SOCIETY FOR EXPERIMENTAL BIOLOGY

*Journal of Experimental Botany*, Vol. 72, No. 7 pp. 2491–2500, 2021
doi:10.1093/jxb/erab006 Advance Access Publication 17 January 2021

RESEARCH PAPER

# Improved *Spirodela polyrhiza* genome and proteomic analyses reveal a conserved chromosomal structure with high abundance of chloroplastic proteins favoring energy production

**Alex Harkess[1,*,†], Fionn McLoughlin[2,*], Natasha Bilkey[2], Kiona Elliott[2], Ryan Emenecker[2], Erin Mattoon[2], Kari Miller[2], Kirk Czymmek[1], Richard D. Vierstra[2], Blake C. Meyers[1,3] and Todd P. Michael[4,‡]**

[1] Donald Danforth Plant Science Center, St Louis, MO 63132, USA
[2] Department of Biology, Washington University, St Louis, MO 63130, USA
[3] Division of Plant Sciences, University of Missouri, Columbia, MO 65211, USA
[4] Department of Informatics, J. Craig Venter Institute (JCVI), San Diego, CA 92037, USA

‡ Present address: The Plant Molecular and Cellular Biology Laboratory, The Salk Institute for Biological Studies, La Jolla, CA 92037, USA
* These authors contributed equally to this work.
† Correspondence: aharkess@hudsonalpha.org

Received 8 June 2020; Editorial decision 7 January 2021; Accepted 13 January 2021

Editor: Rainer Melzer, University College Dublin, Ireland

## Abstract

**Duckweeds are a monophyletic group of rapidly reproducing aquatic monocots in the Lemnaceae family. Given their clonal, exponentially fast reproduction, a key question is whether genome structure is conserved across the species in the absence of meiotic recombination. Here, we studied the genome and proteome of *Spirodela polyrhiza*, or greater duckweed, which has the largest body plan yet the smallest genome size in the family (1C=150 Mb). Using Oxford Nanopore sequencing combined with Hi-C scaffolding, we generated a highly contiguous, chromosome-scale assembly of *S. polyrhiza* line Sp7498 (Sp7498_HiC). Both the Sp7498_HiC and Sp9509 genome assemblies reveal large chromosomal misorientations relative to a recent PacBio assembly of Sp7498, highlighting the need for orthogonal long-range scaffolding techniques such as Hi-C and BioNano optical mapping. Shotgun proteomics of Sp7498 verified the expression of ~2250 proteins and revealed a high abundance of proteins involved in photosynthesis and carbohydrate metabolism among other functions. In addition, a strong increase in chloroplast proteins was observed that correlated to chloroplast density. This Sp7498_HiC genome was generated cheaply and quickly with a single Oxford Nanopore MinION flow cell and one Hi-C library in a classroom setting. Combining these data with a mass spectrometry-generated proteome illustrates the utility of duckweed as a model for genomics- and proteomics-based education.**

**Keywords:** Chloroplast, duckweed, Oxford Nanopore, proteomics, *Spirodela*.

## Introduction

Duckweeds are the fastest reproducing flowering plants, with some species capable of dividing in less than 2 days under optimal conditions (Ziegler *et al.*, 2015). Found on every continent except Antarctica, the Lemnaceae family of duckweeds is taxonomically divided into 36 species across five genera: *Spirodela*, *Lemna*, *Landoltia*, *Wolffia*, and *Wolffiella* (Bog *et al.*, 2020). They are strictly aquatic, typically floating on slow-moving bodies of freshwater.

*Spirodela polyrhiza* has the smallest known duckweed genome size at 158 Mb (Wang *et al.*, 2011). The initial *S. polyrhiza* reference generated by the Department of Energy Joint Genome Initiative (DOE-JGI) was based on clone 7498 (Sp7498) collected from Durham, NC, USA (35°N 75°W) that was received directly from Elias Landolt, an early pioneer of duckweed taxonomy, physiology, and anatomy (Urbanska *et al.*, 2013). The initial Sp7498 genome assembly was based on 454 pyrosequencing reads and BAC-end sequencing. The resulting draft assembly was sufficiently contiguous to reveal that while *S. polyrhiza* retains a core set of plant genes, it shows a reduction in most gene families resulting in a total of 19 623 protein-coding genes (Wang *et al.*, 2014). At the time this was the smallest number of genes found in a flowering plant genome, perhaps consistent with its greatly reduced body plan and rapid growth. Afterwards, the genome of the sea-grass *Zostera marina* was sequenced and also found to harbor a reduced protein-coding gene set totaling 20 450, suggesting that a reduced gene count may be a common feature of aquatic plants (Olsen *et al.*, 2016). However, the Sp7498 genome assembly only covered 90% of the expected genome size, with 10.7% of the assembly in gaps filled with Ns. Instead of being resolved into the expected 20 chromosomes, it was anchored onto 32 pseudo-molecules (Wang *et al.*, 2014).

*Spirodela* species form specialized, dormant winter buds called turions. Turions sink to the bottom of the water where they rest, surviving on starch reserves (40–70% dry weight) until favorable growing conditions stimulate floatation and vegetative growth. To better elucidate turion formation, a second turion-producing *S. polyrhiza*, clone 9509 (Sp9509) from Lotschen, Stadtroda Germany (50N 11W), was sequenced and assembled (Michael *et al.*, 2017). The Sp9509 line was sequenced using a combination of Illumina short read libraries and scaffolded into 20 chromosomes using BioNano Genomics optical maps. In combination with extensive expression (RNA-seq) and DNA methylation (bisulfite-seq) analysis, the minimal set of genes was confirmed, which also revealed that *Spirodela* has the lowest levels of genome-wide DNA methylation of any flowering plant examined (Michael *et al.*, 2017). The Sp9509 assembly was later updated using Oxford Nanopore single molecule long read sequencing and validated using the optical maps, which resulted in a highly contiguous *Spirodela* genome (Hoang *et al.*, 2018).

Recently, the Sp7498 genome assembly was updated using Pacific Biosciences (PacBio) Sequel reads (An *et al.*, 2019). The resulting assembly was less contiguous than Sp9509 with a contig N50 length of 0.83 Mb compared with Sp9509 at 2.87 Mb (Table 1), and did not leverage information from chromosome-scale technologies (Michael *et al.*, 2017; Hoang *et al.*, 2018), such as BioNano optical mapping or Hi-C chromatin conformation sequencing. To identify any possible mis-assemblies in the latest PacBio Sp7498 genome assembly, and more broadly to assess genome stability across the *S. polyrhiza* species, we generated a single MinION flow cell of Oxford Nanopore long-read data for this line, as well as Phase Genomics Hi-C data, as part of a graduate seminar course at Washington University in St Louis (MO, USA). As a novel strategy to more accurately map exomes, we also subjected Sp7498 to shotgun proteome analysis via tandem mass spectrometry (MS/MS) following liquid chromatographic separation of trypsinized soluble proteomes, which helped verify the expression of ~2250 proteins. We combined these data to present a high-quality genome assembly and proteome for Sp7498, and we highlight the usefulness of duckweeds in a classroom setting for genomics and proteomics.

