GENOMICS AND TRANSCRIPTOMICS OF THE GREATER DUCKWEED, *SPIRODELA POLYRHIZA*, A MODEL FOR AQUATIC BIOLOGY

By

WENQIN WANG

A Dissertation submitted to the

Graduate School-New Brunswick

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Plant Biology

Written under the direction of

Dr. Joachim Messing

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

January 2014

ABSTRACT OF THE DISSERTATION

Genomics and transcriptomics of the Greater Duckweed, *Spirodela polyrhiza*, a model for aquatic biology

By WENQIN WANG

Dissertation Director:

Dr. Joachim Messing

My thesis provides the first whole genome analysis of an aquatic plant, *Spirodela polyrhiza* and a reference genome for a new order among the monocotelydonous angiosperms. The *Lemnoideae* belong to the order of the Alismatales and are commonly known as duckweeds, the smallest, fastest growing, and simplest of aquatic plants, thus telling them apart is not a trivial task. Whereas a simple and accessible protocol has been established for land plants by the Consortium for the Barcode of Life with seven universal DNA barcoding markers, we found that *atpF-atpH* noncoding spacer is the most promising marker for duckweed species-level identification. Furthermore, our assembly and annotation of the Spirodela chloroplast and mitochondrial genomes open an opportunity of population-level classification.

A key to our understanding of the evolution of a species and its potential use is the gene content of the organism. Therefore, we sequenced *Spirodela polyrhiza* 7498 that has one of the smallest genomes with 158 Mb within this subfamily of species. The

genome contains 19,623 predicted protein-coding genes, sharing a total of 8,255 common gene families with Arabidopsis, tomato, banana, and rice despite a significantly reduced gene number. Reduced gene families and missing genes reflect changes consistent with its compact and reduced morphogenesis or forever-young life style, aquatic suspension, and suppression of juvenile-to-adult transition.

Spirodela exhibits a remarkable phenotypic plasticity to adapt to cold weather in winter. We identified and functionally annotated 362 differentially expressed genes, which open a major step towards understanding the molecular network underlying vegetative frond dormancy. Moreover, the expression data for lipid and starch biosynthesis together with the turion-specific transcriptional genes from our RNA-Seq data could be ideal targets to develop duckweeds into oil crops.

Thanks to its unique and fascinating biology, applications of duckweed in water remediation and as a renewable energy source are predicted to have a bright future. The genome sequence of Spirodela provides the first step to identify, understand, and improve relevant traits for specific target applications.

Acknowledgement

I thank my advisor, Prof. Joachim Messing, for believing in me and providing me with constant support and guidance. His sharp intuitive and broad knowledge in science especially sequencing and genomics really impresses me. He trains me how to think scientifically and independently. He encourages me how to draft, format and edit manuscripts. He offers me to present my data in meeting and conference getting involved in rapidly moving areas of biological research. It also gives me chance to meet and chat with my former advisor, Dr. Todd Michael. I appreciate he took me as his first student and influenced in my early training in genome size measurement and bioinformatics.

I would like to thank my committee members: Hugo Dooner, Pal Maliga and Chunguang Du. I have had the wonderful privilege to chat, discuss and work with all of them since I physically stay in the junction of two labs, and each has helped in my professional development. I would like to express my gratitude to Dr. Dooner for giving great advices for my research and class, and allowing me to use lab space. I like to discuss with Dr. Maliga about organelle genomes, ask him everything when I meet problems associated with chloroplast in my analysis of organelle assembly and RNA-Seq. I enjoy the critical thinking from Dr. Du's group for the algorithm of transposon prediction and functional network analysis in maize. My conclusion is that being a real scientist really reflects everywhere and pursuit perfection is endless.

My research being interdisciplinary, I am fortunate to have received help from people with different expertise. Thanks to Qinghua Wang, she guides me to construct the Spirodela BAC library. Limei He arranges the best bench and working space for me. I enjoyed the time to discuss bioinformatics analysis with Jianhong Xu, David

Sidote, Kerry Lutz, Anna Zdepski, Jun Huang, Csanad Gurdon, Gregory Thyssen, Yaping Feng, Wei Zhang, Mary Galli over the years. I also owe special thanks for Randall Kerstetter, Brian Gelfand, Mark Diamond and Dibyendu Kumar to patiently explain the experimental setup and to conduct my next-generation sequencing. The research in maize and Sorghum from Yubin Li, Tarinee Tungsuchat, Mihai Miclaus, Martin Calvino, Nelson Garcia and Jose Planta really broaden my view. Jennifer Ayer and Jennifer Tirrito are always there to help me. I'm thankful to all other members from the groups of Dr. Messing, Dr. Dooner, Dr. Maliga, Dr. Dong, Dr. Gallavotti at Waksman, Dr. Huang and Dr. Belanger from Plant Biology and Pathology Department for their advice, support, and friendship. I also thank all collaborators from Dr. Klaus Mayer group (Georg Haberer, Heidrun Gundlach, Christine Gläßer, Thomas Nussbaumer), JGI (Jeremy Schmutz), Dr. Mingchen Luo and Dr. John Vogel.

I thank my husband for his love and good suggestions for my projects. Thanks to my dear son and sweet twin girls, they bring a lot of laugh and joy into my life, giving me strength and inspiration when I needed it the most. I am heavily indebted to my parents and my mother-in-law. Without their supports, I could never have succeeded.

Parts of my work including genome size, DNA barcode, chloroplast and mitochondrial genome, and starch synthesis at turion development have already been published in Journal of Botany, BMC Plant Biology and PLOS ONE.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

CHAPTER 0 Overview

CHAPTER 1 DNA barcode

CHAPTER 2 DNA sizing

CHAPTER 3 Chloroplast genomics

CHAPTER 4 Mitochondrial genomics

CHAPTER 5 Nuclear genomics

CHAPTER 6 Starch synthesis at turion formation

CHAPTER 7 Expression profiling with onset of dormancy

LIST OF TABLES

CHAPTER 1 DNA barcode

CHAPTER 2 DNA sizing

CHAPTER 3 Chloroplast genomics

CHAPTER 4 Mitochondrial genomics

CHAPTER 6 Starch synthesis at turion formation

CHAPTER 7 Expression profiling with onset of dormancy

CHAPTER 0 Overview: duckweed biology, genomics and transcriptomics

0.1 Abstract

Duckweeds have been studied at the botanic and biochemical level. However, their reduced morphology and constant environmental selection have not been subjected to molecular analysis. Therefore, my first but preferential task has been to screen a large duckweed collection from the late Dr. Landolt of Switzerland for genome size variation and to develop DNA bar-coding markers in order to replace the previously ambiguous classical systematics. Comparing different species and ecotypes, I found that *atpF-atpH* was the preferred barcoding marker due to its ease of amplification and sufficient polymorphism.

Furthermore, I was able to choose the species with the smallest genome size for whole-genome sequencing. This was *Spirodela polyrhiza* with 158 Mb. For sequencing genomes and predicting the gene content of Spirodela, I took two approaches. I used a next generation sequencing platform to sequence total plant DNA and assembled the organelle genome sequences with a dosage-sensitive algorithm. I also provided nuclear genomic DNA for community-service sequencing center at the Department of Energy Joint Genome Institute (JGI). To reach chromosome-size pseudo-molecules from the JGI data, I used the genomic DNA to construct a 10X Spirodela bacterial artificial chromosome (BAC) library, which was DNA fingerprinted with a high-throughput method by Dr. Luo at the University of California at Davis. The BACs were also end-sequenced with long reads using the traditional ABI capillary sequencers so that a physical map could be aligned with contigs from sequence assemblies. A key finding of

the analysis of the chromosome-size DNA sequences is that the reduced gene families and missing genes are consistent with its compact morphogenesis, aquatic suspension and suppression of juvenile-to-adult transition. The contraction or expansion of special gene families provides new information of how to design new transgenic duckweed for industrial applications like animal feeding, wastewater treatment, and biofuel.

In respect to gene expression, I provided cDNAs for transcriptome sequencing at JGI to aid in the annotation of the gene content. I also investigated one example of gene expression involved in starch biosynthesis switching from growth to dormant phase of the life cycle. In addition, RNA deep sequencing was performed at the Waksman Genomics Facility and I was able to obtain detailed information for dormancy related gene expression.

Rapid advances in sequencing technologies will continue to promote a proliferation of genome sequences for additional ecotypes as well as other duckweed species. Here, we review the current status of genome research in duckweed for my projects.

0.2 Morphology, life cycle and applications

Lemmnoideae, called duckweeds, are aquatic plants seen on water surfaces broadly distributed around the world. They are remarkably adaptive in aquatic environments and exhibit an extreme compact structure and a fast clonal growth [1]. They include five genera of Spirodela, Landoltia, Lemna, Wolffiella and Wolffia, and a total of 38 species [2]. The leaf-like organ, called frond is the simple version of the

combination of leaf and stem. The whole plant size is extremely small, ranging from 1 to 10 mm (Figure 0.1).

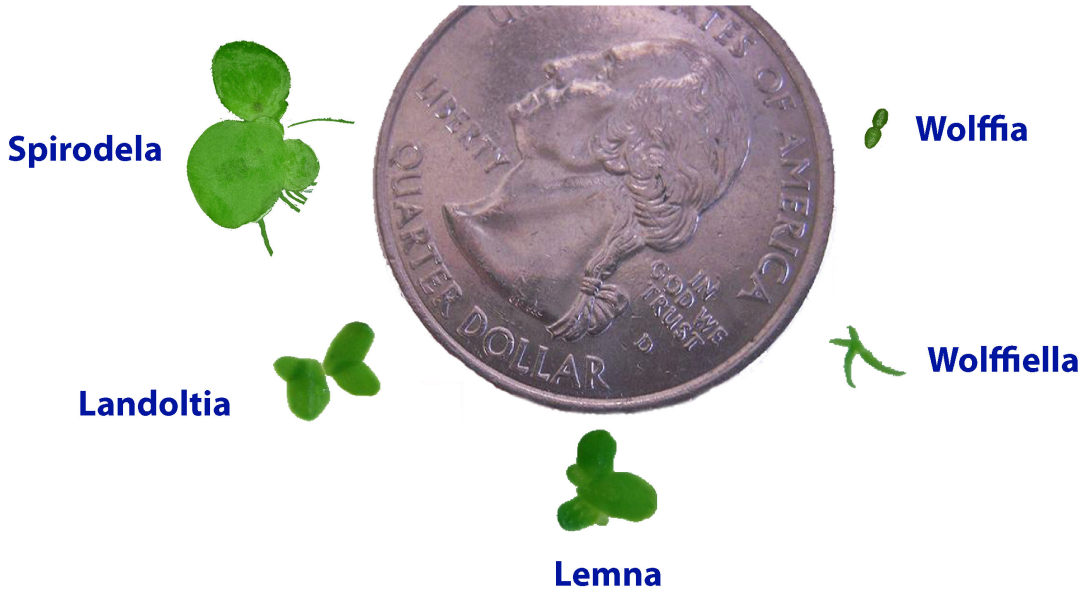


Figure 0.1 Five genera of duckweeds.

The relative size of Spirodela, Landoltia, Lemna, Wolffiella and Wolffia compared to an American quarter. The figure was modified from [3].

