L), $CaCl_2 \cdot 2H_2O$ (0.054 g/L), FeNa EDTA (0.01 g/L), H_3BO_4 (2.86 mg/L), $MnSO_4 \cdot H_2O$ (1.54 mg/L), $ZnSO_4 \cdot 7H_2O$ (0.22 mg/L), $CuSO_4 \cdot 5H_2O$ (0.08 mg/L), $NaMoO_4 \cdot 2H_2O$ (0.06 mg/L), $CoSO_4 \cdot 7H_2O$ (0.09 mg/L). Cultures were grown under a light:dark cycle of 12:12h with a constant light intensity of 120 μE $m^{-2}s^{-1}$. Alga cells were subcultured by inoculation of 100 mL from a 6-week-old culture into 750mL of fresh medium.

DNA isolation and sequencing

Algal cells were harvested by vacuum filtration through a 5-μm or 10-μm nylon microsieve (BioDesign Inc. of New York) for *B. lycopadienor* and B. alkenealis/*B. braunii*, respectively. Cells were either used directly or snap-frozen in liquid nitrogen and stored at -80˚C until further use. High molecular weight (HMW) genomic DNA (gDNA) free of RNA and polysaccharides, was isolated from harvested biomass using a protocol we developed specifically for *Botryococcus* species [35] that was adapted from previous studies [107, 108]. Briefly, harvested algae biomass was macerated by mortar and pestle in the presence of liquid nitrogen to a fine powder. Then 100-110mg of macerated biomass was resuspended in 1 mL of sorbitol wash buffer (100 mM Tris-HCl pH 8.0, 0.35 M sorbitol, 5 mM EDTA pH 8.0, 1% (w/v) polyvinylpyrrolidone (PVP-40)), sonicated at 30% power for 25 seconds (strong enough to homogenize the sample, but not to lyse the cells), centrifuged at 2,500 x *g* for five minutes at room temperature and the supernatant discarded. Due to the hydrocarbons in the sample, the pellet floats on top of the supernatant. Care was taken to remove the supernatant without disturbing the floating pellet. The washed pellet was then resuspended by vortexing for 5 seconds in 700 μL of pre-warmed (65˚C), extraction buffer (100 mM Tris-HCl pH 8.0, 3 M NaCl, 3% CTAB-cetyl trimethylammonium bromide, 20 mM EDTA, 1% (w/v) Polyvinylpyrrolidone). Resuspended pellets were then incubated for 30 minutes in a 65˚C water bath, mixed by inversion every 10 minutes, and allowed to cool to room temperature for 5 minutes. Then 700 μL of 24:1 (v:v) chloroform: isoamyl alcohol (CIA) was added, mixed by vortexing for 10 seconds, and then centrifuged at 2,500 x *g* for 10 minutes at room temperature. The aqueous phase was removed and treated with Rnase A (0.1 mg/mL) at 37˚C for 15 minutes with mixing by inversion every 5 minutes, 500 μL of CIA added with mixing by inversion, centrifuged at 13,000 x *g* for 10 minutes at 4˚C, and the aqueous phase was recovered and transferred to a fresh tube. HMW gDNA was then precipitated by the addition of 1/10 volume of 3M sodium acetate (pH 5.2), and 2/3 volumes of cold (-20˚C) isopropanol, the sample was mixed by inverting 10 times, and incubated overnight at -20˚C. The samples were then centrifuged at 13,000 x *g* for 10 minutes at room temperature, the supernatant discarded, the pellet air-dried, and the pellets washed with 1 mL of 70% ethanol followed by centrifugation at 13,000 x *g* for 10 minutes at 4˚C. The HMW gDNA pellet was then dried under vacuum for 10 minutes, resuspended in 100 μL of nuclease-free water, and either stored at -80˚C or prepped for sequencing.

Library preparation and sequencing for Illumina-based sequencing was performed on the Illumina HiSeq2000 platform by the Texas A&M AgriLife Genomics and Bioinformatics Service (TxGen). Oxford Nanopore-based sequencing was performed by the Sequencing Technologies and Analysis core facility at the Cold Spring Harbor Laboratory. For this sequencing, DNA concentration was determined via fluorometry and small/fragmented gDNA (<25kb) was removed using the Circulomics short read eliminator kit (SS-100-101-01). Samples were barcoded and adaptors ligated with the native barcode kit (Oxford Nanopore Technologies, EXP-NBD104). Sequencing was performed on the Oxford Nanopore PromethION platform, using R9.4.1 flowcell biochemistry and base calling was done with ONT's proprietary Guppy (v4). In total, two flow cells were run due to complications with the quality of sequencing data on the first flow cell, the lower quality flow cell run will be referred to as flow cell 1 and the higher quality flow cell referred to as flow cell 2. Both flow cells' data were used in the final genome assemblies for *B*. *alkenealis* and *B*. *lycopadienor* given recovery of read quality after read trimming. Both sets of reads from each flow cell were subjected to the same pre-assembly trimming process, however, flow cell 1 had sufficiently fewer reads post-trim than flow cell 2 reflecting the overall lower quality of the reads obtained.

De novo genome assembly for *B*. *lycopadienor* (L race)

The raw fastq files from the Nanopore sequences were trimmed with Porechop (v0.2.4) to remove any remaining non-gDNA sequences and poor-quality bases. The L race genome assembly was performed using the Flye long-read assembler [109] with the trimmed long-read sequences as input. The resulting scaffolds and contigs were then polished with the short-read Illumina DNA sequencing data that was trimmed for barcode, adapter, and poor quality sequence using Trimmomatic [110]. Scaffold polishing was done iteratively twice with Pilon [111], with the output of the first round of polishing used as the input for the next round of polishing. Scaffolds/contigs belonging to symbiotic or organelle DNA [112] were identified and removed using Blobtools (v1.1) [113]. Briefly, scaffolds/contigs that displayed GC% and k-mer frequency and were annotated taxonomically as non-green microalgae were individually reviewed for confirmation before removal. The nuclear draft genome was then masked using a combination of RepeatModeler (v1.0) [46] and RepeatMasker (v4.0) [47]. First, a *de novo* repeat library was assembled using RepeatModeler. The repeat library was then used to mask the genome with RepeatMasker using *B*. *braunii* B race CDS

[34] (JGI: Bbraunii_502_v2.1.cds.fa) as trusted sequences to prevent over-masking by RepeatMasker. Finally, assembly metrics and quality were assessed using Quast [114] and BUSCO v5 [115], respectively. ChlorophytaODBv10 [50] was used as the ortholog database for BUSCO analysis.

Hybrid *de novo* genome assembly for *B. alkenealis* (A race)

The genome assembly pipeline used for *B. lycopadienor* was applied to *B. alkenealis* but did not result in a high-quality assembly. Thus, a modified approach was taken to produce a higher quality assembly for *B. alkenealis* to maximize the final N50, minimize the final L50, and reduce the number of "N"'s in the unmasked assembly. Trimming of both the Nanopore and Illumina sequencing data along with contig/scaffold assembly of the long read sequences by Flye remained the same as was done for *B. lycopadienor*. The hybrid assembler MaSuRCA (v4.0.1) [116] was used to generate a hybrid genome assembly from both the long and short read sequences as input. Then both the Flye assembly and MaSuRCA assembly were merged and de-replicated using Quickmerge [117]. The resulting merged assembly was then polished twice iteratively with the short-read sequence using Pilon. Removal of non-nuclear DNA, masking, and quality/metric assessment was done as described for the *B. lycopadienor*.