**Table 1.** Assembly statistics for Sp7498 and Sp9509 genomes

| | Sp7498_PacBio | | Sp7498_HiC | | Sp9509 | |
|---|---|---|---|---|---|---|
| | **Contigs** | **Scaffolds** | **Contigs[a]** | **Scaffolds[a]** | **Contigs** | **Scaffolds** |
| Sequencer | PacBio Sequel | PacBio Sequel | Oxford Nanopore | Oxford Nanopore | Oxford Nanopore | Oxford Nanopore |
| Scaffolding | BAC-FISH | BAC-FISH | Hi-C | Hi-C | BioNano | BioNano |
| Total length (bp) | 138 525 388 | 138 536 245 | 138 486 032 | 138 493 532 | 138 570 896 | 138 592 155 |
| Contig count | 384 | 134 | 99 | 24 | 95 | 20 |
| N50 length (bp) | 837 937 | 7 645 982 | 3 339 466 | 7 689 391 | 2 868 147 | 7 949 387 |
| L50 count | 52 | 8 | 15 | 8 | 16 | 8 |
| Ns count | 0 | 10 857 | 0 | 7 500 | 9 | 21 268 |

The genome assemblies were compared using quast to generate basic statistics.

[a] Collected in this study.

BioNano, BioNano Genomics optical mapping; HiC, high throughput chromatin conformation capture; L50, the number of contigs at half the length of the assembly; N50 length, the length of the contig at half of the assembly; Ns, gap in the assembly filled with N to represent unknown bases; Oxford Nanopore, Oxford Nanopore Technologies; PacBio, Pacific Bioscience.

## Materials and methods

### *Sp7498 Oxford Nanopore genome sequencing, assembly, and annotation*

*Spirodela polyrhiza* line Sp7498 tissue was obtained from the Rutgers Duckweed Stock Cooperative (RDSC; http://www.ruduckweed.org/) and grown in Hoaglands No. 2 Basal Salt Mixture (1.6 g $l^{-1}$) at 25 °C under 16 h days. High molecular mass DNA was isolated from 1.5 g of flash-frozen whole-plant *S. polyrhiza* 7498 tissue using 10 ml CTAB buffer (100 mM Tris–HCl pH 8.0, 1.4 M NaCl, 20 mM EDTA, 2% cetyl trimethyl ammonium bromide (CTAB), 2% polyvinylpyrrolidone, 0.2% 2-mercaptoethanol) in a 65 °C waterbath for 45 min. DNA was purified twice with 1 volume of 24:1 chloroform:isoamyl alcohol, precipitated with isopropanol and spooled out on a glass shepherd's crook, treated with 3 µl RNase A (20 µg $ml^{-1}$) at 37 °C for 30 min, and left on a benchtop at room temperature overnight. The DNA was purified again with 24:1 chloroform:isoamyl alcohol, and resuspended in EB buffer (Qiagen, Hilden, Germany) on a benchtop at room temperature for 48 h. Approximately 1 µg of DNA, measured by Qubit dsDNA BR, was used as input to the LSK-109 library kit (Oxford Nanopore Technologies), loaded onto a 9.4.1 flow cell and sequenced for 48 h. Raw signal data were base-called using Guppy v3.1.5 with the 'flip-flop' algorithm. Reads were assembled with minimap2 (2.17-r941) (Li, 2018) and miniasm (0.2-r128) (Li, 2016) with default options, then error-corrected with three rounds of Racon (Vaser *et al.*, 2017) and one round of Medaka (https://github.com/nanoporetech/medaka).

Hi-C data were generated from frozen whole plant line 7498 tissue sent to Phase Genomics (Seattle, WA, USA) using the Sau3IA cut-site and assembled into pseudomolecules with Proximo software. The resulting assembly and annotation was syntenically compared with Sp9509 and Sp7498_PacBio using CoGE SynMap with default DAGChainer options (−D 20 −A 5) (Lyons and Freeling, 2008). Gene annotations were lifted over from Sp9509 to the Sp7498_HiC assembly within CoGE. Detailed statistics on misassemblies, mismatches, and indels were generated using QUAST v5.0.2 with default options (Gurevich *et al.*, 2013). Long terminal repeat (LTR) retrotransposons were annotated in all three genomes using default parameters in GenomeTools LTRharvest (v1.5.8) (Ellinghaus *et al.*, 2008).

### *Mass spectrometry proteome profiling*

Arabidopsis seedlings were grown in shaking (90 rpm) liquid MS medium (4.4 g $l^{-1}$ Murashige and Skoog basal medium, 1% (w/v) sucrose, 0.05% (w/v) MES (pH 5.7)) at 21–23 °C under a long-day photoperiod (16 h light–8 h darkness) with a light intensity of 75–100 µmol $m^{-2}$ $s^{-1}$. The seedlings were drained from excess liquid using paper towels and frozen in liquid nitrogen. Proteins were extracted at 4 °C from pulverized *S. polyrhiza* 7498 and Arabidopsis Col-0 leaf tissue into 50 mM HEPES (pH 7.5), 5 mM $Na_2EDTA$, 2 mM dithiotreitol, and 1× plant protease inhibitor cocktail (Sigma-Aldrich). The protein extraction buffer was adjusted based on the fresh weight of the tissue (ratio sample:buffer=1:2), effectively normalizing the protein content between the two species. The samples were further homogenized using a Pyrex Potter-Elvehjem tissue grinder (Fisher Scientific) and clarified by centrifugation at 16 000 *g*. Total protein extract (150 µl) was precipitated in 4:1:3 (v/v) methanol–chloroform–water and collected by centrifugation. The resulting pellet was lyophilized to dryness and resuspended into 100 µl of 8 M urea, reduced for 1 h at 22 °C with 10 mM dithiothreitol, followed by alkylation with 20 mM iodoacetamide for 1 h. The reactions were quenched with 20 mM dithiotreitol and diluted with with 900 µl of 25 mM ammonium bicarbonate to reduce the urea concentrations below 1.5 M, and digested overnight at 37 °C with 0.5 µg of sequencing-grade modified porcine trypsin (Promega). The resulting peptides were lyophilized to a final volume of ~250 µl, acidified with 0.5% v/v trifluoroacetic acid (pH<3.0), and desalted and concentrated using a 100 µl Bond Elute OMIX C18 pipette tip (Agilent Technologies) according to the manufacturer's instructions. The peptides were eluted in 50 µl of 75% acetonitrile and 0.1% acetic acid, lyophilized, and resuspended in 20 µl 5% acetonitrile and 0.1% formic acid for MS/MS analysis.

Nano-scale LC separation of the tryptic peptides was performed using a Dionex Ultimate 3000 Rapid Separation system equipped with a 75 µm×25 cm Acclaim PepMap RSLC C18 column (Thermo Fisher Scientific), in combination with a 2 h linear 4% to 36% acetonitrile gradient in 0.1% formic acid and a flow rate of 250 nl $min^{-1}$. Eluted peptides were analysed online by a Q Exactive Plus spectrometer (Thermo Fisher Scientific) in the positive electrospray ionization mode. Data-dependent acquisition of full MS scans (mass range of 380–1500 $m/z$) at a resolution of 70 000 was collected, with the automatic gain control (AGC) target set to $3\times10^6$, and the maximum fill time set to 200 ms. High-energy collision-induced dissociation fragmentation of the 15 strongest peaks was performed with an intensity threshold of $4\times10^4$ counts and an isolation window of 3.0 $m/z$, excluding precursors that had unassigned, +1, +7, +8, or >+8 charge states. MS/MS scans were conducted at a resolution of 17 500, with an AGC target of $2\times10^5$ and a maximum fill time of 100 ms. Dynamic exclusion was performed with a repeat count of 2 and an exclusion duration of 30 s, while the minimum MS ion count for triggering MS/MS was set to $4\times10^3$ counts. Each sample was analysed in quadruplicate to enable broad coverage; the first two runs were performed without an exclusion list, while the third and fourth runs were performed with an exclusion list containing the 5000 most abundant peptides that were detected in the first two runs, to increase sample coverage and maximize suppression of abundant peptides. Raw MS2 files from all four runs were merged, resulting in two technical replicates per sample (McLoughlin *et al.*, 2018).