Duckweeds have a very unique life cycle with seasonal change [4]. Fronds, as a growing state, engaging in photosynthesis grow fast and mainly collect biomass under optimal conditions in spring and summer. When it is getting cold and they are deprived of nutrients at the end of growing season in the fall, fronds shift to a dormant phase termed turions (Figure 0.2). Turions allow duckweeds to endure harsh cold winter and adapt to a life worldwide [5, 6]. Especially, the facts of abundant starch in turions as well as little amount of lignin become a luminous point for biofuel use either by converting starch into ethanol or redirecting carbon flow into oil [7].

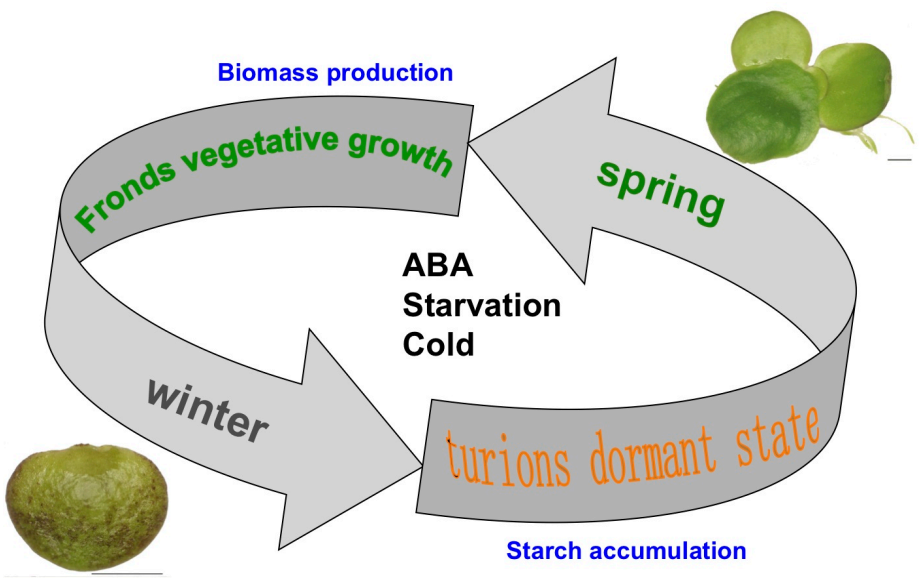


Figure 0.2 Life cycle of Spirodela.

Under the optimal conditions of nutrition and temperature, Spirodela does fast vegetative growing by the format of fronds, while under the poor conditions such as stress of ABA, starvation and cold, fronds are switched into dormant state of turions that accumulate starch and sink to the bottom of the ponds

For long time duckweed has been used to study photosynthesis because of its ability to convert sun energy into biomass efficiently. They are also used as monitoring device for measuring water quality by the Environmental Protection Agency because they take up nutrients directly out of the water [8]. With fast biomass accumulation and the need of little amount of lignin to support their floating body, it has attracted a lot of attention in industrial applications, such as wastewater treatment and biofuel processing [9-11].

0.3 DNA barcode

Their miniature plant size, highly reduced morphology and rare flowering impose a big challenge to distinguish the nearly 38 species. Some progress had been made with

the analysis of metabolites like flavonoids, anthocyanins, and allozymes in combination of morphological traits [2, 12]. However, the high resolution of DNA polymorphism, permits us to use the chloroplast genes of *rbcL* and *matK* together with introns of *trnK* and *rpl16* in phylogenetic and systematic analysis of these species [2]. The Consortium for the Barcode of Life (CBOL) plant-working group proposes seven leading candidate barcoding markers. Four plastid-coding genes are *rpoB*, *rpoC1*, *rbcL* and *matK* and three noncoding spacers are *atpF-atpH*, *psbK-psbI* and *trnH-psbA* [13]. Based on these markers, we evaluated DNA sequence polymorphism in 97 ecotypes from 31 duckweed species. We found that *atpF-atpH* appears to be the most promising DNA barcode marker based on its reliable amplification, straightforward sequence alignment, and rates of DNA variation between species and within species [14] (See Chapter 1).

0.4 DNA sizing

To estimate the genome size and to provide a basic reference for a duckweed genome sequence project, we measured the DNA content for 115 different ecotypes of 23 duckweed species by flow cytometry (FCM). Surprisingly, there is a continuous increase of DNA content that parallels a morphological reduction in size, ranging from 150 Mb in Spirodela to 1,881 Mb in Wolffia (Figure 0.3). There is a significant intraspecific variation in the genus Lemna. However, no such variation was found in the genera of Spirodela and Landoltia. With few samples for the same species in Wolffiella, and Wolffia, it is unclear if the intraspecific variation exists or not [3] (See Chapter 2).

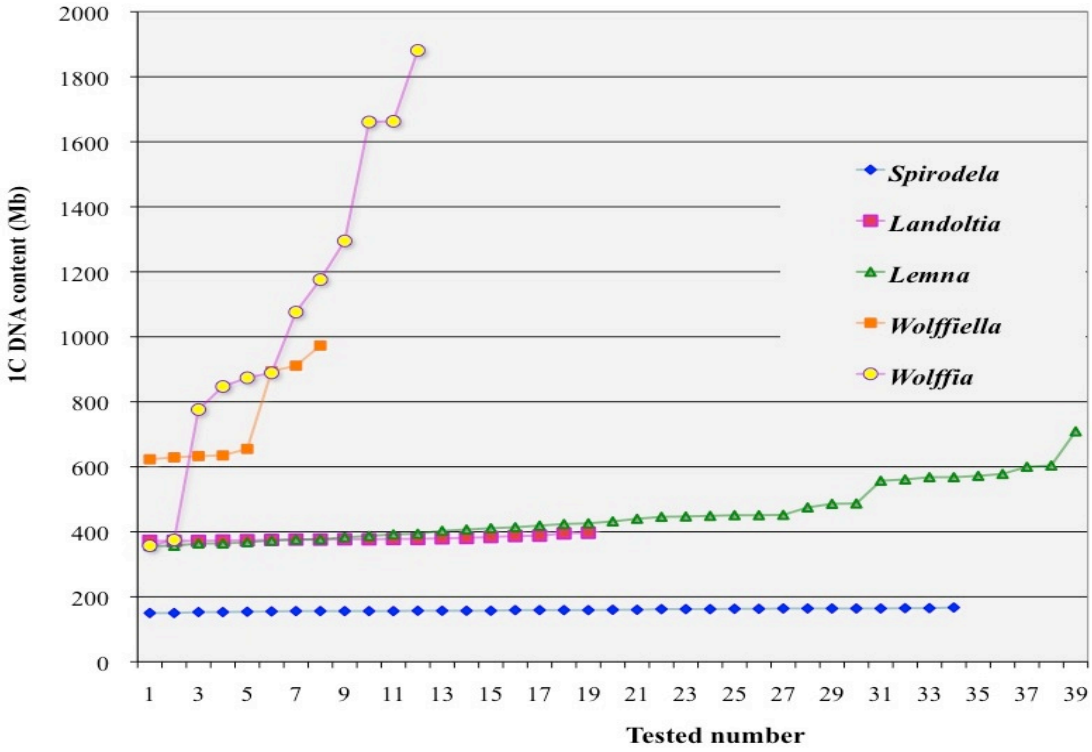


Figure 0.3 Genome sizes of duckweeds.

The x-axis shows the total of tested number for each genus and y-axis shows the genome size in Mb. The tested species are *Spirodela polyrhiza, Landoltia punctata, Lemna aequinoctialis, Lemna valdiviana, Lemna minor, Lemna gibba G-3, Lemna trisulca, Lemna japonica, Lemna obscura, Wolffiella hyaline, Wolffiella gladiata, Wolffiella lingulata, Wolffia brasiliensis, Wolffia borealis, Wolffia australiana, Wolffia microscopica, Wolffia globosa, Wolffia angusta, Wolffia neglecta, Wolffia elongata, Wolffia Columbiana, Wolffia cylindracea, Wolffia arrhiza.* The figure was modified from [3].

0.5 Chloroplast genome

Ecotypes refer to the population of the same species but from different geographical locations. Whereas we find that the same species for different ecotypes has undifferentiated morphology and almost identical DNA barcode sequence of *atpF-atpH*, they still could represent distinct physiological attributes [15]. For example, the ecotypes of *Spirodela polyrhiza* have a very broad range of turion yield from 0.22 to 5.9 times of the vegetative frond. Previous findings also indicate that phenotypic differences are

probably inherited as the result of DNA mutations [16]. Therefore, in an effort to screen and isolate suitable ecotypes with high starch, protein or rapid growth, or with heavy metal tolerance for the application of animal food, biofuel, wastewater treatment, it is a prerequisite to differentiate ecotypes as well [17]. It has so far proved to be difficult to genotype ecotypes only based on limited DNA markers. Additional polymorphism is needed to delineate ecotypes of the same species. Thus, the full plastid genome becomes the best avenue due to their highly conserved sequence but increased resolution and informative sequence variation. Together with the fast improvement of next-generation sequencing technology, it is feasible to get multiple plastid genomes simultaneously using the multiplex bar-coded library system [18].

We have sequenced total frond DNA from Spirodela using SOLiD sequencing platform. It generates about 1,000-times coverage of chloroplast, 100-times of mitochondria and 10-times of nuclear genome at the same time in a quarter slide. Therefore, the plastid genome with its copy number could be *de novo* assembled into contigs by setting up higher coverage threshold in order to computationally filter the nuclear and mitochondrial reads. Although the chloroplast genome is conserved in gene number and organization with respect to other species, higher nucleotide substitutions, abundant deletions and insertions occur in non-coding regions [19], facilitating its utilization for ecotype identification and evolutionary studies. The complete organelle genomes in duckweeds also provide a wealth of information to understand photosynthesis and energy utilization especially under aquatic environments (See Chapter 3).

0.6 Mitochondrial genome

The same plant DNA sequences that have been used to assemble chloroplast genomes can also be used to assemble the *Spirodela polyrhiza* mitochondrial genome by adjusting the parameters for the relative copy number of mitochondria. Furthermore, the assembled chloroplast sequence can be used to filter chloroplast reads computationally prior to the assembly of the mitochondrial sequences. The Spirodela mitochondrial genome is the most compact among monocots with 228,493 bp. It shares the conserved protein-coding genes with other monocots, but after eliminating genes, introns, ORFs, and plastid-derived DNA, nearly four-fifths of the genome is of unknown origin and function [20] (See Chapter 4).

0.7 Nuclear genome

Although a number of genomes of monocot species in particular within the grass family [21-24] and one example outside, banana [25] have been sequenced and annotated, a high quality genome sequence of a different order than the two above will be invaluable for gene discovery and evolutionary analysis of basal monocot species. Therefore, the genome sequence of a duckweed species cannot rely on a pre-existing reference genome for assembly and has to be sequenced based on a physical map that is derived independently from sequencing. Furthermore, to avoid a genome with a high content of repetitive DNA, such as retrotransposons, *Spirodela polyrhiza* 7498 with the small genome size of 158 Mb was selected for such a reference genome [3].