Attempts at genome assembly for the *B. braunii* (B race)

Attempts were made to improve the current *B. braunii* (B race) genome assembly [34]. The newly generated Nanopore sequencing data were used in identical workflows that were used for *B. lycopadienor* and *B. alkenealis* genome assemblies mentioned above. Unfortunately, neither resulted in an improved assembly. While N50 values were slightly higher than the current metrics, both contig number and L50 values were larger than their respective values in the current *B. braunii* (B race) genome assembly. BUSCO analysis also showed a higher amount of fragmented orthologs than the current assembly. An additional 10 assembly approaches were tested including a reference-guided genome assembly approach [118] using the current assembly as "trusted" contigs for inputs into MaSuRCA, Flye, and Canu assemblers. None of these advanced assembly tactics yielded an improved genome assembly for *B. braunii*. Likely, the Nanopore sequencing data did not contain enough sequencing depth to resolve the complexities of the *B. braunii* genome [34]. Future attempts at genome sequencing and assembly into *B. braunii* will need more long-read sequencing data and optical mapping to resolve complex, repeat-rich portions of the genome.

RNA isolation and sequencing

B. alkenealis and *B. lycopadienor* cultures were grown and biomass was harvested as described above. Total RNA was isolated and purified using the TRIzol Reagent (Thermofisher). The resulting RNA pellet was washed with 75% ethanol and then dried using a speed vac. To remove the polysaccharides from the samples, the RNA pellet was resuspended in 2M LiCl and centrifuged at 10,000 x *g* at room temperature for 10 minutes. The supernatant containing polysaccharides was discarded and the pellet retained. This process was repeated until the RNA pellet size remained the same after each wash step. The washed pellet was then resuspended in 1x TE buffer and extracted against and an equal volume of 1:1:1 phenol/chloroform/isoamyl alcohol. The suspension was then centrifuged at 10,000 x *g* at room temperature for 10 minutes with the pellet discarded and the supernatant saved. A final extraction against an equal volume of chloroform was done followed by centrifugation at 10,000 x *g* and the supernatant saved. Finally, the RNA was precipitated with 0.1 volumes of 3M sodium acetate and 2.5 volumes of 100% ethanol. The solution was centrifuged at 10,000 x *g* for 10 minutes at room temperature and the pellet was saved. The RNA pellet was resuspended in Rnase-free water and treated with Dnase (1 unit/μg RNA), incubated at 30˚C for 30 min, treated with 1:1:1 phenol/chloroform/isoamyl alcohol, precipitated with sodium acetate and ethanol as described earlier, the DNA-free final RNA pellet was resuspended in nuclease-free water, and stored at -80˚C until further use. RNA pellets were sent for sequencing through the National Alliance for Advanced Biofuels and Bioproducts (NAABB) to the Los Alamos National Laboratory for sequencing on the Illumina platform [119, 120].

Genome-guided transcript assembly for *B. alkenealis* and *B. lycopadienor*

RNA-seq data were trimmed of adapters and low-quality bases using Trimmomatic. Trimmed reads were then mapped to the corresponding genome assemblies using HISAT2 [121] with default settings. The resulting BAM file was indexed and fed as input for use in the Trinity suite [122] run in genome-guided mode. The Trinity raw assembled transcripts were then filtered for low-expression isoforms using Trinity's built-in scripts, and the removal of redundant transcripts was done with CD-HIT-EST [123]. Final draft transcriptomes were assessed for quality and metrics by Quast and BUSCO, respectively, run in transcript mode. ChlorophytaODBv10 was used as the ortholog database for BUSCO analysis.

Annotation of genome features

Genes were predicted and annotated using the BRAKER2 pipeline [124–133]. Briefly, BRAKER2 was run in "evidence-based" mode where RNA-seq data was first mapped to a hard-masked version of the genome assemblies using HISAT2 [121]. The resulting BAM file was used as input into BRAKER2 with default parameters and the output was saved in GTF format for all three Botryococcus species. BRAKER2 was also used for the untranslated region (UTR) analysis using the GUSHR module. At the time of analysis, BRAKER2 did not support automatic intronic region annotation. So, intronic sequences were detected and annotated using a custom python script with the BRAKER2 GTF output files. Due to BRKAER2's increased capability of fully automatic training of the gene prediction toolsets and leveraging RNA-seq and protein homology information into the final gene prediction sets, the decision was made to re-annotate the *B. braunii* (B race) v2 genome assembly, using this increased capability that was not available at the time of the original release of the v2 assembly. This re-annotation gave a different final number of genes when compared to the v2 genome assembly files. This new annotation predicted 23,685 genes while the v2 genome annotation predicted 20,765 genes as seen in Fig 1. Transposable element content was annotated using RepeatModeler [46], and RepeatMasker [47]. DNA tandem repeats (satellites) were annotated by Tandem Repeats Finder [48] using default parameters. In all three *Botryococcus* species, genes were scanned for coding sequence overlaps with TEs at an overlap threshold of ≥70%. This was achieved by converting the RepeatMasker output file and BRAKER2 GFF/GTF output files to the BED format. BEDTools [134] was then used to identify overlapping regions using the "intersect" program with the "-f 0.70" flag.

Phylogenomic analysis

The approach for a mass phylogenomic analysis using multiple microalgae species was modified slightly from a previous study [91]. Briefly, protein sequence files for each species analyzed in this study (*B. alkenealis*, *B. braunii*, *B. lycapdienor*, *Asterchloris glomerata*, *Auxenochlorella protothecoides*, *Chlamydomonas incerta*, *Chlamydomonas reinhardtii*, *Chlamydomonas schloesseri*, *Chlorella sorokiniana*, *Chlorella variabilis*, *Coccomyxa subellipsoidea*, *Dunaliella salina*, *Micractinium conductrix*, *Picochlorum renovo*, *Picochlorum soloecismus*, *Trebouxia sp. A1-2*, *Volvox carteri*) were screened by BUSCO in protein mode using the ChlorophytaODBv10 database [50]. Single-copy orthologous proteins that were conserved in at least 13 of the 17 species were extracted, counted, and analyzed. Multiple sequence alignments (MSAs) of each ortholog set were produced using MAFFT v7 [135]. Maximum likelihood trees were inferred for each ortholog set from the corresponding MSAs using IQTREE [136] run with the following parameters: '-m MFP -bb 1000 -T 5'. The final set of 1,189 ortholog trees was reconciled into a single alternative species tree using ASTRAL-III [137]. The final phylogenomic tree was rooted in the Chlorophyceae sub-clad*e*.

Comparative genomic analysis

Protein sequence files for *B*. *alkenealis*, *B*. *braunii*, and *B*. *lycopadienor* were generated using BRAKER2 as described above. The remaining 14 species (*A*. *glomerata*, *A*. *protothecoides*, *C*. *incerta*, *C*. *reinhardtii*, *C*. *schloesseri*, *C*. *sorokiniana*, *C*. *variabilis*, *C*. *subellipsoidea*, *D*. *salina*, *M*. *conductrix*, *P*. *renovo*, *P*. *soloecismus*, *Trebouxia sp*. *A1-2*, *V*. *carteri*) used in this study were downloaded from JGI's PhycoCosm [91, 138–143]. Comparative genomic analysis was initiated by first inferring gene family relationships within and between the different species using OrthoFinder [51].

To visualize the distribution of gene families among the 17 species, a custom script was written to parse the assigned orthogroups from OrthoFinder and visualize the number and intersection of orthogroups by generating an UpSet plot. Similarly, the same script also generated weighted Venn diagrams for both the three *Botryococcus* species and the three *Chlamydomonas* species.