The resulting MS/MS datasets were queried by Proteome Discoverer (version 2.0.0.802; Thermo Fisher Scientific) against the *S. polyrhiza* Sp7498 predicted proteome (Spolyrhiza_290_v2.protein.primarytranscriptonly.header.fa; http:/phytozome.jgi.doe.gov (nuclear) and GenBank MN419335.1 (chloroplast encoded) and Arabidopsis (TAIR10_PEP_20101214_UPDATED.fa; http://www.arabidopsis.org (Nuclear) and NCBI Reference NC_000932.1 (Chloroplast encoded)) predicted proteome databases and a list of common protein contaminants.

Peptides were assigned by SEQUEST HT (Eng *et al.*, 1994), allowing a maximum of one missed tryptic cleavage, a minimum peptide length of six, a precursor mass tolerance of 10 ppm, and fragment mass tolerances of 0.02 Da. Carbamidomethylation of cysteines and oxidation of methionine were specified as static and dynamic modifications, respectively.

The target false discovery rates (FDR) of ≤0.01 (strict, high confidence) and ≤0.05 (relaxed, medium confidence) were used as validation for peptide spectral matches (PSMs); protein modifications are only reported at high PSM confidence (FDR 0.01); and peptide and protein grouping was performed excluding all protein groups that are not strictly necessary to explain the identified peptides (strict). Label-free quantification was obtained in Proteome Discoverer™ as previously described (Silva *et al.*, 2006) with a minimum Quan value threshold set to 0.0001 using unique peptides, and '3 Top N' peptides were used for area calculation (Silva *et al.*, 2006). Three biological replicates were each analysed in quadruplicate, and the resulting values were averaged, which resulted in 2163 master protein identifications in common between *S. polyrhiza* and Arabidopsis. Data were $log_2$-transformed and missing values were imputed while assuming a normal distribution within the Perseus computational platform (Tyanova *et al.*, 2016), based on a width distribution shrinkage of 0.3 and a downshift of 1.8 standard deviations if necessary. Peptides were analysed and significance was determined using one-way ANOVA contrasts ($P$-value<0.05), excluding peptides with a fold change (FC)>2. Gene Ontology (GO) enrichments were identified using the AgriGO analysis toolkit (Tian *et al.*, 2017). GO-annotation categories shown here were selected based on their uniqueness, $P$-value significance, and degree of completeness.

*Confocal microscopy and chloroplast density*

Chloroplast densities and distributions were quantified in a mature, fully expanded leaf of Arabidopsis Col-0 and a mature frond of *S. polyrhiza* 7498 by confocal microscopy. Before imaging, the tissues were syringe-infiltrated with perfluoroperhydrophenanthrene (Sigma-Aldrich, cat. no. 58919) to fill air spaces and allow deep imaging into the mesophyll by confocal microscopy (Littlejohn *et al.*, 2014). Three-dimensional volumes were captured on a Leica SP8-X microscope using HC PlanApochromat ×63 water immersion objective lens with the pinhole set to 1 Airy unit and *z*-step of 0.6 µm; 405 and 649 laser excitation with the spectral prism set to collect 414–516 nm and 658–768 nm emission was used for imaging cell wall and chloroplast autofluorescence, respectively. Three-dimensional volumes were processed in FIJI (ImageJ version 1.51w) by thresholding background in the red channel (representing chloroplast autofluorescence) and averaging the calculated percentage area of chloroplasts in each volume with Z-stack slices extending from the epidermis to the mesophyll within the frond tissue.

## Results

To generate a high quality genome assembly for Sp7498, we sequenced long-read DNA on one Oxford Nanopore MinION 9.4.1 flow cell. After filtering poor quality 'failed' reads with the Guppy base caller, we generated 1.4 million clean single molecule nanopore reads with an N50 of 13.5 kb and total length of 6.4 gigabases, providing approximately 46× coverage of the roughly 138 megabase *S. polyrhiza* 7498 genome. Genome assembly with miniasm produced a contig assembly with an N50 of 3.34 Mb (Fig. 1A). The total length of the assembly was 138.49 Mb, which is highly congruent with all existing Sp7498 assemblies regardless of technology (Table 1).

To scaffold the contig assembly into chromosomes, a Phase Genomics Hi-C library was generated and sequenced to 115 741 530 read pairs. Further scaffolding and polishing with the Phase Genomics Hi-C links using Proximo identified no misjoined segments in the assembly and combined the assembly into 20 chromosomes plus four unplaced scaffolds (Fig. 1B, C). The four unplaced scaffolds totalled 2.8 Mb, or 2.0% of the total assembly length. The canonical telomere repeat (5′-TTTAGGG-3′) in Sp7498_HiC was the same as identified in Sp9509 (Michael *et al.*, 2017). All 20 chromosomes in Sp7498_HiC contain a telomere repeat at the distal tip of at least one arm of the chromosome, and 12 chromosomes have telomere repeats at the distal tips of both arms (see Supplementary Fig. S1).

We then tested the degree of synteny of our Sp7498_HiC assembly against the recently published Sp7498_PacBio genome, which should be genotypically identical. The assembled genome sizes were nearly the same, differing by only 42 713 nt more in the Sp7498_HiC assembly. Across the whole genome, GC content was nearly identical as well (see Supplementary Fig. S2). A synteny dot-plot revealed several major discrepancies between the two assemblies, including several whole chromosome arm inversions (Fig. 2A). Across the aligned length of both genomes, 689 misassemblies (breakpoints) were identified when using Sp7498_HiC as the reference. These included eight inversions, 96 774 single nucleotide mismatches and 315 762 indels, 96% of which were less than 6 nt in size. Nearly 1.6 Mb of sequence in the Sp7498_HiC assembly did not share an alignment with the Sp7498_Pacbio genome.

Next, we tested the level of conservation across two unique clones of *S. polyrhiza*, Sp7498 and Sp9509. Both of these genomes (Sp7498_HiC and Sp9509) were assembled with similar methods, based on Oxford Nanopore MinION 9.4.1 flow cells and assembly with miniasm. Overall the two genomes were