A critical factor in keeping sequencing cost down has been the elimination of the huge number of DNA preparations for sequencing with the BAC-by-BAC approach. This

was first illustrated for viral genomes, when Sanger's group sequenced phiX174 fragment-by-fragment [26] and Messing's group Cauliflower Mosaic virus with DNA libraries [27]. This concept of parallelization of sequencing reactions has greatly been accelerated with next generation sequencing platforms that permit us to sequence DNA ligations without the separation of individual templates by cloning [28]. Therefore, we have used in the case of Spirodela the "454" platform, a compromise in throughput and read length. In these ligation reactions, genomic DNA is size-selected so that DNA fragments can be sequenced from both ends [29]. When distances between two sequences are known, assembly of larger contigs can be facilitated. We also sequenced the ends of BAC clones with traditional methods to emphasize read length over longer physical, but linked distances. These BACs were also fingerprinted and assembled into a physical map, which allowed us to align assembled sequences along this map [30]. Given such alignments with the physical map, the assembled scaffolds have been ordered into 32 chromosome-sized pseudomolecules.

A wealth of information on plant biology can be drawn from comparative analysis by examine the expansion or contraction of gene families with other known genomes. The Spirodela genome has only 19,623 predicted protein-coding genes, 28% less than dicotyledonous *Arabidopsis thaliana* and 50% less than monocotyledonous rice. The Spirodela genome contains very similar patterns of orthologous gene sets in comparison to four representative species (Arabidopsis, tomato, banana, and rice), sharing a total of 8,255 common gene families despite a significantly reduced gene number. Reduced gene families and missing genes reflect changes consistent with its compact and reduced morphogenesis or forever-young life style, aquatic suspension, and promotion of

juvenile-to-adult transition. It is based on skipping the development of organs like roots, stems, and flowers. This is reflected in reduced gene numbers for cellulose/lignin biosynthesis, expansins, MADS-box factors, miRNA172, and miRNA169.

Despite a genome-wide reduction in gene number, copy numbers of certain gene families are retained or even amplified in Spirodela. For example, there are up to four times more copies of glutamate synthase for nitrogen absorption in Spirodela compared to Arabidopsis and rice. This could be the reason that Spirodela requires the efficient usage of nutrients matching its high growth and thus has been successfully exploited for wastewater remediation because of its ability to remove excess nitrogen from polluted water. Another good example is miRNA156 gene also with increased copy number known for suppressing the juvenile-to-adult transition (See Chapter 5).

0.8 Starch synthesis at turion formation

To understand the starch accumulation in Spirodela turion development, we used a plant hormone of abscisic acid (ABA) to initiate the process. We investigated ultra-structural characteristics and starch content both in fronds and turions along the time course. Turions, as a dormant state, were rich in anthocyanin pigmentation. When checked under transmission electron microscopy (TEM), turion cells exhibit shrunken vacuoles, smaller intercellular space, and abundant starch granules surrounded by thylakoid membranes (Figure 0.4). They could even collect more than 60% starch in dry mass after two weeks of ABA treatment. One of the key genes involved in starch biosynthesis is ADP-glucose pyrophosphorylases including both large (APLs) and small subunits (APSs). When we quantitatively measured the level of APL expression by

qPCR, we found they were developmentally dependant. APL2 and APL3 were highly expressed in earlier stages of turion development, while APL1 expression was reduced throughout turion development (See chapter 6).

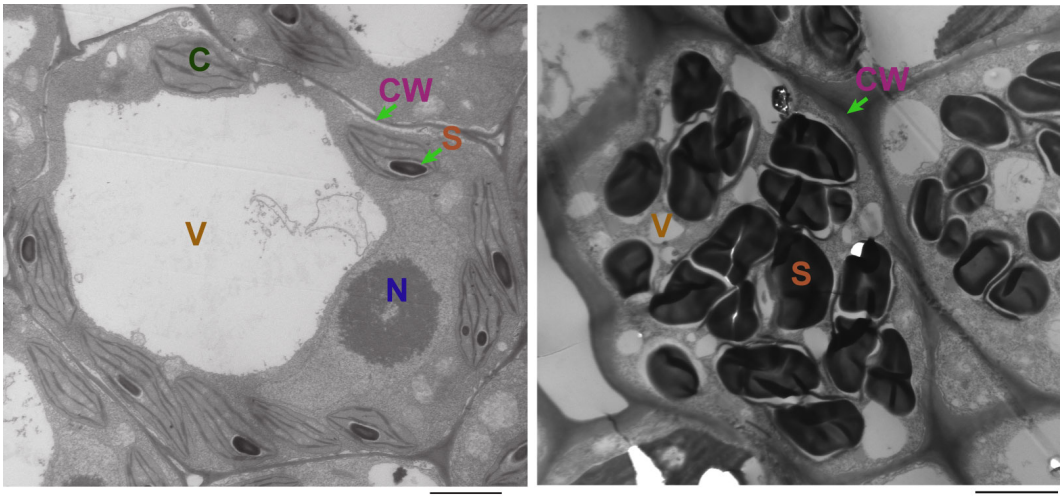


Figure 0.4 Comparsion between frond and mature turion by TEM.

A. A frond cell with a big vacuole and well-shaped chloroplasts but few and less starch granules, Bar = 2 μ m; B. A turion cell with thick cell wall and abundant starch granules, Bar = 2 μ m; Abbreviation: cell wall (CW), chloroplast (C), starch granule (S) and nucleus (N).

0.9 Expression profiling with onset of dormancy

To help annotation and validate genome assembly, the whole-genome transcriptome has been sequenced from RNA pools derived from plants grown under different conditions, in order to maximally capture a representative gene set. Eventually, 379,502 assembled ESTs with high quality were aligned back to the genome to delimit gene regions and also validate the computationally predicted gene content of Spirodela.

Access to the annotated gene sets allows us to addresses significant biological questions regarding environmental adaptation and ecology. As described above,

Spirodela undergoes dormancy without seeds by forming turions. Therefore, expression profiling of this transition not only offers insight into the life-cycle of duckweeds, but also dormancy in the absence of reproduction. To investigate this profile, RNA-Seq data from frond and developing turion was generated with the SOLiD 5500 platform and the cDNA sequence reads were mapped back to the annotated genes. The result shows that the 208 up-regulated genes are associated with signal transduction, seed dehydration, carbohydrate and secondary metabolism, and senescence. On the other side, the 154 down-regulated genes are responsible for rapid growth and biomass accumulation through histone synthesis that packages with DNA after its replication, protein synthesis and carbon fixation. Particularly, we highlighted three turion-specific genes. The understanding of the mechanism of turion formation is helpful to manipulate bud and seed dormancy in agricultural and horticultural fields. These results provide a valuable genomic resource for duckweed and pave the way for the further molecular biological studies and the application of duckweed as a bioenergy crop (See Chapter 7).

0.10 References

1. Landolt E: **The family of Lemnaceae - a monographic study, Vol 1**, vol. 1: Veroffentlichungen des Geobotanischen Institutes der Eidgenossischen Technischen Hochschule, Stiftung Rubel; 1986.
2. Les DH, Crawford DJ, Landolt E, Gabel JD, Kimball RT: **Phylogeny and Systematics of Lemnaceae, the Duckweed Family**. *Systematic Botany* 2002, **27**(2):221-240.
3. Wang W, Kerstetter R, Michael T: **Evolution of genome size in duckweeds (Lemnaceae)**. *Journal of Botany* 2011(Special Issues).
4. Appenroth K-J, Nickel G: **Turion formation in Spirodela polyrhiza: The environmental signals that induce the developmental process in nature**. *Physiologia Plantarum* 2009, **138**(3):312-320.
5. Appenroth KJ: **Co-action of temperature and phosphate in inducing turion formation in Spirodela polyrhiza (Great duckweed)**. *Plant, Cell & Environment* 2002, **25**(9):1079-1085.

6. Appenroth KJ, Teller S, Horn M: **Photophysiology of turion formation and germination in Spirodela polyrhiza**. *Biologia Plantarum* 1996, **38**(1):95-106.
7. Wang W, Messing J: **Analysis of ADP-glucose pyrophosphorylase expression during turion formation induced by abscisic acid in Spirodela polyrhiza (greater duckweed)**. *BMC Plant Biology* 2012, **12**(1):5.
8. Brain RA, Solomon KR: **A protocol for conducting 7-day daily renewal tests with Lemna gibba**. *Nat Protoc* 2007, **2**(4):979-987.
9. Cheng JJ, Stomp AM: **Growing duckweed to recover nutrients from wastewaters and for production of fuel ethanol and animal feed**. *CLEAN - Soil, Air, Water* 2009, **37**(1):17-26.
10. Cheng J, Bergmann BA, Classen JJ, Stomp AM, Howard JW: **Nutrient recovery from swine lagoon water by Spirodela punctata**. *Bioresource Technology* 2002, **81**(1):81-85.
11. Stomp A-M, El-Gewely MR: **The duckweeds: A valuable plant for biomanufacturing**. In: *Biotechnology Annual Review.* vol. Volume 11: Elsevier; 2005: 69-99.
12. Les D, Landolt E, Crawford DJ: **Systematics of theLemnaceae (duckweeds): Inferences from micromolecular and morphological data**. *Plant Systematics and Evolution* 1997, **204**(3-4):161-177.
13. Hollingsworth ML, Andra Clark A, Forrest LL, Richardson J, Pennington RT, Long DG, Cowan R, Chase MW, Gaudeul M, Hollingsworth PM: **Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants**. *Mol Ecol Resour* 2009, **9**(2):439-457.
14. Wang W, Wu Y, Yan Y, Ermakova M, Kerstetter R, Messing J: **DNA barcoding of the Lemnaceae, a family of aquatic monocots**. *BMC Plant Biology* 2010, **10**:205.
15. Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C *et al*: **Whole-genome sequencing of multiple Arabidopsis thaliana populations**. *Nat Genet* 2011, **43**(10):956-963.
16. Kuehdorf K, Jetschke G, Ballani L, Appenroth K: **The clonal dependence of turion formation in the duckweed Spirodela polyrhiza-an ecogeographical approach.** *Physiol Plant* 2013, **10**.
17. Appenroth K, Borisjuk N, Lam E: **Telling duckweed apart: genotyping technologies for the Lemnaceae**. *Chin J Appl Environ Biol* 2013, **19**(1):1-10.
18. Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T: **Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology**. *Nucleic Acids Res* 2008, **36**(19):e122-e122.
19. Wang W, Messing J: **High-Throughput Sequencing of Three Lemnoideae (Duckweeds) Chloroplast Genomes from Total DNA**. *PLoS ONE* 2011, **6**(9):e24670.
20. Wang W, Wu Y, Messing J: **The mitochondrial genome of an aquatic plant, Spirodela polyrhiza**. *PLoS ONE* 2012, **7**(10):e46747.
21. Consortium B: **Genome sequencing and analysis of the model grass Brachypodium distachyon**. *Nature* 2010, **463**(7282):763-768.

22. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA *et al*: **The B73 maize genome: complexity, diversity, and dynamics**. *Science* 2009, **326**(5956):1112-1115.
23. Consortium R: **The map-based sequence of the rice genome**. *Nature* 2005, **436**(7052):793-800.
24. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A *et al*: **The Sorghum bicolor genome and the diversification of grasses**. *Nature* 2009, **457**(7229):551-556.
25. D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M *et al*: **The banana (Musa acuminata) genome and the evolution of monocotyledonous plants**. *Nature* 2012, **488**(7410):213-217.
26. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proceedings of the National Academy of Sciences of the United States of America* 1977, **74**(12):5463-5467.
27. Gardner RC, Howarth AJ, Hahn P, Brown-Luedi M, Shepherd RJ, Messing J: **The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing**. *Nucleic acids research* 1981, **9**(12):2871-2888.
28. Metzker ML: **Sequencing technologies - the next generation**. *Nat Rev Genet* 2010, **11**(1):31-46.
29. Vieira J, Messing J: **The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers**. *Gene* 1982, **19**(3):259-268.
30. Luo MC, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J: **High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis**. *Genomics* 2003, **82**(3):378-389.

CHAPTER 1 DNA BARCODE

1.1 Abstract

Members of the aquatic monocot subfamily *Lemnoideae* (commonly called duckweeds) represent the smallest and fastest growing flowering plants. Their highly reduced morphology and infrequent flowering result in a dearth of characters for distinguishing between the nearly 38 species that exhibit these tiny, closely-related and often morphologically similar features within the same family of plants.

We developed a simple and rapid DNA-based molecular identification system for the *Lemnoideae* based on sequence polymorphisms. We compared the barcoding potential of the seven plastid-markers proposed by the CBOL (Consortium for the Barcode of Life) plant-working group to discriminate species within the land plants in 97 accessions representing 31 species from the family of *Lemnoideae*. A *Lemnoideae*-specific set of PCR and sequencing primers were designed for four plastid coding genes (*rpoB*, *rpoC1*, *rbcL* and *matK*) and three noncoding spacers (*atpF-atpH*, *psbK-psbI* and *trnH-psbA*) based on the *Lemna minor* chloroplast genome sequence. We assessed the ease of amplification and sequencing for these markers, examined the extent of the barcoding gap between intra- and inter-specific variation by pairwise distances, evaluated successful identifications based on direct sequence comparison of the "best close match" and the construction of a phylogenetic tree.

Based on its reliable amplification, straightforward sequence alignment, and rates of DNA variation between species and within species, we propose that the *atpF-atpH*

noncoding spacer could serve as a universal DNA barcoding marker for species-level identification of duckweeds.

1.2 Introduction

The cost of DNA purification and sequencing has dropped considerably in recent years so that identification of individual species by DNA barcoding has become an independent, subtler method than solely morphological-based classification to distinguish closely related species, which also defines the systematic relationships by analysis of genetic distance. The key element for a robust barcode is a suitable threshold between inter- and intra-specific genetic distances. Sequence variation between species has to be high enough to tell them apart while the distances within species must be low enough for them to cluster together [1]. The mitochondrial cytochrome c oxidase subunit I (COI) gene has proven to be a reliable, cost-effective, and easily recovered barcode marker to successfully identify animal species [2], but its application in the plant kingdom is impeded by a slow nucleotide substitution rate, which is insufficient for the diagnosis of individual species [3, 4]. However, the Consortium for the Barcode of Life (CBOL) plant-working group recently proposed seven leading candidate sequences for use as barcoding markers [5]. Four plastid coding genes (*rpoB*, *rpoC1*, *rbcL* and *matK*) and three noncoding spacers (*atpF-atpH*, *psbK-psbI* and *trnH-psbA*) have been selected based on previous investigations among different plant families [6-8]. However, the utility of each of these sequences for individual families of species within the plant kingdom is hardly predictable [9, 10].

Although there have been attempts to use the single-locus of *matK* [6], a combination of two loci, *rbcL* and *trnH-psbA* [7], and even multi-loci combinations [11] as barcoding sequences, the use of a unified barcode for the identification of all the land plants would be difficult due to conflicting needs of different researchers. For example, an optimal barcode marker that has been determined empirically to distinguish plants at the family level may prove less useful for making accurate species level identifications. Most of the proposed plant barcode markers were designed primarily for identifying distantly related organisms in biodiversity hotspots such as Panama [12] and Kruger National Park in South Africa [6]. So far, little attention and only a few studies have been devoted to developing unified barcodes suitable for making identifications within a family, within a genus, or between closely related sister species. A test of seven other candidate barcoding sequences in the family of Myristicaceae was applied to eight species within a genus and yielded two suitable barcodes [13]. Recently, it has been shown that all three markers (*rbcL*, *trnH-psbA* and *matK*) can discriminate 4 sister species of Acacia across three continents [14]. The marker *matK* has been reported to distinguish 5 Dendrobium species [15]. More complex approaches have been developed at the subfamily level identification of larger groups of related plants [16]. Although an extensive barcode study for 31 Carex species suggested that a single locus or even multiple loci cannot provide a resolution of greater than 60%, it did not include some of the new markers (*atpF-atpH* and *psbK-psbI*) [17]. When *atpF-atpH* and *psbK-psbI* were included for distinguishing Carex and Kobresia, it could be shown that *matK* identifies 95% as single-locus or 100% of the species when combined with another marker. However, this study used material from a well defined regional perspective, the Canadian

Arctic Archipelago, where the number of co-existing closely related species is limited [18]. Our objective was to determine whether one or more of the markers proposed by the CBOL plant-working group would serve as an optimal marker for species-level identification within the family *Lemnoideae*.

Until now, the most readily observed anatomical feature of the minute and highly reduced duckweeds are their fronds with or without roots. These few and somewhat variable morphological characters and rarely emerging flowers or fruits make identification of duckweeds extremely difficult even for professional taxonomists [19]. Complementing traditional classification methods with a DNA-based method would be highly applicable for such a family of species. It would permit these species to be classified in a highly reproducible and cost effective manner because DNA-based methods are independent of morphology, integrity, and developmental stage of the organism and can distinguish among species that superficially look alike [20].

Here, we present a simple and accessible protocol to barcode duckweeds and establish a sequence database against which unknown species may be compared and tentative species identifications can be validated. This database also provides a high-resolution phylogenetic resource for this important plant monocot family.

1.3 Results

Sampling criteria

The duckweed family consists of 38 species classified into 5 genera [21]. A worldwide collection has been characterized by genome sizes [22]. From this collection,

97 ecotypes were sampled for the current work representing all five genera and 31 species (81.6% of the known species) (Table 1.1). The ecotypes selected encompass the worldwide geographical distribution of duckweeds originating from different climates and geographical regions, ranging from N60^{o} to S42^{o} latitude and 9 m to 1287 m in altitude (Figure 1.1) (Table 1.1). 85 ecotypes from 19 species were used for statistical calculations and candidate barcode evaluations. An additional 12 single-ecotype species were examined to determine the broader applicability of the barcode markers for identification.

Figure 1.1 Google map of duckweed collection.
The distribution of duckweeds was made by GPS with corresponding latitude and longitude.

Ecotype	Altitude & GPS	*psbK-psbI*	*trnH-*	*matK*	*atpF-*	*rpoB*	*rpoC1*	*rbcL*
Spirodela intermedia 7125	547 S34º W56º	454290		454125	454194	454030	453933	454387
Spirodela intermedia 7178	548 S34º W58º	454291		454126	454195	454031	453934	454388
Spirodela intermedia 7291	550 S3º W60º	454292	454484	454127	454196	454032	453935	454389
Spirodela intermedia 7355	567 N5º W55º	454293	454485	454128	454197	454033	453936	454390
Spirodela intermedia 7450	555 N28º E77º	454294		454129	454198	454034	453937	454391
Spirodela intermedia 7747	558 S11º W76º	454295	454486	454130	454199	454035	453938	454392
Spirodela intermedia 8410	563 N30º W85º	454296		454131	454200	454036	453939	454393
Spirodela polyrrhiza 7205	643 N22º E114º	454297	454487	454132	454201	454037	453940	454394
Spirodela polyrrhiza 7212	1253 N23º E87º	454298	454488	454133	454202	454038	453941	454395

Spirodela polyrrhiza 7222	646 N3º E101º	454299	454489	454134	454203	454039	453942	454396
Spirodela polyrrhiza 7498	658 N35º W78º	454300	454490	454135	454204	454040	453943	454397
Spirodela polyrrhiza 7657	663 N18º W94º	454301	454491	454136	454205	454041	453944	454398
Spirodela polyrrhiza 8790	702 N54º W124º	454302	454492	454137	454206	454042	453945	454399
Spirodela polyrrhiza 9203	1202 N6º W72º	454303		454138	454207	454043	453946	454400
Spirodela polyrrhiza SJ		454304	454493	454139	454208	454044	453947	454401
Landoltia punctata 7248	589 S33º E18º	454305	454494	454140	454209	454045	453948	454402
Landoltia punctata 7260	590 S38º E141º	454306	454495	454141	454210	454046	453949	454403
Landoltia punctata 7449	597 N28º E77º	454307	454496	454142	454211	454047	453950	454404
Landoltia punctata 7487	600 N27º W82º	454308	454497	454143	454212	454048	453951	454405
Landoltia punctata 8721	616 S17º E145º	454309	454498	454144	454213	454049	453952	454406
Landoltia punctata 9278	N30º E114º	454310	454499	454145	454214	454050	453953	454407
Lemna aequinoctialis 6612	131 N36º W120º	454311	454500	454146	454215	454051	453954	454408
Lemna aequinoctialis 6746	132 N37º W121º	454312	454501	454147	454216	454052	453955	454409
Lemna aequinoctialis 7126	136 N30º W97º	454313	454502	454148	454217	454053	453956	454410
Lemna disperma 7269	273 S42º E147º	454314	454503	454149	454218	454054	453957	454411
Lemna gibba 7589	317 N34º W117º	454315	454504	454150	454219	454055	453958	454412
Lemna gibba 7741	324 N37º E15º	454316	454505	454151	454220	454056	453959	454413
Lemna gibba 7784	327 N9º E38º	454317	454506	454152	454221	454057	453960	454414
Lemna gibba 8703	344 N36º E138º	454318	454507	454153	454222	454058	453961	454415
Lemna gibba JS 6F7-11		454319	454508	454154	454223	454059	453962	454416
Lemna gibba JS parent line		454320	454509	454155	454224	454060	453963	454417
Lemna japonica 7182	357 N33º E130º	454321	454510	454156	454225	454061	453964	454418
Lemna minor 7018	9 N39º E39º	454322	454511	454157	454226	454062	453965	454419
Lemna minor 7136	15 N39º W89º	454323	454512	454158	454227	454063	453966	454420
Lemna minor 7210	21 S33º E26º	454324	454513	454159	454228	454064	453967	454421
Lemna minor 9016	70 N36º E138º	454325	454514	454160	454229	454065	453968	454422
Lemna minor 9253	1242 N60º E24º	454326	454515	454161	454230	454066	453969	454423
Lemna minor 9417	N50º E8º	454327	454516	454162	454231	454067	453970	454424
Lemna minuta 7284	1219 S34º W56º	454328	454517	454163	454232	454068	453971	454425
Lemna minuta 7726	97 S33º W71º	454329	454518	454164	454233	454069	453972	454426
Lemna minuta 8065	99 N29º W95º	454330	454519	454165	454234	454070	453973	454427
Lemna obscura 7856	1177 N30º W91º	454331	454520	454166	454235	454071	453974	454428
Lemna trisulca 7579	429 N43º W79º	454332	454521	454167	454236	454072	453975	454429
Lemna trisulca 8137	440 N35º W115º	454333	454522	454168	454237	454073	453976	454430
Lemna trisulca UTCC 399		454334	454523	454169	454238	454074	453977	454431
Lemna turionifera 8339	362 N32º E118º	454335	454524	454170	454239	454075	453978	454432
Lemna turionifera 8760	1239 N49º E15º	454336	454525	454171	454240	454076	453979	454433
Lemna valdiviana 7288	500 S1º W63º	454337	454526	454172	454241	454077	453980	454434
Lemna valdiviana 8634	106 N18º W77º	454338	454527	454173	454242	454078	453981	454435
Lemna valdiviana 9229	1257 S0º W78º	454339	454528	454174	454243	454079	453982	454436
Lemna valdiviana 9232	1258 S0º W78º	454340	454529	454175	454244	454080	453983	454437
Wolffiella denticulata 8221	984 S28º E30º	454341	454530	454176	454245	454081	453984	454438
Wolffiella gladiata 7595	990 N37º W77º	454342	454531	454177	454246	454082	453985	454439
Wolffiella gladiata 7852	991 N30º W91º	454343	454532	454178	454247	454083	453986	454440
Wolffiella gladiata 8066	992 N28º W96º	454344	454533	454179	454248	454084	453987	454441
Wolffiella gladiata 8261	993 N41º W80º	454345	454534		454249	454085	453988	454442
Wolffiella gladiata 8350	994 N39º W88º	454346	454535		454250	454086	453989	454443
Wolffiella hyalina 8640	1003 S2º E36º	454347	454536	454180	454251	454087	453990	454444