Gene gain and gene loss analysis was performed as previously described [61] with slight modifications. Briefly, the gene trees for each orthogroup from the OrthoFinder analysis were reconciled using Notung [60] to determine gene loss and gene gain over an evolutionary period. A custom script was written to extract these losses and gains at each node of the reconciled tree and mapped to the gene duplication species tree from the OrthoFinder analysis. The total number of organism gene duplication events was also obtained from the gene duplication specie trees.

Identification of hydrocarbon genes within each *Botryococcus* species

Fatty acid elongase gene analysis was performed by by first obtaining FAE (Pfam: PF08392) and ELO (Pfam: PF01151) hidden-markov models (HMM) from Pfam [144, 145]. These HMM profiles were then used to search for homologous sequences using the HMMER suite [146] with an E-value cutoff of 10^{-3}. Sequences that had significant homology to the FAE and ELO profiles in each *Botryococcus* species were then aligned using MAFFT and phylogeny was inferred with IQTREE using the same parameters mentioned above. KCR and HCD genes do not currently have a Pfam domain entry. Instead, single-copy protein sequences from *A*. *thaliana* were obtained from NCBI (NP_564905.1, and NP_193180.1 respectively) and used as query sequences in a blastp analysis of the *Botryococcus* species' proteomes. An e-value cutoff of 10^{-3} was used. The Pfam ECR hmm domain (Pfam: PF12241) was used in a similar approach as the FAE and ELO searches. Due to the single-copy nature of KCR, HCD, and ECR genes, phylogenetic trees were not inferred. Squalene synthase (SS) and squalene synthase-like (SSL) gene analysis in the three *Botryococcus* species was performed in the same manner as the fatty acid elongases using the SQS-PYS HMM profile (Pfam: PF00494).

Identification of meiosis genes within each *Botryococcus* species

Potential meiosis genes were identified based on sequence homology to meiosis genes identified in *C*. *subellipsoidea* and *C*. *variabilis* [79]. Briefly, DMC1, HOP1, HOP2, MER3, MND1, MSH4, MSH5, REC8, and SPO11 protein targets were identified in the BRAKER2 protein annotation sets by protein-protein BLAST+ homology alignment. Targets identified by alignment over an E-value threshold of 0.001, were then searched for Pfam domains using InterPro [144] against the Pfam database. In both homology searches, only the top alignment (smallest E-value) is reported. *C. subellipsoidea* protein sequences used in the homology searches were HOP1 (XP_005651810), HOP2 (XP_005643862), MER3 (XP_005651102, XP_005649357, XP_005649238), MND1 (XP_005649122), MSH4 (XP_005646103), MSH5 (XP_005644118), REC8 (XP_005651890), and SPO11 (XP_005647003). The *C. variabilis* meiosis gene used in the homology searches was DMC1 (XP_005848077).

Supporting information

S1 File.
(PDF)

Acknowledgments

Illumina-based sequencing was performed by the Texas A&M AgriLife Genomics and Bioinformatics Service (TxGen) as a fee-for-service. Oxford Nanopore-based sequencing was performed by Cold Spring Harbor Laboratories as a fee-for-service. Portions of this research were conducted with the advanced computing resources provided by Texas A&M High Performance Research Computing. We would like to thank Luke Friedrich (CSE Software) for insight and advice on custom scripts.

Author Contributions

Conceptualization: Devon J. Boland, Timothy P. Devarenne.

Data curation: Devon J. Boland.

Formal analysis: Devon J. Boland.

Funding acquisition: Timothy P. Devarenne.

Investigation: Devon J. Boland, Daniel R. Browne, Rebecca L. Murphy.

Methodology: Devon J. Boland, Ivette Cornejo-Corona, Daniel R. Browne.

Project administration: John Mullet, Shigeru Okada, Timothy P. Devarenne.

Resources: Timothy P. Devarenne.

Supervision: John Mullet, Shigeru Okada, Timothy P. Devarenne.

Visualization: Devon J. Boland.

Writing – original draft: Devon J. Boland, Rebecca L. Murphy.

Writing – review & editing: Devon J. Boland, Ivette Cornejo-Corona, Daniel R. Browne, John Mullet, Shigeru Okada, Timothy P. Devarenne.

References

1. Banerjee A, Sharma R, Chisti Y, Banerjee UC. *Botryococcus braunii*: A renewable source of hydrocarbons and other chemicals. Critical Reviews in Biotechnology. 2002; 22(3):245–79. https://doi.org/10.1080/07388550290789513 PMID: 12405558
2. Metzger P, Largeau C. *Botryococcus braunii*: a rich source for hydrocarbons and related ether lipids. Appl Microbiol Biotechnol. 2005; 66(5):486–96. Epub 20041204. https://doi.org/10.1007/s00253-004-1779-z PMID: 15630516.

3. Hillen LW, Pollard G, Wake LV, White N. Hydrocracking of the Oils of *Botryococcus braunii* to Transport Fuels. Biotechnology and Bioengineering. 1982; 24:193–205. https://doi.org/10.1002/bit.260240116 PMID: 18546110

4. Zhang K, Zhang X, Tan TW. The production of bio-jet fuel from *Botryococcus braunii* liquid over a Ru/CeO2 catalyst. Rsc Adv. 2016; 6(102):99842–50. https://doi.org/10.1039/c6ra22517a WOS:000386439800028.

5. van Dyk S, Su J, McMillan JD, Saddler J. Potential synergies of drop-in biofuel production with further co-processing at oil refineries. Biofuels, Bioproducts and Biorefining. 2019; 13(3):760–75. https://doi.org/10.1002/bbb.1974

6. Brown AC, Knights BA, Conway E. Hydrocarbon content and its relationship to physiological state in the green alga *Botryococcus braunii*. Phytochemistry. 1969; 8(3):543–7. https://doi.org/10.1016/S0031-9422(00)85397-2

7. Gelpi E, Schneider H, Mann J, Oró J. Hydrocarbons of Geochemical Significance in Microscopic Algae. Phytochemistry. 1970; 9(3):603–12. https://doi.org/10.1016/S0031-9422(00)85700-3

8. Metzger P, Casadevall E. Lycopadiene, a tetraterpenoid hydrocarbon from new strains of the green alga *Botryococcus braunii*. Tetrahedron Letters. 1987; 28(34):3931–4. https://doi.org/10.1016/S0040-4039(00)96423-2

9. Kawachi M, Tanoi T, Demura M, Kaya K, Watanabe MM. Relationship between hydrocarbons and molecular phylogeny of *Botryococcus braunii*. Algal Research. 2012; 1(2):114–9. https://doi.org/10.1016/j.algal.2012.05.003

10. Largeau C, Casadevall E, Berkaloff C. The biosynthesis of long-chain hydrocarbons in the green alga *Botryococcus braunii*. Phytochemistry. 1980; 19(6):1081–5. ISI:A1980KA49600013.

11. Templier J, Largeau C, Casadevall E. Mechanism of non-isoprenoid hydrocarbon biosynthesis in *Botryococcus braunii*. Phytochemistry. 1984; 23(5):1017–28. https://doi.org/10.1016/S0031-9422(00)82602-3.

12. Metzger P, Casadevall E, Coute A. Botryococcene distribution in strains of the green alga *Botryococcus braunii*. Phytochemistry. 1988; 27(5):1383–8. https://doi.org/10.1016/0031-9422(88)80199-7

13. Templier J, Largeau C, Casadevall E. Biosynthesis of n-alkatrienes in *Botryococcus braunii*. Phytochemistry. 1991; 30(7):2209–15. https://doi.org/10.1016/0031-9422(91)83616-S

14. Metzger P, Rager M-N, Largeau C. Polyacetals based on polymethylsqualene diols, precursors of algaenan in *Botryococcus braunii* race B. Organic Geochemistry. 2007; 38:566–81. https://doi.org/10.1016/j.orggeochem.2006.12.003

15. Metzger P, Rager MN, Fosse C. Braunicetals: acetals from condensation of macrocyclic aldehydes and terpene diols in *Botryococcus braunii*. Phytochemistry. 2008; 69(12):2380–6. Epub 20080717. https://doi.org/10.1016/j.phytochem.2008.06.004 PMID: 18639308.