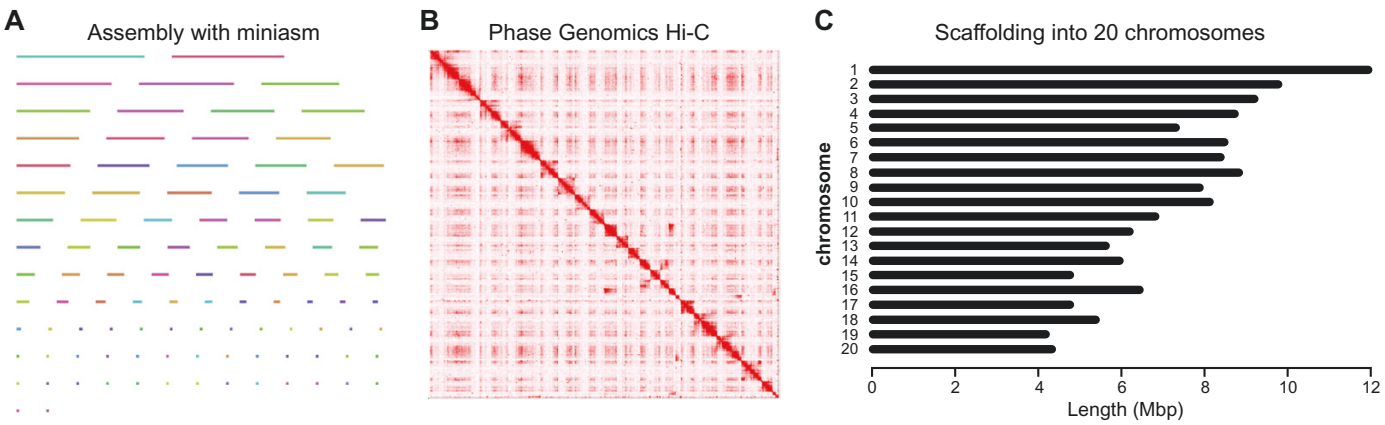


**Fig. 1.** Assembly and scaffolding strategy overview for the Sp7498 genome. (A) Contigs were assembled from raw Oxford Nanopore reads using miniasm to generate 99 contigs. Contigs were polished with three iterations of Racon, followed by one iteration of Medaka, and visualized with Bandage. (B) Chromosome scaffolding using Hi-C links. The Hi-C technique uses chromatin isolation, proximity ligation, and sequencing to identify DNA–DNA interactions over short and long distances on chromatin. Hi-C sequencing results in read pairs that represent these distant chromatin interactions, allowing for unordered contigs to be scaffolded into full-length chromosome pseudomolecules. These short- and long-distance interactions can be represented by a dot-plot matrix as shown here. For Sp7498, a Phase Genomics Hi-C library was sequenced, followed by Proximo scaffolding and manual polishing in JuiceBox into chromosome pseudomolecules. (C) Chromosome naming and length distribution. All 20 chromosomes were named according to synteny against the Sp9509 assembly.

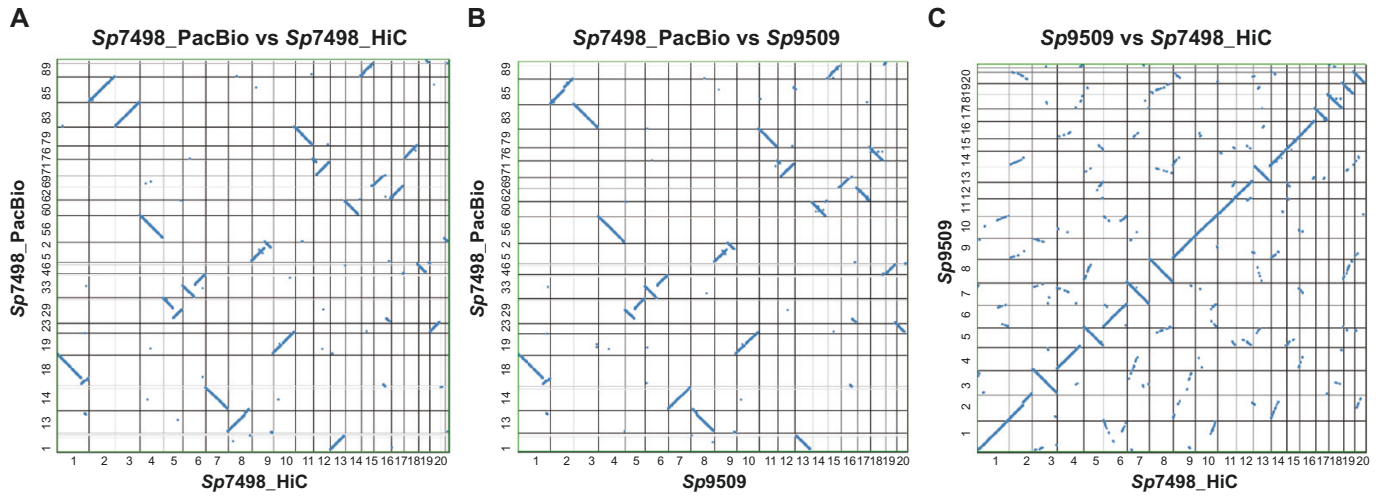


**Fig. 2.** Syntenic dot-plot comparisons of Sp7498 and Sp9509 genomes. Each dot represents a syntenic block between both genomes, containing at least five syntenic genes over a 20 gene stretch. All comparisons were performed in CoGE using default SynMap parameters. Any comparison involving Sp7498_PNAS was performed using gene predictions provided by the Sp9509 genome assembly since no gene annotations were publicly available. (A) Sp7498_HiC versus Sp7498_PNAS, (B) Sp9509 versus Sp7498_PNAS, and (C) Sp9509 versus Sp7498_HiC.

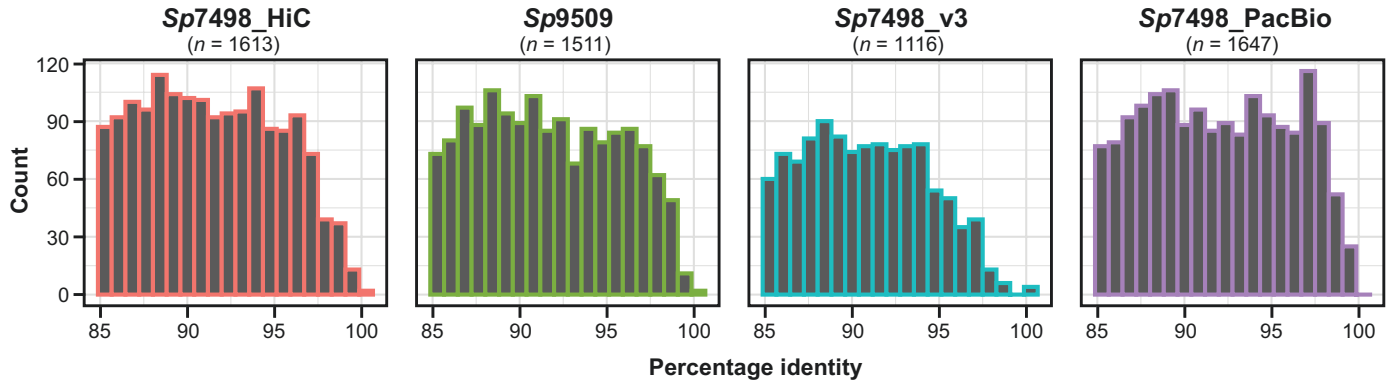


**Fig. 3.** Comparison of LTR retrotransposon content and relative insertion timing across four *Spirodela polyrhiza* genome assemblies. Percentage identity was calculated by pairwise comparisons of LTR ends for each full-length LTR retrotransposon. Recently inserted retrotransposons have 100% pairwise LTR identity that typically decreases over time as the ends diverge.

highly syntenic across the whole genome, and shared nearly the same contig orders and orientations for all 20 chromosomes (Fig. 2C). Both the Sp7498_HiC and Sp9509 genomes shared the same large structural variation disagreements with the Sp7498_HiC genome (Fig. 2B). Of the 22 605 proteins annotated in the Sp9509 genome, we were able to lift over 20 530 proteins onto the Sp7498_HiC genome.

We next asked whether repeat content variation could explain some of this variation between genome assemblies. In all three assemblies (Sp7498_HiC, Sp7498_PacBio, Sp9509), repeat content of LTR retrotransposons was low, with between 1511 and 1647 annotations (Fig. 3). Sp7498_PacBio contained the greatest number of LTR annotations ($n$=1647), whereas Sp9509 had the fewest ($n$=1511). We compared these three long-read assemblies with the first *S. polyrhiza* genome (Sp7498_v3), which was produced with a combination of 454 pyrosequencing, Sanger reads, and BAC-end sequencing. The Sp7498_v3 assembly showed a marked decrease in total number of LTR annotations, as well as a reduction in young (i.e. high percentage identity of LTR ends) retrotransposons in the assembly.

To help validate expression of the predicted protein-coding regions we surveyed *S. polyrhiza* 7498 by shotgun MS/MS analysis of the total protein extracts following trypsinization, which resulted in the detection of 2289 proteins. Although the protein-coding loci were generally uniformly mapped to all chromosomes, clusters were observed on chromosomes 4, 6, 7, 10, and 13, which could represent highly active euchromatic regions (Fig. 4A). In contrast, chromosomes 15 and 17 had a lower number of MS/MS-confirmed protein-coding loci, suggesting that these two chromosomes are less active in expressing abundant *S. polyrhiza* proteins.