Wolffiella lingulata 7289	1007 S1º W63º	454348	454537		454252	454088	453991	454445
Wolffiella lingulata 7464	1011 N10º W66º	454349	454538	454181	454253	454089	453992	454446
Wolffiella lingulata 7655	1013 N18º W92º	454350	454539	454182	454254		453993	454447
Wolffiella lingulata 7725	1015 S28º W58º	454351	454540	454183	454255	454090	453994	454448
Wolffiella lingulata 8742	1023 S27º W58º	454352	454541			454091	453995	454449
Wolffiella neotropica 7290	1056 S1º W63º	454353	454542	454184	454256	454092	453996	454450
Wolffiella neotropica 7609	1057 S19º W40º	454354	454543	454185	454257	454093	453997	454451
Wolffiella neotropica 8848	1058 S23º W43º	454355	454544	454186	454258	454094	453998	454452
Wolffiella oblonga 7164	1063 N29º W90º	454356	454545	454187	454259	454095	453999	454453
Wolffiella oblonga 7201	1065 S34º W58º	454357	454546		454260	454096	454000	454454
Wolffiella oblonga 7343	1066 S34º W58º	454358	454547		454261	454097	454001	454455
Wolffiella oblonga 8072	1075 N28º W96º	454359	454548		454262	454098	454002	454456
Wolffiella oblonga 9136	1086 S17º W57º	454360	454549	454188	454263	454099	454003	454457
Wolffiella rotunda 9072	1285 S19º E29º	454361	454550	454189	454264	454100	454004	454458
Wolffiella rotunda 9121	1216 S16º E28º	454362	454551	454190	454265	454101	454005	454459
Wolffia angusta 7476	711 S36º E145º	454363	454552		454266	454102	454006	454460
Wolffia arrhiza 8872	753 N46º E20º	454364	454553		454267	454103	454007	454461
Wolffia australiana 7733	766 S34º E138º	454365	454554	454191	454268	454104	454008	454462
Wolffia australiana 8730	769 S32º E151º	454366	454555	454192	454269	454105	454009	454463
Wolffia borealis 9123	1207 N33º W117º	454367	454556		454270	454106	454010	454464
Wolffia brasiliensis 7150	784 N29º W98º	454368	454557		454271	454107	454011	454465
Wolffia brasiliensis 7306	786 N19º W99º	454369	454558		454272	454108	454012	454466
Wolffia brasiliensis 8743	809 S27º W58º	454370	454559		454273	454109	454013	454467
Wolffia columbiana 7310	862 N19º W99º	454371	454560		454274	454110	454014	454468
Wolffia columbiana 7972	880 N30º W87º	454372	454561		454275	454111	454015	454469
Wolffia columbiana 8265	884 N41º W80º	454373	454562		454276	454112	454016	454470
Wolffia columbiana 8856	888 S24º W64º	454374	454563		454277	454113	454017	454471
Wolffia columbiana 8890	890 N15º W88º	454375	454564		454278	454114	454018	454472
Wolffia cylindracea 9080	913 S19º E29º	454376	454565		454279	454115	454019	454473
Wolffia elongata 9188	1211 S10º W74º	454377	454566		454280	454116	454020	454474
Wolffia globosa 8152	1151 N36º W120º	454378	454567		454281	454117	454021	454475
Wolffia globosa 8441	949 N13º E100º	454379	454568		454282	454118	454022	454476
Wolffia globosa 8691	950 N35º E136º	454380	454569		454283	454119	454023	454477
Wolffia globosa 8789	953 N27º E84º	454381	454570		454284	454120	454024	454478
Wolffia globosa 8973	960 N15º E100º	454382	454571		454285	454121	454025	454479
Wolffia globosa 9196	964 N9º W75º	454383	454572	454193	454286	454122	454026	454480
Wolffia globosa 9317	N26º E73º	454384	454573		454287	454123	454027	454481
Wolffia microscopica 9276	1287 N28º E77º	454385	454574		454288	454124	454028	454482
Wolffia neglecta 9149	969 N24º E67º	454386	454575		454289		454029	454483

Table 1.1 Information of sampled duckweeds.

The table lists the altitude and GPS locations of collected samples as well as Genbank IDs. To save the space, all Genbank IDs omit the initial "GU".

Validation of DNA barcoding markers

To simplify identification of different species by DNA barcodes, a target DNA sequence marker has to meet two basic requirements: the first is a high success rate during PCR amplification and DNA sequencing, the second is sufficient DNA sequence polymorphism to permit different species to be distinguished and evolutionary distances between them to be calculated [1]. The CBOL plant-working group proposed 7 leading candidates [5], i.e., 4 coding genes (*rpoB*, *rpoC1*, *rbcL* and *matK*) and 3 noncoding spacers (*atpF-atpH*, *psbK-psbI* and *trnH-psbA*). To evaluate the seven markers, genomic DNA extracted from the 97 ecotypes was subjected to PCR amplification with the primer pairs based on the chloroplast sequence of *Lemna minor*. The PCR primers were also used for sequencing (See Materials and methods). PCR and sequencing were generally successful (≥95%) for all the barcode candidates except *matK* (71%) (Table 1.2). The maximal and minimal alignment length of PCR product for *rpoB*, *rpoC1*, *rbcL* and *matK* were identical, while that of *atpF-atpH*, *psbK-psbI* and *trnH-psbA* were quite variable, with a range of 579-622 bp, 185-576 bp and 286-504 bp, respectively. It was not unexpected that the coding markers (*rpoB*, *rpoC1*, *rbcL* and *matK*) were conserved in PCR product length, while the noncoding spacers (*atpF-atpH*, *psbK-psbI* and *trnH-psbA*) displayed more variability due to extensive insertions/deletions (Table 1.2). These results indicate that the selection of markers by the COBL plant-working group should provide a reasonable level of success for new untested plant families.

	psbK-psbI	*trnH-psbA*	*matK*	*atpF-atpH*	*rpoB*	*rpoC1*	*rbcL*
Max. length of product∗	576	504	725	622	389	450	522
Min. length of product∗	185	286	719	579	389	450	522
# tested Samples	97	97	97	97	97	97	97
% Success of PCR and sequencing	100%	95%	71%	99%	98%	100%	100%

Table 1.2 Success ratios of PCR amplification and sequencing

∗ The analyzed product length becomes shorter than corresponding one's due to removal of the end of ambiguous nucleotides

Intra- and inter-specific DNA sequence polymorphism

To assess the degree of DNA polymorphism between DNA samples, sequence divergences between and within species were calculated by Kimura 2-parameter (K2P) and uncorrected p-distance, respectively. Both models exhibited the same tendency: higher average interspecific diversity and lower intraspecific distance. For example, the K2P distance within and between species is as follows: *psbK-psbI* (0.1648 and 0.0072), *trnH-psbA* (0.1133 and 0.0058), *matK* (0.0715 and 0.0019), *atpF-atpH* (0.0633 and 0.0008) *rpoB* (0.0388 and 0.0069), *rpoC1* (0.0303 and 0.0006), *rbcL* (0.0216 and 0.0004). The noncoding spacer *psbK-psbI* showed the highest interspecific diversity (66 average substitution sites among 675 bp), while the coding marker *rbcL* is the most conserved one (11 average substitution sites among 522 bp) (Table 1.3). The most variable barcode between species was *psbK-psbI*, followed by *trnH-psbA*, *matK* and *atpF-atpH* (Table 1.3). The lowest intraspecific distance was provided by *atpF-atpH* and *rbcL*, whereas the highest is *trnH-psbA*, *psbK-psbI* and *matK* (Table 1.3). Although none

of the seven proposed markers possessed both the highest variation between species and the lowest distance within a species, *atpF-atpH* seemed to show sufficient interspecific but relatively low intraspecific divergence, compared to the other six markers (Table 1.3).

Region	*psbK-psbI*	*trnH-psbA*	*matK*	*atpF-atpH*	*rpoB*	*rpoC1*	*rbcL*
Aligned length (bp) *	675	520	725	674	389	450	522
Interspecific substitution	66	32	48	44	13	13	11
Interspecific K2P	0.1648±0.0221	0.1133±0.0120	0.0715±0.0061	0.0633±0.0068	0.0338±0.0051	0.0303±0.0050	0.0216±0.0038
Intraspecific K2P	0.0072±0.0015	0.0058±0.0014	0.0019±0.0003	0.0008±0.0002	0.0069±0.0008	0.0006±0.0002	0.0004±0.0002
Interspecific P-distances	0.1435±0.0156	0.0986±0.0095	0.0671±0.0052	0.0601±0.0059	0.0327±0.0048	0.0295±0.0048	0.0212±0.0037
Intraspecific P-distance	0.0066±0.0012	0.0057±0.0014	0.0019±0.0003	0.0008±0.0002	0.0062±0.0007	0.0006±0.0002	0.0004±0.0002

Table 1.3 Measurement of inter- and intra-specific divergences

∗ Aligned length becomes longer than corresponding ones due to addition of the gap.
K2P= Kimura 2-parameter distances.