16. Weiss TL, Spencer Johnston J, Fujisawa K, Sumimoto K, Okada S, Chappell J, Devarenne TP. Phylogenetic Placement, Genome Size, and Gc Content of the Liquid-Hydrocarbon-Producing Green Microalga *Botryococcus braunii* Strain Berkeley (Showa) (Chlorophyta). Journal of Phycology. 2010; 46 (3):534–40. https://doi.org/10.1111/j.1529-8817.2010.00820.x

17. Weiss TL, Johnston JS, Fujisawa K, Okada S, Devarenne TP. Genome size and phylogenetic analysis of the A and L races of *Botryococcus braunii*. J Appl Phycol. 2011; 23(5):833–9. https://doi.org/10.1007/s10811-010-9586-7 WOS:000295822300005.

18. Thapa HR, Naik MT, Okada S, Takada K, Molnar I, Xu Y, Devarenne TP. A squalene synthase-like enzyme initiates production of tetraterpenoid hydrocarbons in *Botryococcus braunii* Race L. Nat Commun. 2016; 7(1):11198. Epub 20160406. https://doi.org/10.1038/ncomms11198 PMID: 27050299; PubMed Central PMCID: PMC4823828.

19. Tatli M, Ishihara M, Heiss C, Browne DR, Dangott LJ, Vitha S, et al. Polysaccharide associated protein (PSAP) from the green microalga *Botryococcus braunii* is a unique extracellular matrix hydroxyproline-rich glycoprotein. Algal Res. 2018; 29:92–103. https://doi.org/10.1016/j.algal.2017.11.018 WOS:000425552500010.

20. Hirano K, Hara T, Ardianor Nugroho RA, Segah H, Takayama N, et al. Detection of the oil-producing microalga Botryococcus braunii in natural freshwater environments by targeting the hydrocarbon biosynthesis gene SSL-3. Scientific reports. 2019; 9(1):16974. Epub 20191118. https://doi.org/10.1038/s41598-019-53619-y PMID: 31740707; PubMed Central PMCID: PMC6861321.

21. Heiss C, Black I, Ishihara M, Tatli M, Devarenne TP, Azadi P. Structure of the polysaccharide sheath from the B race of the green microalga *Botryococcus braunii*. Algal Res. 2021; 55:102252. https://doi.org/10.1016/j.algal.2021.102252 WOS:000642455500015.

22. Okada S, Devarenne TP, Murakami M, Abe H, Chappell J. Characterization of botryococcene synthase enzyme activity, a squalene synthase-like activity from the green microalga *Botryococcus* *braunii*, Race B. Arch Biochem Biophys. 2004; 422(1):110–8. https://doi.org/10.1016/j.abb.2003.12.004 PMID: 14725863.

23. Kawamura K, Nishikawa S, Hirano K, Ardianor A, Nugroho RA. BoCAPS: Rapid screening of chemical races in *Botryococcus braunii* with direct PCR-CAPS. Algal Research. 2022; 66:102789. https://doi.org/10.1016/j.algal.2022.102789.

24. Mallet J. A species definition for the modern synthesis. Trends in Ecology & Evolution. 1995; 10 (7):294–9. https://doi.org/10.1016/0169-5347(95)90031-4 PMID: 21237047

25. Harada A, Yamagishi T. Meiosis in *Spirogyra* (Chlorophyceae). Japanese Journal Phycology. 1984; 32:10–8.

26. Ramawat KG, Merillon J-M, Shivanna KR. Reproductive Biology of Plants. Boca Raton: CRC Press; 2014.

27. Ferris PJ, Goodenough UW. The mating-type locus of *Chlamydomonas reinhardtii* contains highly rearranged DNA sequences. Cell. 1994; 76(6):1135–45. https://doi.org/doi.org/10.1016/0092-8674(94)90389-1

28. Ferris PJ, Woessner JP, Goodenough UW. A sex recognition glycoprotein is encoded by the plus mating-type gene fus1 of *Chlamydomonas reinhardtii*. Molecular Biology of the Cell. 1996; 7(8):1235–48. https://doi.org/10.1091/mbc.7.8.1235 PMID: 8856667.

29. Ferris PJ, Pavlovic C, Fabry S, Goodenough UW. Rapid evolution of sex-related genes in *Chlamydomonas*. Proc Natl Acad Sci U S A. 1997; 94(16):8634–9. https://doi.org/10.1073/pnas.94.16.8634 PMID: 9238029.

30. Ferris P, Olson BJ, De Hoff PL, Douglass S, Casero D, Prochnik S, et al. Evolution of an expanded sex-determining locus in *Volvox*. Science. 2010; 328(5976):351–4. https://doi.org/10.1126/science.1186222 PMID: 20395508; PubMed Central PMCID: PMC2880461.

31. Funk ER, Mason NA, Palsson S, Albrecht T, Johnson JA, Taylor SA. A supergene underlies linked variation in color and morphology in a *Holarctic* songbird. Nat Commun. 2021; 12(1):6833. Epub 20211125. https://doi.org/10.1038/s41467-021-27173-z PMID: 34824228; PubMed Central PMCID: PMC8616904.

32. Caputo A, Dubourg G, Croce O, Gupta S, Robert C, Papazian L, et al. Whole-genome assembly of *Akkermansia muciniphila* sequenced directly from human stool. Biol Direct. 2015; 10(1):5. Epub 20150219. https://doi.org/10.1186/s13062-015-0041-1 PMID: 25888298; PubMed Central PMCID: PMC4333879.

33. Caputo A, Lagier JC, Azza S, Robert C, Mouelhi D, Fournier PE, Raoult D. *Microvirga massiliensis* sp. nov., the human commensal with the largest genome. Microbiologyopen. 2016; 5(2):307–22. Epub 20160108. https://doi.org/10.1002/mbo3.329 PMID: 26749561; PubMed Central PMCID: PMC4831475.

34. Browne DR, Jenkins J, Schmutz J, Shu S, Barry K, Grimwood J, et al. Draft Nuclear Genome Sequence of the Liquid Hydrocarbon-Accumulating Green Microalga *Botryococcus braunii* Race B (Showa). Genome Announc. 2017; 5(16):e00215–17. Epub 20170420. https://doi.org/10.1128/genomeA.00215-17 PMID: 28428306; PubMed Central PMCID: PMC5399265.

35. Cornejo-Corona I, Boland DJ, Devarenne TP. Method for isolation of high molecular weight genomic DNA from *Botryococcus* biomass. In Press, PLoS ONE. 2024.

36. Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, Witman GB, et al. The *Chlamydomonas* Genome Reveals the Evolution of Key Animal and Plant Functions. Science. 2007; 318 (5848):245–50. https://doi.org/10.1126/science.1143609 PMID: 17932292

37. Blanc G, Agarkova I, Grimwood J, Kuo A, Brueggeman A, Dunigan DD, et al. The genome of the polar eukaryotic microalga *Coccomyxa subellipsoidea* reveals traits of cold adaptation. Genome Biol. 2012; 13(5):R39. Epub 20120525. https://doi.org/10.1186/gb-2012-13-5-r39 PMID: 22630137; PubMed Central PMCID: PMC3446292.