To identify unique characteristics of the *S. polyrhiza* 7498 proteome, a comparative analysis was conducted with the model species Arabidopsis. As a first pass analysis, we identified more than 11 000 and 10 000 unique peptides in the MS/MS datasets for *S. polyrhiza* and Arabidopsis, respectively, which were assigned to ~2250 protein groups in both species (Fig. 4B; Supplementary Tables S1, S2). Of this total, homologs for ~1300 were detected in the other species, while ~950 were tentatively assigned as unique to one or the other (Fig. 4C, Supplementary Table S3). Because presence or absence of a homology by MS is often challenged by the lack of detection in complex samples and by variations in protein sequence, we restricted our analysis to exactly matched peptides between *S. polyrhiza* and Arabidopsis.

To hone-in on processes or cellular components that are either over- or under-represented in *S. polyrhiza* 7498 versus Arabidopsis, we selected 938 peptides that were perfectly conserved between the two species, and estimated their relative abundances based on the precursor ion intensities of the peptides from the MS1 scans. When their abundances were plotted in a volcano plot based on fold change differences between the two species and *P*-values of significance, 70 peptides were considered to be significantly more abundant in Arabidopsis and 142 in *S. polyrhiza*. (Fig. 4D). GO analysis

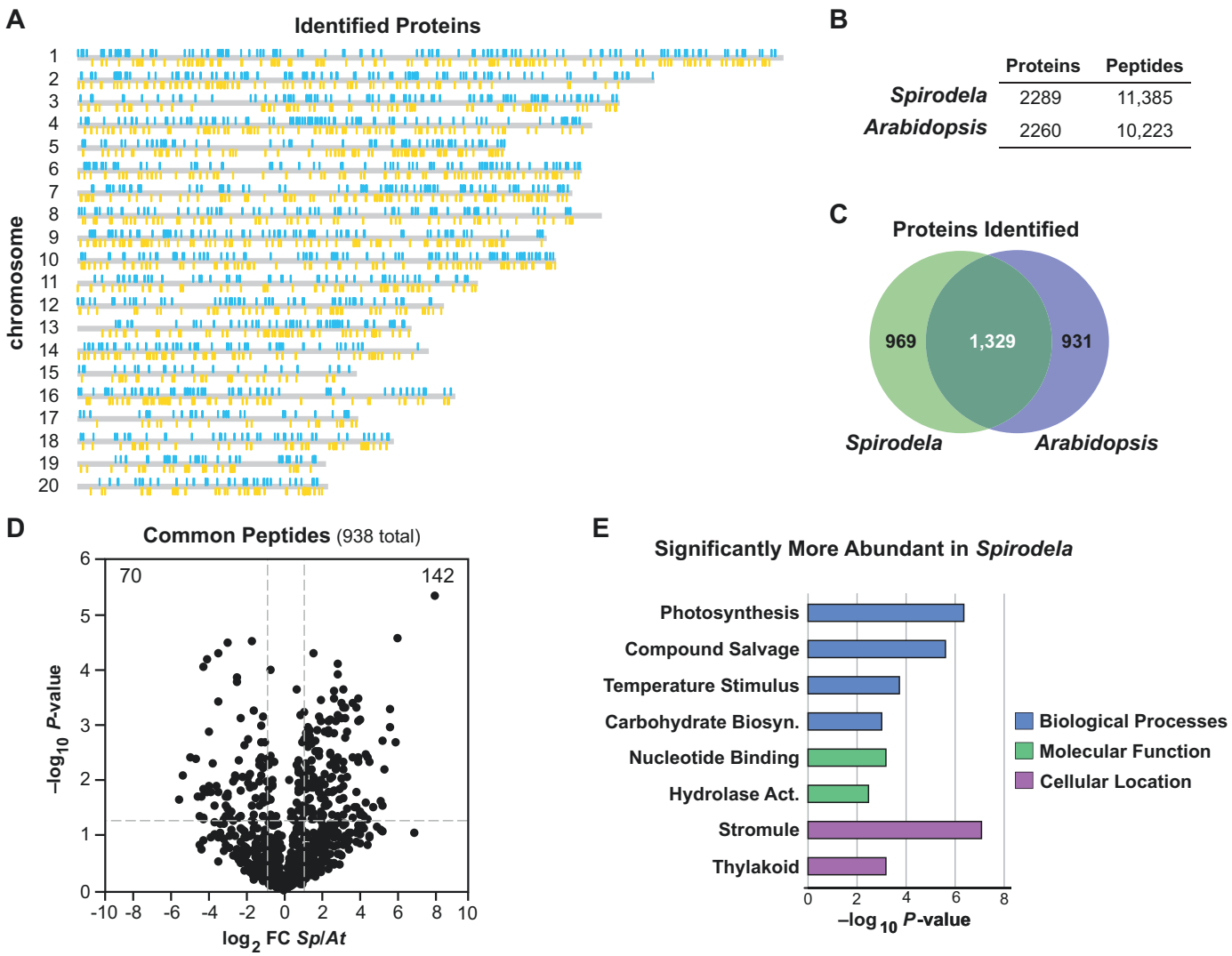


**Fig. 4.** Analysis of the *S. polyrhiza* 7498 proteome. Proteins were identified and quantified by tandem MS from a total soluble protein extract following trypsinization. (A) Confirmed protein-coding regions highlighted on chromosome maps. Blue and yellow identify protein-coding sequence orientations on the Watson and Crick strands, respectively. (B) Numbers of peptides and proteins identified that resulted in the identification of ~2250 proteins. (C) Venn diagram displaying the overlap of predicted protein homologs in *Spirodela* and Arabidopsis. (D) Quantitative analysis of 938 conserved peptides between *Spirodela* and Arabidopsis that were identified and quantified using MS1 precursor ion intensities. Significance in abundance was called at $P<0.05$, FC>2× ($n=3$). (E) Comparisons of estimated protein abundances of proteins shared between *S. polyrhiza* 7498 and Arabidopsis. GO analysis was conducted on proteins that harbor at least two significantly affected peptides. Fold change in abundances and *P*-value of each protein were plotted as were –log transformed values. Dashed lines indicate a $\log_2$FC of 1 and –log10 *P*-values of significance < or >1.3.

was then conducted on a subset of these proteins for which at least two peptides were identified that were significantly affected (see Supplementary Table S4), using all the commonly detected proteins as a reference (Fig. 4E). Although there were no significantly affected categories in Arabidopsis as compared with *S. polyrhiza* 7498, a significant over-representation of proteins involved in various processes directly related to energy production (e.g. photosynthesis and carbohydrate metabolism) was observed for *S. polyrhiza* 7498. The enrichment of the energy production-related protein clusters was consistent and could contribute to the rapid reproduction rate of *Spirodela*.

To validate the increases seen for several proteins related to energy production in *S. polyrhiza* 7498, we plotted the relative abundance of peptides conserved between Arabidopsis and *Spirodela* (Fig. 5A). Strikingly, among the proteins more abundant in *Spirodela* were ribulose-bisphosphate carboxylase (RuBisCO; carbon fixation), glyceraldehyde-3-phosphate dehydrogenase (GAPDHa and GAPDHb; glycolysis), photosystem I reaction center subunit (PSAN; light driven oxidoreductase) and phosphoribulokinase (PRK; Calvin cycle). These increases were consistent for all the selected proteins and suggests an emphasis on various energy-related processes in *Spirodela*.