The accuracy of barcoding for species identification depended to a large extent on the barcoding gap between intraspecific and interspecific sequence variations. Effective barcoding became weaker when interspecific and intraspecific distances overlapped. To evaluate whether there was a significant barcoding gap, we calculated the distribution of divergences for the seven markers (Figure 1.2). Median and Mann–Whitney U tests inferred that the mean of intraspecific divergence was significantly lower than that of interspecific distance in each case ($p<0.0001$). Even though *psbK-psbI* and *trnH-psbA* exhibited the highest rates of divergence between species, they were also most diverged within species, which could easily result in misidentification (Table 1.3) (Figure 1.2)(Figure 1.3). On the other hand, the adequate variation and the narrow overlapping

distance of the *atpF-atpH* marker would ensure accurate ecotype and species identification (Table 1.3)(Figure 1.2).

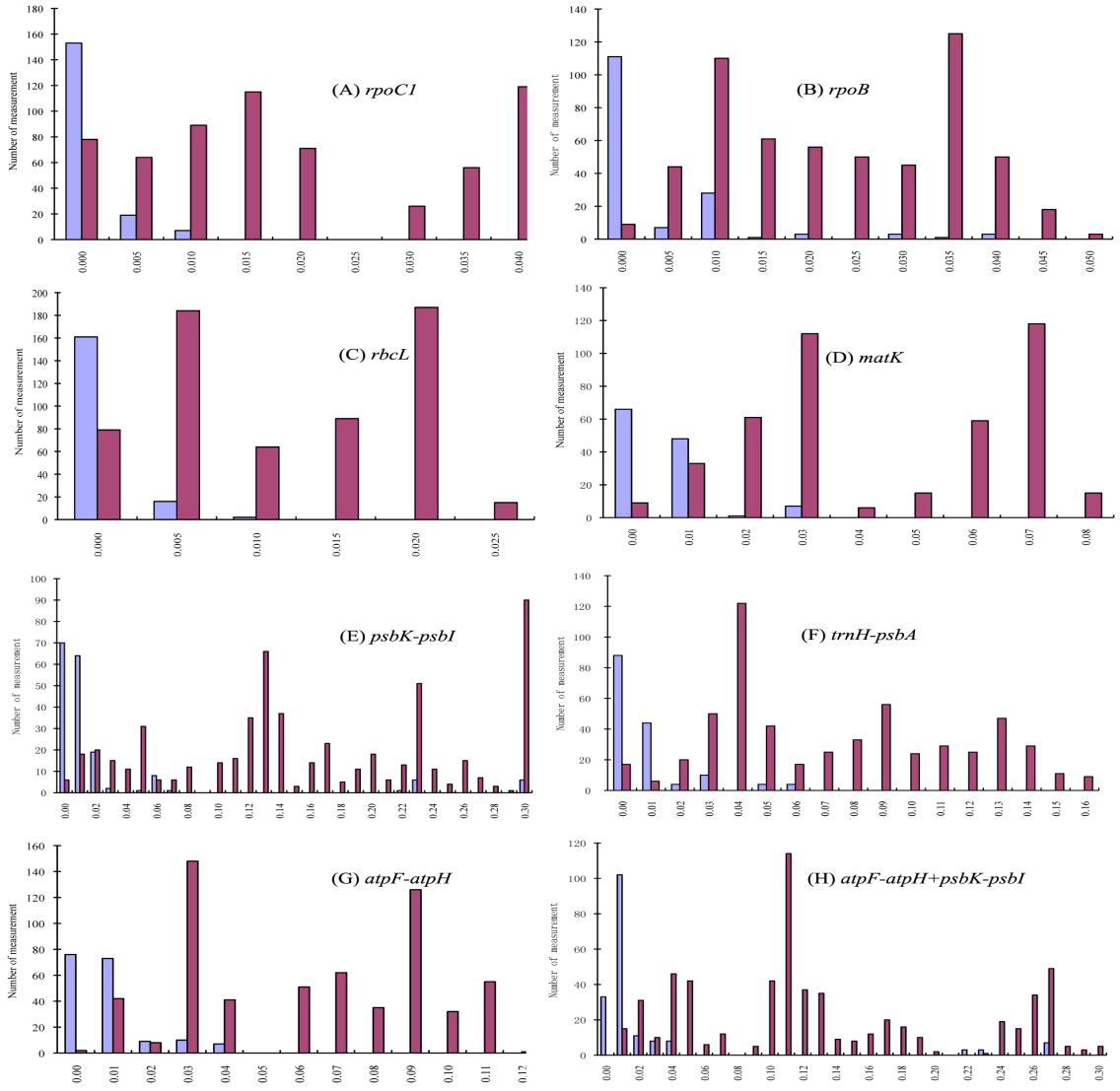


Figure 1.2 Relative distribution of intra- and inter-specific divergence.

(A) *rpoC1*. (B) *rpoB*. (C) *rbcL*. (D) *matK*. (E) *psbK-psbI*. (F) *trnH-psbA*. (G) *atpF-atpH*. (H) *atpF-atpH+ psbK-psbI*. X axis is uncorrected p-distance with corresponding increment unit based on variation of each marker. Y axis is the number of occurrences. Barcoding gaps were evaluated with high significance ($p<0.0001$) by Median and Mann–Whitney U tests for all markers. Blue bars indicate intraspecific distance and red bars are interspecific distance.

DNA sequence similarity–based identification

In order to test whether accurate species identification can be made in our samples, we adopted the "best match" function in the program TAXONDNA [23]. The rank order for the correct identification is *atpF-atpH* (92.85%) *psbK-psbI* (84.7%), *trnH-psbA* (82.5%), *matK* (77.77%), *rpoB* (77.5%), *rpoC1* (70.58%), *rbcL* (70.58%) (Table 1.4). Generally, the three noncoding spacers produced higher rates of successful identifications than those of the four coding markers. Consistent with Table 1.5, *atpF-atpH* yielded the best result with 92.85% successful identifications. Among 84 ecotypes (not including species with single sampled ecotypes), 78 samples were successfully discriminated, three were ambiguous and three were incorrectly identified using *atpF-atpH*. When we combined *atpF-atpH* with one of the other five barcoding markers, the percentage of correct identification dropped, except for *psbK-psbI*, which gave an increase of 1.19% (Table 1.4). The markers *matK* + *atpF-atpH* were not counted because of the small number of sequence comparisons done with *matK*.

	psbK-psbI	trnH-psbA	matK	atpF-atpH	rpoB	rpoC1	rbcL	psbK-psbI + atpF-atpH	trnH-psbA + atpF-atpH	rpoB + atpF-atpH	rpoC1 + atpF-atpH	rbcL + atpF-atpH
Correct	72 (84.7%)	66 (82.5%)	49 (77.77%)	78 (92.85%)	62 (77.5%)	60 (70.58%)	60 (70.58%)	79 (94.04%)	71 (89.87%)	77 (91.66%)	77 (91.66%)	77 (91.66%)
Ambiguous	8 (9.41%)	11 (13.75%)	10 (15.87%)	3 (3.57%)	12 (15.0%)	21 (24.7%)	21 (24.7%)	0 (0.0%)	3 (3.79%)	2 (2.38%)	4 (4.76%)	4 (4.76%)
Incorrect	5 (5.88%)	2 (2.5%)	4 (6.34%)	3 (3.57%)	6 (7.5%)	4 (4.7%)	2 (2.35%)	5 (5.95%)	5 (6.32%)	5 (5.95%)	3 (3.57%)	3 (3.57%)
No match	0 (0.0%)	1 (1.25%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	2 (2.35%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
Threshold	22.12%	4.01%	2.62%	2.96%	2.57%	0.44%	0.38%	22.16%	2.08%	2.44%	1.77%	1.67%

Table 1.4 Identification success based on "best close match" tools

"best close match" was analyzed by TAXONDNA program [23] with single region or two-region combinations. The ecotypes was classified into correct, ambiguous, incorrect and no match group. The group number was shown in each well. Number in bracket indicates percentage in all barcoding ecotypes. *matK* + *atpF-atpH* was not counted due to the small number of sequence comparison done for *matK*. Percentage in the bracket was calculated by dividing each item by all tested sample.

Tree-based sequence classification

As an alternative to sequence similarity-based identification, we estimated the proportion of recovered monophyly from multiple conspecific ecotypes per species in the phylogenetic tree for each barcoding marker. Here, we need to stress that the primary purpose of the tree is not so much the evolutionary relationship, but the species identification. The *atpF-atpH* attained the highest score of monophyletic species (73.7%, i.e., 14 correctly identified out of 19 species) (Table 1.5). The number of successfully identified species with the other six markers was *rpoB* (11), *rpoC1* (11), *rbcL* (11), *trnH-psbA* (10), *psbK-psbI* (8). The *atpF-atpH* marker did not distinguish closely-related pairs of sister species such as *W. gladiata* and *W. oblonga* and *L. minuta* and *L. valdiviana*.

Although the location of most grouped ecotypes in the taxonomic trees did not change in regard to each marker, a close examination consistently revealed two interesting connections. First, despite the fact that very little is known about how cross pollination in these tiny flowering plants occurs, *L. japonica* has been suspected to originate from a hybridization event between *L. minor* and *L. turionifera* based on morphological characters [24]. Our data indicates that sequence from each of the seven

tested markers of *L. japonica* 7182 was always identical to and clustered with *L. minor*. Since the chloroplast is maternally inherited in many (but not all) plants, our data is consistent with *L. japonica* arising from a cross between *L. minor* and *L. turionifera.*

The second connection was *S. polyrhiza* 9203, which consistently clusters with *S. intermedia* rather than other *S. polyrhiza* in all seven tested markers. We examined 34 ecotypes of *S. polyrhiza* from the collection using the *atpF-atpH* marker and found four additional ecotypes that grouped closely with *S. intermedia* (Figure 1.3). This suggested that these accessions might have been misidentified as *S. polyrhiza* due to the overlap in morphological characteristics between these species.

Loci	UPGMA	MP
psbK-psbI	8(93.3)	8 (87.5)
trnH-psbA	10 (87.5)	10 (85.7)
matK	/	/
atpF-atpH	14 (100)	14 (94.1)
rpoB	11 (83.3)	11 (68.8)
rpoC1	11 (85.7)	11 (68.8)
rbcL	11 (85.7)	12 (68.8)

Table 1.5 Number of monophyletic species recovered with the best two phylogenetic methods

The number of monophyletic species out 19 species was shown in each well. Proportions supported by bootstrap >50% are in brackets.