38. Blanc G, Duncan G, Agarkova I, Borodovsky M, Gurnon J, Kuo A, et al. The *Chlorella variabilis* NC64A genome reveals adaptation to photosymbiosis, coevolution with viruses, and cryptic sex. Plant Cell. 2010; 22(9):2943–55. Epub 20100917. https://doi.org/10.1105/tpc.110.076406 PMID: 20852019; PubMed Central PMCID: PMC2965543.

39. Grigoriev IV, Hayes RD, Calhoun S, Kamel B, Wang A, Ahrendt S, et al. PhycoCosm, a comparative algal genomics resource. Nucleic Acids Res. 2021; 49(D1):D1004–D11. https://doi.org/10.1093/nar/gkaa898 PMID: 33104790; PubMed Central PMCID: PMC7779022.

40. Polle JEW, Barry K, Cushman J, Schmutz J, Tran D, Hathwaik LT, et al. Draft Nuclear Genome Sequence of the Halophilic and Beta-Carotene-Accumulating Green Alga *Dunaliella salina* Strain CCAP19/18. Genome Announc. 2017; 5(43). Epub 20171026. https://doi.org/10.1128/genomeA.01105-17 PMID: 29074648; PubMed Central PMCID: PMC5658486.

41. Dahlin LR, Gerritsen AT, Henard CA, Van Wychen S, Linger JG, Kunde Y, et al. Development of a high-productivity, halophilic, thermotolerant microalga *Picochlorum renovo*. Commun Biol. 2019;2:388-. https://doi.org/10.1038/s42003-019-0620-2 PMID: 31667362.

42. Foflonker F, Price DC, Qiu H, Palenik B, Wang S, Bhattacharya D. Genome of the halotolerant green alga *Picochlorum* sp. reveals strategies for thriving under fluctuating environmental conditions. Environ Microbiol. 2015; 17(2):412–26. Epub 20140724. https://doi.org/10.1111/1462-2920.12541 PMID: 24965277.

43. Foflonker F, Ananyev G, Qiu H, Morrison A, Palenik B, Dismukes GC, Bhattacharya D. The unexpected extremophile: Tolerance to fluctuating salinity in the green alga *Picochlorum*. Algal Research. 2016; 16:465–72. https://doi.org/10.1016/j.algal.2016.04.003.

44. Foflonker F, Mollegard D, Ong M, Yoon HS, Bhattacharya D. Genomic Analysis of *Picochlorum* Species Reveals How Microalgae May Adapt to Variable Environments. Molecular Biology and Evolution. 2018; 35(11):2702–11. https://doi.org/10.1093/molbev/msy167 PMID: 30184126

45. Weissman JC, Likhogrud M, Thomas DC, Fang W, Karns DAJ, Chung JW, et al. High-light selection produces a fast-growing *Picochlorum celeri*. Algal Research. 2018; 36:17–28. https://doi.org/10.1016/j.algal.2018.09.024

46. Smit A, Hubley R. RepeatModeler Open-1.0. 2008-2015.

47. Smit A, Hubley R, Green P. RepeatMasker Open-4.0. 2013-2015.

48. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999; 27(2):573–80. https://doi.org/10.1093/nar/27.2.573 PMID: 9862982; PubMed Central PMCID: PMC148217.

49. Olmstead RG, Sweere JA. Combining Data in Phylogenetic Systematics: An Empirical Approach Using Three Molecular Data Sets in the Solanaceae. Systematic Biology. 1994; 43(4):467–81. https://doi.org/10.1093/sysbio/43.4.467

50. Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simao FA, Zdobnov EM. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Res. 2019; 47(D1):D807–D11. https://doi.org/10.1093/nar/gky1053 PMID: 30395283; PubMed Central PMCID: PMC6323947.

51. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biol. 2015; 16(1):157. Epub 20150806. https://doi.org/10.1186/s13059-015-0721-2 PMID: 26243257; PubMed Central PMCID: PMC4531804.

52. Lex A, Gehlenborg N, Strobelt H, Vuillemot R, Pfister H. UpSet: Visualization of Intersecting Sets. IEEE Transactions on Visualization and Computer Graphics. 2014; 20(12):1983–92. https://doi.org/10.1109/TVCG.2014.2346248 PMID: 26356912

53. Luna SK, Chain FJJ. Lineage-Specific Genes and Family Expansions in Dictyostelid Genomes Display Expression Bias and Evolutionary Diversification during Development. Genes (Basel). 2021; 12(10):1628. Epub 20211016. https://doi.org/10.3390/genes12101628 PMID: 34681022; PubMed Central PMCID: PMC8535579.

54. Fitch WM. Distinguishing Homologous from Analogous Proteins. Systematic Biology. 1970; 19(2):99–113. https://doi.org/10.2307/2412448 PMID: 5449325

55. Fitch WM. Homology: a personal view on some of the problems. Trends in Genetics. 2000; 16(5):227–31. https://doi.org/10.1016/s0168-9525(00)02005-9 PMID: 10782117

56. Leebens-Mack JH, Barker MS, Carpenter EJ, Deyholos MK, Gitzendanner MA, Graham SW, et al. One thousand plant transcriptomes and the phylogenomics of green plants. Nature. 2019; 574(7780):679–85. https://doi.org/10.1038/s41586-019-1693-2 PMID: 31645766

57. Conant GC, Wolfe KH. Turning a hobby into a job: How duplicated genes find new functions. Nature Reviews Genetics. 2008; 9(12):938–50. https://doi.org/10.1038/nrg2482 PMID: 19015656

58. Freeling M, Scanlon MJ, Fowler JE. Fractionation and subfunctionalization following genome duplications: mechanisms that drive gene content and their consequences. Current Opinion in Genetics & Development. 2015; 35:110–8. https://doi.org/10.1016/j.gde.2015.11.002 PMID: 26657818

59. Soltis PS, Marchant DB, Van de Peer Y, Soltis DE. Polyploidy and genome evolution in plants. Current Opinion in Genetics & Development. 2015; 35:119–25. https://doi.org/10.1016/j.gde.2015.11.003 PMID: 26656231

60. Stolzer M, Lai H, Xu M, Sathaye D, Vernot B, Durand D. Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees. Bioinformatics. 2012; 28(18):i409–i15. https://doi.org/10.1093/bioinformatics/bts386 PMID: 22962460

61. Wu G, Hufnagel DE, Denton AK, Shiu SH. Retained duplicate genes in green alga *Chlamydomonas reinhardtii* tend to be stress responsive and experience frequent response gains. BMC Genomics.

2015; 16(1):149. Epub 20150304. https://doi.org/10.1186/s12864-015-1335-5 PMID: 25880851; PubMed Central PMCID: PMC4364661.

62. Metzger P, Berkaloff C, Casadevall E, Coute A. Alkadiene- and botryococcene-producing races of wild strains of *Botryococcus braunii*. Phytochemistry. 1985; 24(10):2305–12. https://doi.org/10.1016/S0031-9422(00)83032-0

63. Haslam TM, Kunst L. Extending the story of very-long-chain fatty acid elongation. Plant science: an international journal of experimental plant biology. 2013; 210:93–107. Epub 2013/07/16. https://doi.org/10.1016/j.plantsci.2013.05.008 PMID: 23849117.

64. Fehling E, Mukherjee KD. Acyl-CoA elongase from a higher plant (*Lunaria annua*): metabolic intermediates of very-long-chain acyl-CoA products and substrate specificity. Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism. 1991; 1082(3):239–46. https://doi.org/10.1016/0005-2760(91)90198-q PMID: 2029543

65. Toke DA, Martin CE. Isolation and characterization of a gene affecting fatty acid elongation in *Saccharomyces cerevisiae*. J Biol Chem. 1996; 271(31):18413–22. https://doi.org/10.1074/jbc.271.31.18413 PMID: 8702485.