Because these processes are primarily focused within chloroplasts, we reasoned that the increased protein levels might reflect a higher density or an enlargement of chloroplasts in *S. polyrhiza* 7498. This would be consistent with the increase in chlorophyll content that was apparent by the darker green color of the protein extracts when compared on a per-gram fresh weight basis (see Supplementary Fig. S3). To test this notion, we quantified chloroplast volumes in *S. polyrhiza*

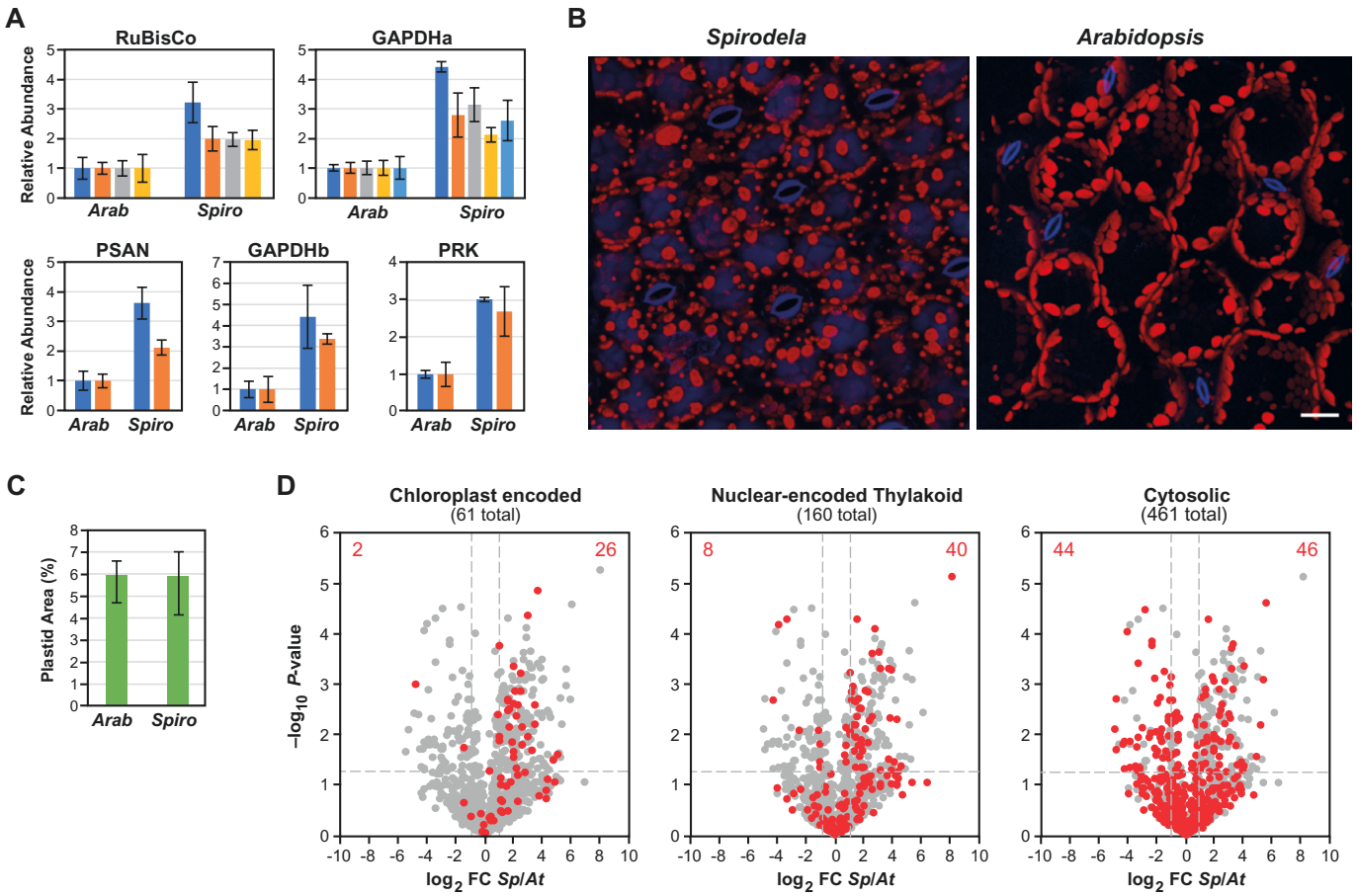


**Fig. 5.** Quantification of peptides from various chloroplast proteins and the chloroplast abundance and distribution. (A) Estimated protein abundances for selected chloroplast proteins based on peptide abundances. Peptide abundance was determined using the precursor ion intensities provided by the MS1 scans, and normalized to the average value obtained from Arabidopsis. Each bar represents the mean value (±SD) of unique and conserved peptides for each protein. (B) Imaging of chloroplasts from *S. polyrhiza* 7498 and Arabidopsis by confocal fluorescence microscopy. UV autofluorescence of cell walls and vacuoles, and chlorophyll autofluorescence are shown in blue and red colors, respectively. Scale bar: 20 μm. (C) Average percentage intracellular volumes of chloroplasts in mesophyll cells from *S. polyrhiza* 7498 and Arabidopsis. The error bars represent the SD. (D) Quantitative analysis of conserved chloroplast-encoded (left panel), nuclear-encoded thylakoid (middle panel) and cytosolic (right panel) peptides between *Spirodela* and Arabidopsis. Chloroplast-encoded peptides were identified in a separate analysis and conserved peptides were identified and quantified using MS1 precursor intensity and compared with nuclear-encoded proteins (gray). Nuclear-encoded thylakoid and cytosolic peptides were assigned using GO analysis and highlighted in the middle and right panel, respectively. Significance in abundance was called at $P<0.05$, FC>2× ($n=3$).

7498 and Arabidopsis using Z-stacks of confocal images from a fully expanded frond of Sp7498 in comparison with a mature Arabidopsis leaf. As shown in Fig. 5B, *S. polyrhiza* 7498 frond cells displayed a higher number of smaller chloroplasts than Arabidopsis, but the total intracellular volume of chloroplast was nearly equal between the two species (Fig. 5B, C). When we quantified the abundances of 61 chloroplast-encoded (left panel) and 160 thylakoid-targeted (middle panel) peptides in common between the two species (Fig. 5D), a strong and consistent increase in chloroplastic protein content was seen for *S. polyrhiza* 7498, which was not evident when comparing cytosolic peptides (right panel). Consequently, although *S. polyrhiza* 7498 accumulates smaller but higher numbers of chloroplasts, they appear to contain higher protein densities. This increase could translate to a greater photosynthetic efficiency relative to Arabidopsis and thus underpin its more rapid reproduction rate.

Lastly, we highlight that the generation of long-read DNA sequencing data and the shotgun proteome data were integrated into a graduate-level teaching module. For Nanopore long-read DNA sequencing, students in the course loaded the flow cell with a pre-prepared library, initiated sequencing and live base-calling, and then bioinformatically assembled a circular chloroplast live in class with the first few thousand reads. For the shotgun proteome, students isolated proteins from duckweed and Arabidopsis and prepared them for proteomic analysis. Firstly, this approach familiarized the students with protein mass spectrometry and potentially removed barriers to conduct similar experiments to address future biological questions. Secondly, the obtained data were used to introduce large data analysis in a relatable manner and stimulate the students to explore various analysis tools that are publicly available. Altogether, this approach turned out to be a very engaging and creative way to introduce students to various techniques that were until recently unavailable this early in their careers.

## Discussion

Here, we present an independently assembled reference genome for *Spirodela polyrhiza* clone 7498 (Sp7498) using Oxford Nanopore and Phase Genomics Hi-C sequencing (Fig. 1). We draw comparisons with another assembly of the same clone, as well as assess the structural diversity with another clone, Sp9509. Our Sp7498 assembly shows large-scale discrepancies compared with a different assembly of the same line, yet high chromosomal conservation compared with Sp9509. The conserved genome structure between Sp7498 and Sp9509 suggests the two clones are highly similar in terms of chromosome organization.