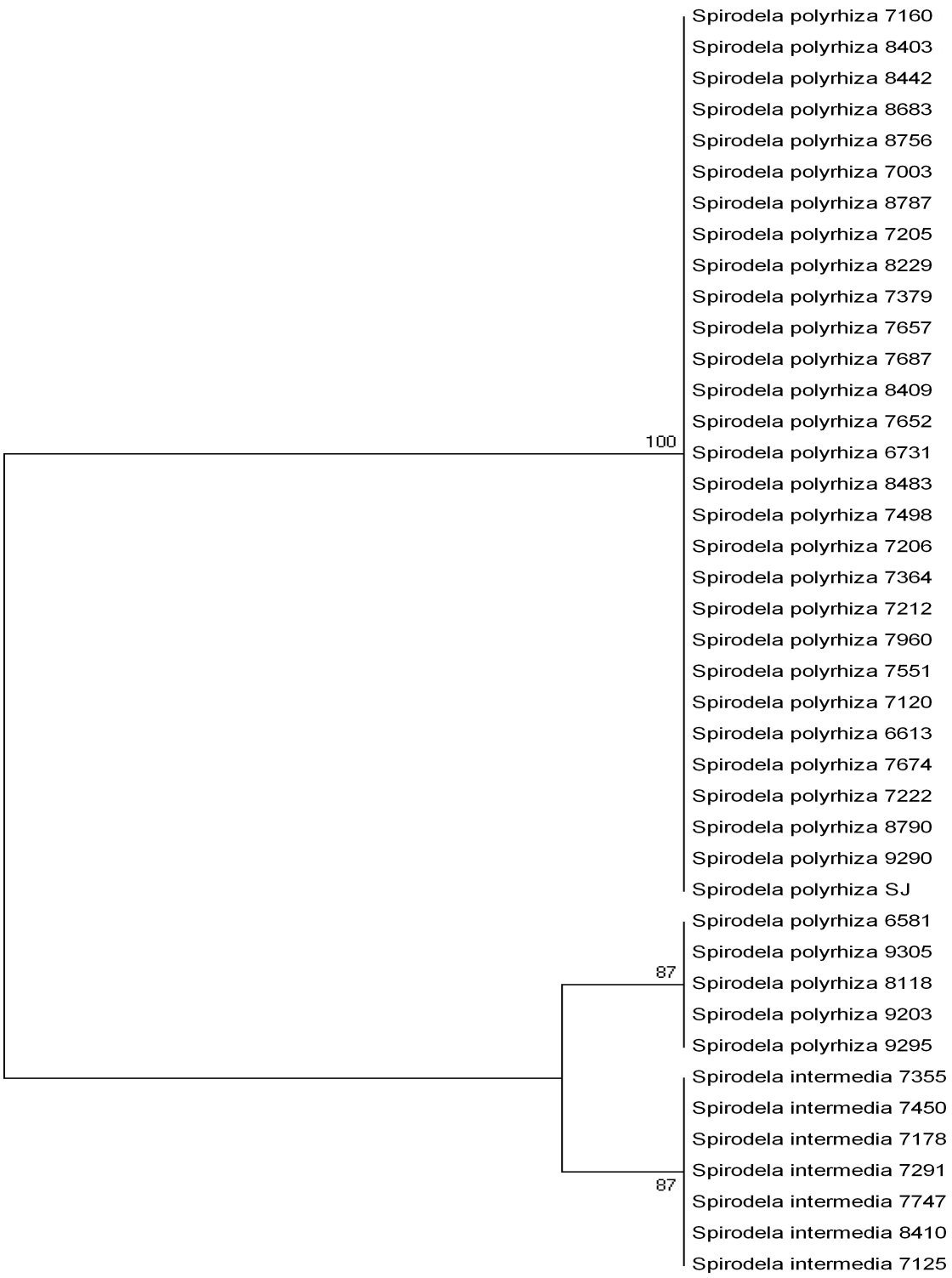


Figure 1.3 UPGMA tree for Spirodela based *atpF-atpH* sequences.

1.4 Discussion

Here, we present data validating the most useful DNA barcoding markers for the family of *Lemnoideae* from among those proposed by the CBOL plant-working group. Such a fundamental, whole family-wide analysis lays the groundwork for phylogenetic and genomic studies. Our samples represent a worldwide collection from the same family with many sister species (Figure 1.1) (Table 1.1). Specimens in previous taxonomic classifications using barcoding markers were mainly from distantly related groups from broadly different families that originated from the local or more defined regions, such as the National Park [6], the Amazon [25], and the Panama region [12]. Because of the diversity of the collection that has accumulated over the years, duckweeds provide a unique system to test the proposed barcoding markers for closely related species. Furthermore, it is difficult to classify members of this family by morphology alone. Therefore, we can not only validate the universal application of barcoding markers, but also apply it to species that may be solely dependent on such an approach for conservation. The advantage of universal barcoding markers is the design of universal primers for barcoding markers from reference sequences, which in this case was *L. minor* [26]. The primers worked very well for all the samples (31 species and 97 ecotypes) with PCR amplification and the sequencing success rates better than 95%, except in the case of *matK*, which yielded a rate as low as 71% (Table 1.2). In addition, a lower PCR annealing temperature than optimal for *Lemna minor* permits primers to anneal to the target sequences despite sequence polymorphism in related species. It is interesting that most PCR failure existed in the *Wolffioideae* subfamily (Table 1.1). The locus *matK* has been shown to be very variable in numerous phylogenetic studies [27, 28]} and other

studies have also noted the difficulties of its utilization due to PCR failure and lack of truly universal primer sites [7, 8]. Further improvement of primer designs for *matK* for other targets could increase amplification success, but might fail because of less conserved sites near the most variable sequences of the locus. Although *matK* DNA sequences exhibited the highest interspecific variation among the four coding markers (Table 1.3), the low percentage of successful PCR amplification and sequencing in duckweeds would restrict its extensive use.

It was not surprising that the noncoding spacers showed dramatically higher sequence variability than the coding markers (Table 1.3). Given the slow evolutionary rate of *rpoB*, *rpoC1* and *rbcL* (especially for *rbcL*, which is strongly recommended for barcoding across all land plants), they work well to distinguish distantly related species either alone or when combined with other more variable regions [4, 7]. However, their sequence polymorphisms might not be sufficient to distinguish closely related species. The non-coding spacers of *psbK-psbI* and *trnH-psbA* were the most polymorphic plastid sequences with variable sequence length in duckweeds (Table 1.2). The size of *trnH-psbA* in Spirodela (~504 bp) was 218 bp longer than in the other four genera (~286 bp). The length of the *psbK-psbI* sequence was the most variable, ranging from ~185 bp in *S. polyrhiza* to ~479 bp in *S. intermedia* even though they were sister-species (Table 1.2). These significant length variations caused by deletion/insertion, simple sequence repeats and rearrangements were problematic for accurate alignment, but could potentially be adapted for simple diagnostic tests that would not require DNA sequencing. Furthermore, the high sequence polymorphisms of the aligned sequences of *psbK-psbI* and *trnH-psbA* could offer greater distinction between species in a diverse set of genera in certain

families [3, 6]. Still, one has to use caution for intraspecies comparison where the relatively higher intraspecific distance compromised their power in barcoding duckweed species. One nearly has to cluster samples into two groups, one for ecotypes of the same species and one for species to species comparison (Table 1.3). Failure to do so would prevent the detection of true differences between congeneric species and conspecific ecotypes and therefore impede the use of a universal duckweed barcode (Figure 1.2).

Although previous studies showed that *atpF-atpH* as a barcoding marker was inferior to *psbK-psbI*, *trnH-psbA* and *matK* based on distantly related species [3, 6, 7], our data suggested that it was the most promising barcoding marker for duckweeds with respect to high PCR amplification, ease of alignment, and sufficient sequence divergence (Figure 1.2)(Table 1.2)(Table 1.3)(Table 1.4)(Table 1.5). Therefore, our data differed from the conclusions of evaluating barcoding markers made from unrelated species. Although it was shown that barcoding plants by more than one region tended to be more effective [9-11], combination of *atpF-atpH* with any of the other markers resulted in only slight increases or drops of the rate of successful identification of species compared to itself alone (Table 1.4), indicating that the discriminatory power of *atpF-atpH* has already reached an optimum. When the *atpF-atpH* marker was combined with other markers, the reduced resolution lowered the differential value without complementary benefits. A similar finding that a combination of *matK* and *trnH-psbA* did not improve species identification has been reported as well [6].

Generally speaking of members of the duckweed family, the more derived they are, the simpler their morphologies. The reduction in size and simplification in structure

make the fronds more mobile and better successfully adapt to variable conditions [24]. *S. intermedia* was characterized by a slight degree of primitivism of more nerves, roots, and ovules compared to *S. polyrhiza*, which suggested that *S. intermedia* was differentiated into *S. polyrhiza* potentially through gradual morphological reduction and isolation. However, gradual differences were sometimes difficult to distinguish from each other due to overlapping characteristics [24]. Our studies for 34 ecotypes of *S. polyrhiza* using *atpF-atpH* markers showed five ecotypes that have been clustered with *S. intermedia* (Figure 1.3), which is mainly restricted to South America [24]. Among five ecotypes, three are derived from South America, while another two are from India. Therefore, a refined classification is necessary to determine whether another four ecotypes except *S. polyrhiza* 9203 should be classified as *S. intermedia* rather than *S. polyrhiza*.

Both phylogenetic data [21] and our barcoding data showed that closely related species *W. gladiata* and *W. oblonga*, *L. minuta* and *L. valdiviana* could not be separated from each other. These sister-species share identical sequences for barcoding markers, which would require a search for additional barcoding markers with greater sequence polymorphism. In fact, a universal DNA barcoding marker has not been reported to distinguish more than 90% of species tested until now [6, 25]. Elucidation of recently evolved species sharing identical barcoding sequences still needs further taxonomic or case-by-case morphological, flavonoid, and allozyme analyses. On the other hand, use of next-generation sequencing technologies and corresponding software applications are emerging where low pass coverage of different specimen could provide the necessary resolution.

1.5 Materials and methods

Plant materials

The *Lemnoideae* collection originated from the Institut für Integrative Biologie (Zürich, Switzerland), the BIOLEX company (North Carolina, USA), and the University of Toronto Culture Collection of Algae and Cyanobacteria (UTCC, Toronto, Canada) where it was maintained for many years. Detailed information about many of these accessions is included in Dr. Landolt's monographic study [19]. In total, 97 ecotypes representing 31 species (81.6% of the known species) were sampled in this study. Since the intraspecific distance is very important for evaluating a suitable barcoding marker, 2 to 8 representatives per species are included for 19 species, whereas another 12 species are represented by a single ecotype. Moreover, the selected ecotypes represent a worldwide geographical distribution (Figure 1.1).

DNA amplification, sequencing and alignment

All duckweed fronds were grown aseptically in half-strength Schenk and Hildebrandt medium (Sigma, S6765). Total DNA was extracted using CTAB [29]. The chloroplast markers *rpoB*, *rpoC1*, *rbcL*, *matK*, *atpF-atpH*, *trnH-psbA*, and *psbK-psbI*, which were proposed by the CBOL plant-working group, were amplified with a set of modified primers based on reference sequences from *Lemna minor* [26]. The amplicon sizes were also estimated according to *Lemna minor* (Table 1.6). PCR reaction conditions also followed guidelines from the CBOL plant-working group. Briefly, 50-100 ng genomic DNA and 5 pmol of each primer are added with the JumpStartTM REDTaq® ReadyMixTM Reaction Mix (P1107, Sigma) Redix in 25 ml of final volume. To improve

the universal application of these primers, they were designed to have an annealing temperature (Ta) of 50°C, which is 1 to 6 °C lower than the optimal Ta of *Lemna minor* (Table 1.6). The program uses the following formula: optimal Ta = 0.3 × Tm (primer) + 0.7 Tm (product) -14.9 [30]. After the ambiguous nucleotides (~30bp) at the ends of reads were removed, the length of products was measured and multiple DNA sequence alignments were generated using ClustalW in MEGA 4.1 [31].