66. Joubes J, Raffaele S, Bourdenx B, Garcia C, Laroche-Traineau J, Moreau P, et al. The VLCFA elongase gene family in *Arabidopsis thaliana*: phylogenetic analysis, 3D modelling and expression profiling. Plant Mol Biol. 2008; 67(5):547–66. Epub 20080509. https://doi.org/10.1007/s11103-008-9339-z PMID: 18465198.

67. Millar AA, Kunst L. Very-long-chain fatty acid biosynthesis is controlled through the expression and specificity of the condensing enzyme. Plant J. 1997; 12(1):121–31. https://doi.org/10.1046/j.1365-313x.1997.12010121.x PMID: 9263455.

68. Jakobsson A, Westerberg R, Jacobsson A. Fatty acid elongases in mammals: their regulation and roles in metabolism. Prog Lipid Res. 2006; 45(3):237–49. Epub 20060306. https://doi.org/10.1016/j.plipres.2006.01.004 PMID: 16564093.

69. Robinson GW, Tsay YH, Kienzle BK, Smith-Monroy CA, Bishop RW. Conservation between human and fungal squalene synthetases: similarities in structure, function, and regulation. Mol Cell Biol. 1993; 13(5):2706–17. https://doi.org/10.1128/mcb.13.5.2706-2717.1993 PMID: 8474436; PubMed Central PMCID: PMC359645.

70. Zhang D, Jennings SM, Robinson GW, Poulter CD. Yeast squalene synthase: expression, purification, and characterization of soluble recombinant enzyme. Arch Biochem Biophys. 1993; 304(1):133–43. https://doi.org/10.1006/abbi.1993.1331 PMID: 8323279.

71. Devarenne TP, Shin DH, Back K, Yin S, Chappell J. Molecular Characterization of Tobacco Squalene Synthase and Regulation in Response to Fungal Elicitor. Archives of Biochemistry and Biophysics. 1998; 349(2):205–15. https://doi.org/10.1006/abbi.1997.0463 PMID: 9448707

72. Hayashi H, Hirota A, Hiraoka N, Ikeshiro Y. Molecular cloning and characterization of two cDNAs for *Glycyrrhiza glabra* squalene synthase. Biol Pharm Bull. 1999; 22(9):947–50. https://doi.org/10.1248/bpb.22.947 PMID: 10513618.

73. Busquets A, Keim V, Closa M, del Arco A, Boronat A, Arro M, Ferrer A. *Arabidopsis thaliana* contains a single gene encoding squalene synthase. Plant Mol Biol. 2008; 67(1-2):25–36. Epub 2008/02/01. https://doi.org/10.1007/s11103-008-9299-3 PMID: 18236008.

74. Nguyen HT, Neelakadan AK, Quach TN, Valliyodan B, Kumar R, Zhang Z, Nguyen HT. Molecular characterization of *Glycine max* squalene synthase genes in seed phytosterol biosynthesis. Plant Physiol Biochem. 2013; 73:23–32. Epub 20130819. https://doi.org/10.1016/j.plaphy.2013.07.018 PMID: 24036394.

75. Unland K, Pütter KM, Vorwerk K, van Deenen N, Twyman RM, Prüfer D, Schulze Gronover C. Functional characterization of squalene synthase and squalene epoxidase in *Taraxacum koksaghyz*. Plant Direct. 2018; 2(6):e00063. https://doi.org/10.1002/pld3.63 PMID: 31245726

76. Dolence JM, Poulter CD. 5.13 - Electrophilic Alkylations, Isomerizations, and Rearrangements. In: Barton SD, Nakanishi K, Meth-Cohn O, editors. Comprehensive Natural Products Chemistry. Oxford: Pergamon; 1999. p. 315–41.

77. Okada S, Devarenne TP, Chappell J. Molecular Characterization of Squalene Synthase from the Green Microalga *Botryococcus braunii*, Race B. Archives of Biochemistry and Biophysics. 2000; 373(2):307–17. https://doi.org/10.1006/abbi.1999.1568

78. Niehaus TD, Okada S, Devarenne TP, Watt DS, Sviripa V, Chappell J. Identification of unique mechanisms for triterpene biosynthesis in *Botryococcus braunii*. Proc Natl Acad Sci U S A. 2011; 108(30):12260–5. Epub 20110711. https://doi.org/10.1073/pnas.1106222108 PMID: 21746901; PubMed Central PMCID: PMC3145686.

79. Fučíková K, Pažoutová M, Rindi F. Meiotic genes and sexual reproduction in the green algal class Trebouxiophyceae (Chlorophyta). Journal of Phycology. 2015; 51(3):419–30. https://doi.org/10.1111/jpy.12293 PMID: 26986659

80. Molles MC. Ecology: Concepts and Applications: McGraw-Hill; 2005.

81. Hardtke CS, Müller J, Berleth T. Genetic similarity among *Arabidopsis thaliana* ecotypes estimated by DNA sequence comparison. Plant Mol Biol. 1996; 32(5):915–22. https://doi.org/10.1007/bf00020488 PMID: 8980542.

82. Town CD, Cheung F, Maiti R, Crabtree J, Haas BJ, Wortman JR, et al. Comparative genomics of Brassica oleracea and *Arabidopsis thaliana* reveal gene loss, fragmentation, and dispersal after polyploidy. Plant Cell. 2006; 18(6):1348–59. Epub 20060421. https://doi.org/10.1105/tpc.106.041665 PMID: 16632643; PubMed Central PMCID: PMC1475499.

83. Guo C, Spinelli M, Ye C, Li QQ, Liang C. Genome-Wide Comparative Analysis of Miniature Inverted Repeat Transposable Elements in 19 *Arabidopsis thaliana* Ecotype Accessions. Sci Rep. 2017; 7 (1):2634. Epub 20170601. https://doi.org/10.1038/s41598-017-02855-1 PMID: 28572566; PubMed Central PMCID: PMC5454002.

84. Barah P, Jayavelu N, Mundy J, Bones A. Genome scale transcriptional response diversity among ten ecotypes of *Arabidopsis thaliana* during heat stress. Frontiers in Plant Science. 2013;4. https://doi.org/10.3389/fpls.2013.00532 PMID: 24409190

85. Kützing FT. Species algarum. Auctore Friderico Traug. Kützing. Lipsiae: F. A. Brockhaus; 1849.

86. Guiry MD, Guiry GM. AlgaeBase. National University of Ireland, Galway. 2023;(World-wide electronic publication).

87. Turland NJ, Wiersema JH, Barrie FR, Grueter W, Hawksworth DL, Herendeen PS, et al. International Code of Nomenclature for algae, fungi, and plats (Shenzhen Code) adopted by the Nineteenth International Botanical Congress Shenzhen, China.: Glashütten: Koeltz Botanical Books; 2018.

88. Maxwell JR, Douglas AG, Eglinton G, McCormick A. The Botryococcenes-hydrocarbons of novel structure from the alga *Botryococcus braunii*, Kützing. Phytochemistry. 1968; 7(12):2157–71. https://doi.org/10.1016/S0031-9422(00)85672-1

89. Lee SI, Kim NS. Transposable elements and genome size variations in plants. Genomics Inform. 2014; 12(3):87–97. Epub 20140930. https://doi.org/10.5808/GI.2014.12.3.87 PMID: 25317107; PubMed Central PMCID: PMC4196380.