The observed differences between the genome assembly reported here (Sp7498_HiC) and that described by An *et al.* (2019) for the same Sp7498 clone (Sp7498_PacBio) (Fig. 2) could be explained by several hypotheses. First, genome assemblies are subject to variation depending on sequencing technology, including read lengths and error profiles. Similarly, the choice of long-read genome assembler and subsequent polishing steps can influence assembly outcome. Additionally, the method of long-range scaffolding (e.g. Hi-C, linked reads, BioNano optical mapping) can alter the scaffolding and error-correction of the final assembly. Both the Sp7498_HiC and Sp9509 assemblies were produced using Oxford Nanopore long-reads generated with MinION 9.4.1 flow cells, then assembled and polished similarly using miniasm, medaka, and RACON. Whereas the Sp9509 assembly was scaffolded with BioNano optical mapping, the Sp7498_HiC assembly was scaffolded using Phase Genomics Hi-C and their proprietary Proximo software. On the other hand, the Sp7498_PacBio assembly was generated with PacBio Sequel I reads, assembled with FALCON (Chin *et al.*, 2016), and ordered using BAC-FISH. Oxford Nanopore and PacBio Sequel reads have different biases during nucleotide sequencing, which certainly could impact the full-length assembly and accurate polishing of homopolymer repeats and satellite repeats in particular. The abundance of mismatched and short indels between Sp7498_HiC and Sp7498_PacBio could be explained in part by these sequencing technology biases. Given the high degree of synteny between Sp7498_HiC and Sp9509, and mutual disagreements with large structural variations like chromosome arm inversions in Sp7498_PacBio, it is likely that BAC-FISH scaffolding introduced order and orientation errors in the Sp7498_PacBio genome.

Based on the LTR retrotransposon annotations, there is slight variation in the number of LTR retrotransposons accurately assembled between the genome assemblies (Fig. 3). Identifying LTRs in the original 454 pyrosequencing plus Sanger assembly (Sp7498_v3) (Wang *et al.*, 2014) shows that these short and mid-length reads led to a collapse of retrotransposons, especially recently amplified ones, in the final assembly (Fig. 3). Both PacBio (Sp7498_PacBio) and Oxford Nanopore (Sp7498_HiC and Sp9509) assemblies appear to have corrected this issue. The Sp7498_PacBio PacBio-based assembly resulted in a slightly higher number of retrotransposon annotations, though fairly similar in number and insertion time compared with Sp9509 and Sp7498_HiC. However, it remains possible that this variation between Sp7498_PacBio and Sp7498_HiC reflects true biological variation caused by genetic drift.

A second hypothesis is related to how *S. polyrhiza* plants are clonal, and reproduce exponentially, although rarely (if ever) do they flower in nature or in culture. In the presumed absence of meiotic recombination that normally occurs during sexual reproduction, *S. polyrhiza* plants across their global range exhibit low genetic variation, and intriguingly also low spontaneous mutation rates (Xu *et al.*, 2019). The estimation by Xu *et al.* of the *S. polyrhiza* mutation rate is similar to that of eubacteria and unicellular eukaryotes. Some of the mismatch, indel, and structural variation content between Sp7498_HiC and Sp7498_PacBio could be explained by these infrequent spontaneous mutations that have since accumulated between

different laboratories' stocks of Sp7498 that was originally collected in Durham, NC, USA by Elias Landoldt, compared with Sp9509 collected in 2002 from a population in Lotschen, Stadtroda, Germany.

Overall, the differences that we pinpoint across these three genomes (Sp7498_HiC, Sp7498_PacBio, and Sp9509) are relatively minor, depending on the intended usage of the genome. For instance, the set of gene annotations are highly similar in number, meaning that the gene space of the assembly is largely the same. Each of these three genomes highlights that developing high-accuracy and syntenic contigs for *S. polyrhiza* is relatively simple using either PacBio or Oxford Nanopore long-reads. The low heterozygosity and low repeat content of the genome is likely responsible for yielding such long contigs regardless of long-read sequencing technology used. However, the choice of long-range scaffolding technology is the major determinant of the quality of assembly order and orientation. In this case, both BioNano optical mapping and Hi-C scaffolding resulted in similar chromosome-scale scaffolds between Sp9509 and Sp7498_HiC, both of which differed in the ordering and orientation of several large chromosomal pieces of Sp7498_PacBio.

Global analysis of the *Spirodela* proteome revealed a high abundance of proteins involved in generating energy in *Spirodela* (Figs 4, 5). In addition to comparing the proteome with other species, this model system is very well suited to analysing proteomic responses to various environmental stress conditions like heat and heavy metal stress amongst others. Intriguingly, much of the highly expressed portion of the proteome is derived from chloroplasts or associated with energy production. Comparisons of expanded leaf chloroplast density, size, and arrangement between *Spirodela* and Arabidopsis show a similar overall area of chloroplasts, but the organelles are smaller and more abundant in *Spirodela*. This is perhaps related to the two-dimensional growth pattern of duckweeds, rapid proliferation, and the need to harvest light from a single plane. Surprisingly, *S. polyrhiza* 7498 had increased levels of chloroplast proteins per fresh weight, implying that the internal volume of these organelles has higher protein densities. It is unknown if this is a common feature of duckweeds, but could certainly be a key factor in their rapid growth given the high abundance of chloroplast-derived proteins necessary for photosynthesis. In addition, we note that the chloroplasts in a related duckweed species, *Lemna trisulca*, are mobile in response to heavy metals (Samardakiewicz *et al.*, 2015), suggesting that chloroplast dynamics in duckweeds favor the ability to quickly reorganize in response to environmental stress. Taken together, as the ability to modulate and control RuBisCO and other plastid-associated proteins appears central to crop improvement (Parry *et al.*, 2003), the relatively high expression of these proteins in duckweeds might serve as a foundation for exploring the genetic, developmental, physiological, and regulatory mechanisms that underlie enhanced photosynthetic potential. However, there are many caveats that could also explain the variation in protein content and chloroplast organization and size between *S. polyrhiza* and Arabidopsis, including a large phylogenetic distance, gene content variation, paleopolyploidy, chloroplast genome variation, developmental stage variation, ecological variation, and more. This phylogenetically distant comparison is most useful for the transfer of gene annotations and functional pathways from the model species Arabidopsis, and cannot be interpreted as a direct comparison to infer relative photosynthetic efficiency without more detailed physiological experiments.

Given the portability and speed of the Oxford Nanopore MinION platform, the computational ease of generating highly contiguous genome assemblies, the ease of generating in-depth, quantified proteomic datasets, and the natural abundance of duckweed species across all continents except Antarctica, there is a substantial opportunity to bring low-cost genome sequencing and proteomics of duckweed into classroom settings that culminate in valuable, publishable discoveries. The Oxford Nanopore MinION sequencer has been successfully deployed in undergraduate and graduate education across disciplines (Zaaijer *et al.*, 2016; Zeng & Martin, 2017, Preprint), and we expect that ongoing updates to library preparation, pore lifespan, and sequencing devices will continue to drive the proliferation of classroom-based education tools. While portable nanopore sequencing devices certainly bring cutting-edge sequencing resources to laboratory and field scientists, they perhaps more importantly democratize the ability for valuable science to be performed by students from middle school to graduate school.

## Supplementary data

The following supplementary data are available at *JXB* online.

Fig. S1. Identification and mapping of telomere repeats (5′-TTTAGGG-3′) across all 20 chromosomes of Sp7498_HiC using CoGE-BLAST.