90. Dai S-f, Zhu X-g, Hutang G-r,Li J-y, Tian J-q, Jiang X-h, et al. Genome Size Variation and Evolution Driven by Transposable Elements in the Genus *Oryza*. Frontiers in Plant Science. 2022;13. https://doi.org/10.3389/fpls.2022.921937 PMID: 35874017

91. Craig RJ, Hasan AR, Ness RW, Keightley PD. Comparative genomics of *Chlamydomonas*. Plant Cell. 2021; 33(4):1016–41. https://doi.org/10.1093/plcell/koab026 PMID: 33793842; PubMed Central PMCID: PMC8226300.

92. Zuccolo A, Sebastian A, Talag J, Yu Y, Kim H, Collura K, et al. Transposable element distribution, abundance and role in genome size variation in the genus *Oryza*. BMC Evolutionary Biology. 2007; 7 (1):152. https://doi.org/10.1186/1471-2148-7-152 PMID: 17727727

93. Hollister JD, Smith LM, Guo Y-L, Ott F, Weigel D, Gaut BS. Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. Proceedings of the National Academy of Sciences. 2011; 108(6):2322–7. https://doi.org/10.1073/pnas.1018222108 PMID: 21252301

94. Hufford Matthew B, Seetharam Arun S, Woodhouse Margaret R, Chougule Kapeel M, Ou S, Liu J, et al. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science. 2021; 373(6555):655–62. https://doi.org/10.1126/science.abg5289 PMID: 34353948

95. Tennant RK, Lux TM, Sambles CM, Kuhn NJ, Petticrew EL, Oldfield R, et al. Palaeogenomics of the Hydrocarbon Producing Microalga *Botryococcus braunii*. Sci Rep. 2019; 9(1):1776. Epub 20190211. https://doi.org/10.1038/s41598-018-38236-5 PMID: 30742038; PubMed Central PMCID: PMC6370823.

96. Volkman JK. Acyclic isoprenoid biomarkers and evolution of biosynthetic pathways in green microalgae of the genus *Botryococcus*. Organic Geochemistry. 2014; 75:36–47. https://doi.org/10.1016/j.orggeochem.2014.06.005

97. Lespinet O, Wolf YI, Koonin EV, Aravind L. The Role of Lineage-Specific Gene Family Expansion in the Evolution of Eukaryotes. Genome Research. 2002; 12(7):1048–59. https://doi.org/10.1101/gr.174302 PMID: 12097341

98. Oliver KR, Greene WK. Transposable elements: powerful facilitators of evolution. Bioessays. 2009; 31 (7):703–14. https://doi.org/10.1002/bies.200800219 PMID: 19415638.

99. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 Maize Genome: Complexity, Diversity, and Dynamics. Science. 2009; 326(5956):1112–5. https://doi.org/10.1126/science.1178534 PMID: 19965430

100. Komárek J, Marvan P. Morphological differences in natural populations of the genus *Botryococcus* chlorophyceae. Archiv für Protistenkunde. 1992; 141:65–100.

101. Plain N, Largeau C, Derenne S, Coute A. Variabilité morphologique de *Botryococcus braunii* (Chlorococcales, Chlorophyta): corrélations avec les conditions de croissance et la teneur en lipides; Morphological variability of *Botryococcus braunii* (Chlorococcales, Chlorophyta): correlations with growth conditions and lipid content. Phycologia. 1993; 32(4):259–65.

102. de Queiroz Mendes MC, González AA, Moreno ML, Figueira CP, de Castro Nunes JM. Morphological and Ultrastrucral Features of a Strain of *Botryococcus terribilis* (Trebouxiophyceae) From Brazil. J Phycol. 2012; 48(5):1099–106. Epub 20120525. https://doi.org/10.1111/j.1529-8817.2012.01181.x PMID: 27011271.

103. Hegedűs A, Mocan A, Barbu-Tudoran L, Coman C, Drugă B, Sicora C, Dragoş N. Morphological, biochemical, and phylogenetic assessments of eight *Botryococcus terribilis* strains collected from freshwaters of Transylvania. J Appl Phycol. 2015; 27(2):865–78. https://doi.org/10.1007/s10811-014-0387-2

104. Oliva AK, Bejaoui M, Hirano A, Arimura T, Linh TN, Uchiage E, et al. Elucidation of the Potential Hair Growth-Promoting Effect of *Botryococcus terribilis*, Its Novel Compound Methylated-Meijicoccene, and C32 Botryococcene on Cultured Hair Follicle Dermal Papilla Cells Using DNA Microarray Gene Expression Analysis. Biomedicines. 2022; 10(5):1186. https://doi.org/10.3390/biomedicines10051186 PMID: 35625924

105. Okada S, Murakami M, Yamaguchi K. Hydrocarbon composition of newly isolated strains of the green microalga *Botryococcus braunii*. J Appl Phycol. 1995; 7(6):555–9. https://doi.org/10.1007/BF00003942

106. Nonomura AM. *Botryococcus braunii* var. showa (Chlorophyceae) from Berkeley, California, United States of America. Jpn J Phycol. 1988; 36:285–91.

107. Healey A, Furtado A, Cooper T, Henry RJ. Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species. Plant Methods. 2014; 10:21. Epub 20140627. https://doi.org/10.1186/1746-4811-10-21 PMID: 25053969; PubMed Central PMCID: PMC4105509.

108. Inglis PW, Pappas MdCR, Resende LV, Grattapaglia D. Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. PLOS ONE. 2018; 13(10):e0206085. https://doi.org/10.1371/journal.pone.0206085 PMID: 30335843

109. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019; 37(5):540–6. Epub 20190401. https://doi.org/10.1038/s41587-019-0072-8 PMID: 30936562.

110. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30(15):2114–20. Epub 20140401. https://doi.org/10.1093/bioinformatics/btu170 PMID: 24695404; PubMed Central PMCID: PMC4103590.

111. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014; 9(11):e112963. Epub 20141119. https://doi.org/10.1371/journal.pone.0112963 PMID: 25409509; PubMed Central PMCID: PMC4237348.

112. Blifernez-Klassen O, Wibberg D, Winkler A, Blom J, Goesmann A, Kalinowski J, Kruse O. Complete Chloroplast and Mitochondrial Genome Sequences of the Hydrocarbon Oil-Producing Green Microalga *Botryococcus braunii* Race B (Showa). Genome Announc. 2016; 4(3):e00524–16. Epub 20160609. https://doi.org/10.1128/genomeA.00524-16 PMID: 27284138; PubMed Central PMCID: PMC4901229.

113. Laetsch D, Blaxter M. BlobTools: Interrogation of genome assemblies [version 1; peer review: 2 approved with reservations]. F1000Research. 2017; 6(1287). https://doi.org/10.12688/f1000research.12232.1

114. Mikheenko A, Prjibelski A, Saveliev V, Antipov D, Gurevich A. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics. 2018; 34(13):i142–i50. https://doi.org/10.1093/bioinformatics/bty266 PMID: 29949969; PubMed Central PMCID: PMC6022658.

115. Seppey M, Manni M, Zdobnov EM. BUSCO: Assessing Genome Assembly and Annotation Completeness. Methods in Molecular Biology: Springer New York; 2019. p. 227–45.

116. Zimin AV, Marcais G, Puiu D, Roberts M, Salzberg SL, Yorke JA. The MaSuRCA genome assembler. Bioinformatics. 2013; 29(21):2669–77. Epub 20130829. https://doi.org/10.1093/bioinformatics/btt476 PMID: 23990416; PubMed Central PMCID: PMC3799473.

117. Solares EA, Chakraborty M, Miller DE, Kalsow S, Hall K, Perera AG, et al. Rapid Low-Cost Assembly of the Drosophila melanogaster Reference Genome Using Low-Coverage, Long-Read Sequencing. G3 (Bethesda). 2018; 8(10):3143–54. Epub 20181003. https://doi.org/10.1534/g3.118.200162 PMID: 30018084; PubMed Central PMCID: PMC6169397.