Fig. S2. GC content across three *Spirodela polyrhiza* genomes.

Fig. S3. Comparison of raw protein extracts from three replicates of *Spirodela polyrhiza* 7498 and Arabidopsis Col-0.

Table S1. Raw MS predicted proteome quantification for *Spirodela polyrhiza* 7498.

Table S2. Raw MS predicted proteome quantification for Arabidopsis Col-0.

Table S3. *Spirodela polyrhiza* functional gene annotation.

Table S4. *Spirodela polyrhiza* detected chloroplast-derived protein quantification.

Table S5. Arabidopsis detected chloroplast-derived protein quantification.

Table S6. Chloroplast-derived protein annotation.

## Acknowledgements

We thank the Department of Biology at Washington University in St Louis for funding the Oxford Nanopore genome sequencing and shotgun

proteome of line Sp7498. AH was supported by a post-doctoral fellowship from the NSF National Plant Genome Initiative (IOS-1611853). FM was supported by a grant from the NSF Plant Genome Research Program (IOS-1840687). We also acknowledge imaging support from the Advanced Bioimaging Laboratory at the Danforth Plant Science Center and usage of the Leica SPX-8 acquired through a NSF Major Research Instrumentation grant (DBI-1337680).

## Author contributions

AH, FM, RDV, BCM, and TPM conceptualized the study. AH, NB, KE, RE, EM, and KM sequenced the genome. FM, NB, KE, RE, EM, and KM generated the proteome. AH and TPM assembled the genome and performed comparative genomic analyses. FM annotated the proteome and performed the proteomic analyses. KC performed confocal microscopy and chloroplast area calculations. AH and FM wrote the manuscript with help from RDV, TPM, and BCM.

## Data availability

The Sp7498_HiC genome browser is available at CoGE (https://genomevolution.org/coge/GenomeView.pl?embed=&gid=55812). All of the assemblies described here are also found at spirodelagenome.org. The raw Oxford Nanopore reads and Hi-C Illumina reads are available at BioProject PRJNA631098. The raw sequence files for the MS datasets are available in the ProteomeXchange database under the submission number PXD17093 within the PRIDE repository (http://proteomecentral.proteomexchange.org/cgi/GetDataset). Accession numbers and unmodified MS output files from the nuclear and chloroplast encoded proteome are available in Supplementary Tables S1–S6.

## References

**An D, Zhou Y, Li C, *et al.*** 2019. Plant evolution and environmental adaptation unveiled by long-read whole-genome sequencing of *Spirodela*. Proceedings of the National Academy of Sciences, USA **116**, 18893–18899.

**Bog M, Sowjanya Sree K, Fuchs J, Hoang PTN, Schubert I, Kuever J, Rabenstein A, Paolacci S, Jansen MAK, Appenroth K.** 2020. A taxonomic revision of *Lemna* sect. *Uninerves* (Lemnaceae). Taxon **69**, 56–66.

**Chin CS, Peluso P, Sedlazeck FJ, *et al.*** 2016. Phased diploid genome assembly with single-molecule real-time sequencing. Nature Methods **13**, 1050–1054.

**Ellinghaus D, Kurtz S, Willhoeft U.** 2008. *LTRharvest*, an efficient and flexible software for *de novo* detection of LTR retrotransposons. BMC Bioinformatics **9**, 18.

**Eng JK, McCormack AL, Yates JR.** 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. Journal of the American Society for Mass Spectrometry **5**, 976–989.

**Gurevich A, Saveliev V, Vyahhi N, Tesler G.** 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics **29**, 1072–1075.

**Hoang PNT, Michael TP, Gilbert S, Chu P, Motley ST, Appenroth KJ, Schubert I, Lam E.** 2018. Generating a high-confidence reference genome map of the Greater Duckweed by integration of cytogenomic, optical mapping, and Oxford Nanopore technologies. The Plant Journal **96**, 670–684.

**Li H.** 2016. Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. Bioinformatics **32**, 2103–2110.

**Li H.** 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics **34**, 3094–3100.

**Littlejohn GR, Mansfield JC, Christmas JT, Witterick E, Fricker MD, Grant MR, Smirnoff N, Everson RM, Moger J, Love J.** 2014. An update: improvements in imaging perfluorocarbon-mounted plant leaves with implications for studies of plant pathology, physiology, development and cell biology. Frontiers in Plant Science **5**, 140.

**Lyons E, Freeling M.** 2008. How to usefully compare homologous plant genes and chromosomes as DNA sequences. The Plant Journal **53**, 661–673.

**McLoughlin F, Augustine RC, Marshall RS, Li F, Kirkpatrick LD, Otegui MS, Vierstra RD.** 2018. Maize multi-omics reveal roles for autophagic recycling in proteome remodelling and lipid turnover. Nature Plants **4**, 1056–1070.

**Michael TP, Bryant D, Gutierrez R, *et al.*** 2017. Comprehensive definition of genome features in *Spirodela polyrhiza* by high-depth physical mapping and short-read DNA sequencing strategies. The Plant Journal **89**, 617–635.

**Olsen JL, Rouzé P, Verhelst B, *et al.*** 2016. The genome of the seagrass *Zostera marina* reveals angiosperm adaptation to the sea. Nature **530**, 331–335.

**Parry MA, Andralojc PJ, Mitchell RA, Madgwick PJ, Keys AJ.** 2003. Manipulation of Rubisco: the amount, activity, function and regulation. Journal of Experimental Botany **54**, 1321–1333.

**Samardakiewicz S, Krzeszowiec-Jeleń W, Bednarski W, Jankowski A, Suski S, Gabryś H, Woźny A.** 2015. Pb-induced avoidance-like chloroplast movements in fronds of *Lemna trisulca* L. PLoS One **10**, e0116757.

**Silva JC, Gorenstein MV, Li GZ, Vissers JP, Geromanos SJ.** 2006. Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition. Molecular & Cellular Proteomics **5**, 144–156.

**Tian T, Liu Y, Yan H, You Q, Yi X, Du Z, Xu W, Su Z.** 2017. agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Research **45**, W122–W129.

**Tyanova S, Temu T, Cox J.** 2016. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nature Protocols **11**, 2301–2319.

**Urbanska KM, Crawford DJ, Appenroth K, Les DH.** 2013. Elias Landolt (July 21, 1926–April 1, 2013). Aquatic Botany **111**, A1–A2.

**Vaser R, Sović I, Nagarajan N, Šikić M.** 2017. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Research **27**, 737–746.

**Wang W, Haberer G, Gundlach H, *et al.*** 2014. The *Spirodela polyrhiza* genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. Nature Communications **5**, 3311.

**Wang W, Kerstetter RA, Michael TP.** 2011. Evolution of genome size in duckweeds (Lemnaceae). Journal of Botany **2011**, 570319.

**Xu S, Stapley J, Gablenz S, Boyer J, Appenroth KJ, Sree KS, Gershenzon J, Widmer A, Huber M.** 2019. Low genetic variation is associated with low mutation rate in the giant duckweed. Nature Communications **10**, 1243.

**Zaaijer S, Columbia University Ubiquitous Genomics 2015 class, Erlich Y.** 2016. Using mobile sequencers in an academic classroom. eLife **5**, e14258.

**Zeng Y, Martin CH.** 2017. Oxford Nanopore sequencing in a research-based undergraduate course. BioRxiv, 227439. [Preprint].

**Ziegler P, Adelmann K, Zimmer S, Schmidt C, Appenroth KJ.** 2015. Relative in vitro growth rates of duckweeds (Lemnaceae) – the most rapidly growing higher plants. Plant Biology **17 (Suppl 1)**, 33–41.