118. Lischer HEL, Shimizu KK. Reference-guided de novo assembly approach improves genome reconstruction for related species. BMC Bioinformatics. 2017; 18(1). https://doi.org/10.1186/s12859-017-1911-6 PMID: 29126390

119. Olivares JA, Baxter I, Brown J, Carleton M, Cattolico RA, Taraka D, et al. National Alliance for Advanced Biofuels and Bio-Products Final Technical Report. United States: 2014 Contract No.: DOE-DANF-03046.

120. Unkefer CJ, Sayre RT, Magnuson JK, Anderson DB, Baxter I, Blaby IK, et al. Review of the algal biology program within the National Alliance for Advanced Biofuels and Bioproducts. Algal Res. 2017; 22:187–215. https://doi.org/10.1016/j.algal.2016.06.002 WOS:000397461000020.

121. Kim D, Paggi JM, Park C, Bennett C, Salzberg SL. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol. 2019; 37(8):907–15. Epub 20190802. https://doi.org/10.1038/s41587-019-0201-4 PMID: 31375807; PubMed Central PMCID: PMC7605509.

122. 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. Nature biotechnology. 2011; 29(7):644–52. Epub 20110515. https://doi.org/10.1038/nbt.1883 PMID: 21572440; PubMed Central PMCID: PMC3571712.

123. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006; 22(13):1658–9. Epub 20060526. https://doi.org/10.1093/bioinformatics/btl158 PMID: 16731699.

124. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008; 24(5):637–44. Epub 20080124. https://doi.org/10.1093/bioinformatics/btn013 PMID: 18218656.

125. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25(16):2078–9. Epub 20090608. https://doi.org/10.1093/bioinformatics/btp352 PMID: 19505943; PubMed Central PMCID: PMC2723002.

126. Barnett DW, Garrison EK, Quinlan AR, Stromberg MP, Marth GT. BamTools: a C++ API and toolkit for analyzing and managing BAM files. Bioinformatics. 2011; 27(12):1691–2. Epub 20110414. https://doi.org/10.1093/bioinformatics/btr174 PMID: 21493652; PubMed Central PMCID: PMC3106182.

127. Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. 2014; 42(15):e119. Epub 20140702. https://doi.org/10.1093/nar/gku557 PMID: 24990371; PubMed Central PMCID: PMC4150757.

128. Hoff KJ, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics (Oxford, England). 2016; 32 (5):767–9. Epub 2015/11/11. https://doi.org/10.1093/bioinformatics/btv661 PMID: 26559507.

129. Keilwagen J, Wenk M, Erickson JL, Schattat MH, Grau J, Hartung F. Using intron position conservation for homology-based gene prediction. Nucleic Acids Research. 2016; 44(9):e89–e. https://doi.org/10.1093/nar/gkw092 PMID: 26893356

130. Keilwagen J, Hartung F, Paulini M, Twardziok SO, Grau J. Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. BMC bioinformatics [Internet]. 2018 2018/05//; 19 (1):[189 p.]. Available from: https://doi.org/10.1186/s12859-018-2203-5 PMID: 29843602

131. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. Whole-Genome Annotation with BRAKER. In: Kollmar M, editor. Gene Prediction: Methods and Protocols. New York, NY: Springer New York; 2019. p. 65–95.

132. Keilwagen J, Hartung F, Grau J. GeMoMa: Homology-Based Gene Prediction Utilizing Intron Position Conservation and RNA-seq Data. In: Kollmar M, editor. Gene Prediction: Methods and Protocols. New York, NY: Springer New York; 2019. p. 161–77.

133. Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. BRAKER2: Automatic Eukaryotic Genome Annotation with GeneMark-EP+ and AUGUSTUS Supported by a Protein Database. bioRxiv. 2020:2020.08.10.245134. https://doi.org/10.1101/2020.08.10.245134

134. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010; 26(6):841–2. https://doi.org/10.1093/bioinformatics/btq033 PMID: 20110278

135. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013; 30(4):772–80. Epub 20130116. https://doi.org/10.1093/molbev/mst010 PMID: 23329690; PubMed Central PMCID: PMC3603318.

136. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015; 32(1):268–74. Epub 20141103. https://doi.org/10.1093/molbev/msu300 PMID: 25371430; PubMed Central PMCID: PMC4271533.

137. Zhang C, Rabiee M, Sayyari E, Mirarab S. ASTRAL-III: polynomial time species tree reconstruction from partially resolved gene trees. BMC Bioinformatics. 2018; 19(Suppl 6):153. Epub 20180508. https://doi.org/10.1186/s12859-018-2129-y PMID: 29745866; PubMed Central PMCID: PMC5998893.

138. Prochnik SE, Umen J, Nedelcu AM, Hallmann A, Miller SM, Nishii I, et al. Genomic analysis of organismal complexity in the multicellular green alga *Volvox carteri*. Science. 2010; 329(5988):223–6. Epub 2010/07/10. https://doi.org/10.1126/science.1188800 PMID: 20616280.

139. Gao C, Wang Y, Shen Y, Yan D, He X, Dai J, Wu Q. Oil accumulation mechanisms of the oleaginous microalga *Chlorella protothecoides* revealed through its genome, transcriptomes, and proteomes. BMC Genomics. 2014; 15(1):582. Epub 20140710. https://doi.org/10.1186/1471-2164-15-582 PMID: 25012212; PubMed Central PMCID: PMC4111847.

140. Arriola MB, Velmurugan N, Zhang Y, Plunkett MH, Hondzo H, Barney BM. Genome sequences of *Chlorella sorokiniana* UTEX 1602 and *Micractinium conductrix* SAG 241.80: implications to maltose excretion by a green alga. Plant J. 2018; 93(3):566–86. Epub 20180110. https://doi.org/10.1111/tpj.13789 PMID: 29178410.

141. Gonzalez-Esquer CR, Twary SN, Hovde BT, Starkenburg SR. Nuclear, Chloroplast, and Mitochondrial Genome Sequences of the Prospective Microalgal Biofuel Strain *Picochlorum soloecismus*. Genome Announc. 2018; 6(4). Epub 20180125. https://doi.org/10.1128/genomeA.01498-17 PMID: 29371352; PubMed Central PMCID: PMC5786678.

142. Armaleo D, Müller O, Lutzoni F, Andrésson Ó S, Blanc G, Bode HB, et al. The lichen symbiosis re-viewed through the genomes of *Cladonia grayi* and its algal partner *Asterochloris glomerata*. BMC Genomics. 2019; 20(1):605. Epub 20190723. https://doi.org/10.1186/s12864-019-5629-x PMID: 31337355; PubMed Central PMCID: PMC6652019.

143. Greshake Tzovaras B, Segers F, Bicker A, Dal Grande F, Otte J, Anvar SY, et al. What Is in *Umbilicaria pustulata*? A Metagenomic Approach to Reconstruct the Holo-Genome of a Lichen. Genome Biol Evol. 2020; 12(4):309–24. https://doi.org/10.1093/gbe/evaa049 PMID: 32163141; PubMed Central PMCID: PMC7186782.

144. Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar Gustavo A, et al. InterPro in 2022. Nucleic Acids Research. 2022; 51(D1):D418–D27. https://doi.org/10.1093/nar/gkac993 PMID: 36350672

145. Eddy SR. Multiple alignmnet using hidden Markov models. Proc Third Int Conf Intelligent Systems for Molecular Biology. 1995:114–20.

146. Eddy SR. Accelerated Profile HMM Searches. PLOS Computational Biology. 2011; 7(10):e1002195. https://doi.org/10.1371/journal.pcbi.1002195 PMID: 22039361