Genetic distance analysis

Genetic distance was calculated using pairwise alignments of sequences between and within species (Table 1.3). The average intraspecific distance was calculated with the mean pairwise distance in each species with more than one representative, which eliminated biases due to unbalanced sampling among taxa. Median and Mann–Whitney U tests were executed to examine the extent of DNA barcoding gap/overlap between intra- and inter-specific divergences [6].

Evaluation of DNA barcoding markers based on sequence similarity

For assessing success in species assignment or identification among our data set, we adopted the "best match" function in the program TAXONDNA [23]. We calculated pairwise distances as uncorrected pairwise distances and compared two sequences over at least 300 bp except for *psbK-psbI* (230 bp). Since the best match was based on direct sequence comparison with other conspecific ecotypes, the analysis only counted species with multiple ecotypes per species.

Evaluation of DNA barcoding markers using phylogenetic analysis

The other criterion used to measure success of species identification was based on generating a phylogenetic tree. We built trees with MEGA 4.1 by using the best algorithms methods of UPGMA and MP compared with other tree building techniques for DNA barcoding [6].

Marker	Primer sequence	Amplicon size (*Lemna minor*)	Ta Optimum (*Lemna minor*)
psbK-psbI	Forward: 5'-TTAGCATTTGTTTGGCAAG-3';	544 bp	51 °C
	Reverse: 5'- AAAGTTTGAGAGTAAGCAT -3'		
trnH-psbA	Forward: 5'-GTTATGCACGAACGTAATGCTC-3';	300 bp	55 °C
	Reverse: 5'- CGCGCGTGGTGGATTCACAATCC-3'		
matK	Forward: 5'-CGTACTGTACTTTTATGTTTACGAG-3';	862 bp	55 °C
	Reverse: 5'- ATCCGGTCCATCTAGAAATATTGGTTC -3'		
atpF-atpH	Forward: 5'-ACTCGCACACACTCCCTTTCC-3';	675 bp	53 °C
	Reverse: 5'- GCTTTTATGGAAGCTTTAACAAT -3'		
rpoB	Forward: 5'-ATGCAGCGTCAAGCAGTTCC-3';	406 bp	55 °C
	Reverse: 5'- TCGGATGTGAAAAGAAGTATA -3'		
rpoC1	Forward: 5'-GGAAAAGAGGGAAGATTCCG-3';	509 bp	56 °C
	Reverse: 5'- CAATTAGCATATCTTGAGTTGG -3'		
rbcL	Forward: 5'-GTAAAATCAAGTCCACCACG-3';	580 bp	56 °C
	Reverse: 5'-ATGTCACCACAAACAGAGACTAAAGC -3'		

Table 1.6 List of primers for the seven proposed DNA barcoding markers.

Acknowledgement

This work has been published in BMC Plant Biology in 2010 by authors of Wang W, Wu Y, Yan Y, Ermakova M, Kerstetter R, Messing J with the title of "DNA barcoding of the Lemnaceae, a family of aquatic monocots".

1.6 References

1. Meyer CP, Paulay G: **DNA barcoding: error rates based on comprehensive sampling**. *PLoS Biol* 2005, **3**(12):e422.
2. Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM: **Identification of birds through DNA barcodes**. *PLoS Biol* 2004, **2**(10):e312.
3. Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH: **Use of DNA barcodes to identify flowering plants**. *PNAS* 2005, **102**(23):8369-8374.
4. Chase MW, Salamin N, Wilkinson M, Dunwell JM, Kesanakurthi RP, Haidar N, Savolainen V: **Land plants and DNA barcodes: short-term and long-term goals**. *Philos Trans R Soc Lond B Biol Sci* 2005, **360**(1462):1889-1895.
5. Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, Chase MW, Cowan RS, Erickson DL, Fazekas AJ *et al*: **A DNA barcode for land plants**. *Proceedings of the National Academy of Sciences* 2009, **106**(31):12794-12797.
6. Lahaye R, Savolainen, Vincent, Duthoit, Sylvie, Maurin, Olivier , and van der Bank, Michelle. : **A test of psbK-psbI and atpF-atpH as potential plant DNA barcodes using the flora of the Kruger National Park (South Africa) as a model system.** *Nature Precedings* 2008.
7. Kress WJ, Erickson DL: **A two-locus global DNA barcode for land plants: The coding rbcL gene complements the non-coding trnH-psbA spacer region**. *PLoS ONE* 2007, **2**(6):e508.
8. Chase MW, Cowan RS, Hollingsworth PM, van den Berg C, Madri, n S, Petersen G, Seberg O, rgsensen T, Cameron KM *et al*: **A proposal for a standardised protocol to barcode all land plants**. *Taxon* 2007, **56**:295-299.
9. Pennisi E: **TAXONOMY: Wanted: A barcode for plants**. *Science* 2007, **318**(5848):190-191.
10. Chase MW, Fay MF: **Barcoding of plants and fungi**. *Science* 2009, **325**(5941):682-683.
11. Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG, Husband BC, Percy DM, Hajibabaei M, Barrett SCH: **Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well**. *PLoS ONE* 2008, **3**(7):e2802.
12. Kress WJ, Erickson DL, Jones FA, Swenson NG, Perez R, Sanjur O, Bermingham E: **Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama**. *Proc Natl Acad Sci U S A* 2009, **106**(44):18621-18626.
13. Newmaster SG, Fazekas AJ, Steeves RAD, Janovec J: **Testing candidate plant barcode regions in the Myristicaceae**. *Molecular Ecology Resources* 2008, **8**(3):480-490.
14. Steven G N, Subramanyam R: **Testing plant barcoding in a sister species complex of pantropical Acacia (Mimosoideae, Fabaceae)**. *Molecular Ecology Resources* 2009, **9**:172-180.
15. Asahina H, Shinozaki J, Masuda K, Morimitsu Y, Satake M: **Identification of medicinal Dendrobium species by phylogenetic analyses using matK and rbcL sequences**. *Journal of Natural Medicines* 2010, **64**(2):133-138.

16. Ward J, Gilmore SR, Robertson J, Peakall R: **A grass molecular identification system for forensic botany: A critical evaluation of the strengths and limitations***. *Journal of Forensic Sciences* 2009, **54**(6):1254-1260.
17. Starr JR, Naczi RFC, Chouinard BN: **Plant DNA barcodes and species resolution in sedges (Carex, Cyperaceae)**. *Mol Eco Resources* 2009, **9**:151-163.
18. Clerc-Blain JL, Starr JR, Bull RD, Saarela JM: **A regional approach to plant DNA barcoding provides high species resolution of sedges (Carex and Kobresia, Cyperaceae) in the Canadian Arctic Archipelago**. *Molecular Ecology Resources* 2010, **10**(1):69-91.
19. Landolt E: **Biosystematic investigations in the family of duckweeds (Lemnaceae)**, vol. 1: Veroffentlichungen des Geobotanischen Institutes der Eidgenossischen Technischen Hochschule, Stiftung Rubel; 1980.
20. Hebert PDN, Gregory TR: **The promise of DNA barcoding for taxonomy**. *Syst Biol* 2005, **54**(5):852-859.
21. Les DH, Crawford DJ, Landolt E, Gabel JD, Kimball RT: **Phylogeny and systematics of Lemnaceae, the duckweed family**. *Systematic Botany* 2002, **27**(2):221-240.
22. Wang W, Kerstetter R, Michael T: **Evolution of genome size in duckweeds (Lemnaceae)**. *Journal of Botany* 2011(Special Issues).
23. Meier R, Shiyang K, Vaidya G, Ng PKL: **DNA barcoding and taxonomy in Diptera: A tale of high intraspecific variability and low identification success**. *Syst Biol* 2006, **55**(5):715-728.
24. Landolt E: **The family of Lemnaceae - a monographic study, Vol 1**, vol. 1: Veroffentlichungen des Geobotanischen Institutes der Eidgenossischen Technischen Hochschule, Stiftung Rubel; 1986.
25. Gonzalez MA, Baraloto C, Engel J, Mori SA, Petronelli P, Riera B, Roger A, Thebaud C, Chave J: **Identification of Amazonian trees with DNA barcodes**. *PLoS One* 2009, **4**(10):e7483.
26. Mardanov A, Ravin N, Kuznetsov B, Samigullin T, Antonov A, Kolganova T, Skyabin K: **Complete sequence of the duckweed (Lemna minor) chloroplast genome: structural organization and phylogenetic relationships to other angiosperms**. *Journal of Molecular Evolution* 2008, **66**(6):555-564.
27. Shaw J, Lickey EB, Beck JT, Farmer SB, Liu W, Miller J, Siripun KC, Small RL: **The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis**. *Am J Bot* 2005, **92**(1):142-166.
28. Hilu KW, Borsch T, Muller K, Soltis DE, Soltis PS, Savolainen V, Chase MW, Powell MP, Alice LA, Evans R *et al*: **Angiosperm phylogeny based on matK sequence information**. *Am J Bot* 2003, **90**(12):1758-1776.
29. Murray MG, Thompson WF: **Rapid isolation of high molecular weight plant DNA**. *Nucl Acids Res* 1980, **8**(19):4321-4326.
30. Rychlik W, Spencer WJ, Rhoads RE: **Optimization of the annealing temperature for DNA amplification in vitro**. *Nucleic Acids Res* 1990, **18**(21):6409-6412.
31. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0**. *Mol Biol Evol* 2007, **24**(8):1596-1599.

CHAPTER 2 DNA sizing

2.1 Abstract

To extensively estimate the DNA content and to provide a basic reference for duckweed genome sequence research, the nuclear DNA content for 115 different accessions of 23 duckweed species was measured by flow cytometry (FCM) stained with propidium iodide as DNA stain. The 1C-value of DNA content in duckweed family varied nearly thirteen-fold, ranging from 150 megabases (Mbp) in *Spirodela polyrhiza* to 1,881 Mbp in *Wolffia arrhiza*. There is a continuous increase of DNA content in *Spirodela*, *Lemna*, *Wolffiella* and *Wolffia* that parallels a morphological reduction in size. There is a significant intraspecific variation in the genus *Lemna*. However, no such variation was found in other studied species with multiple accessions of genera *Spirodela, Landoltia, Wolffiella,* and *Wolffia*.

2.2 Introduction

The advent of high-throughput sequencing technologies has enabled a new generation of model plant systems [1]. In an effort to initiate duckweed genomic research we endeavoured to identify species with small genomes that would be ideal for sequencing. First, we queried the Kew plant genome database (http://data.kew.org/cvalues/) and found that only 6 duckweed accessions had been measured by the Feulgen method [2, 3]. DNA content of single species from each genus

was determined and showed obvious difference. Due to it being laborious and time consuming, the popularity of Feulgen technique has waned. Feulgen has been largely replaced by flow cytometry (FCM) [4], a faster, easier, and more accurate method, and the current preferred technique for genome size estimations and DNA ploidy analyses in plants [5].

In order to find the smallest duckweed genome for sequencing and also explore previous observations about genome complexity in duckweeds, we estimated the genome size of all of the five duckweed genera using FCM. These genome size measurements will form the foundation for future work in sequencing duckweed genome, and enabling duckweeds as a model and applied system.

2.3 Results

Intra- and inter-species variations of genome sizes

The genome sizes of 115 accessions from 23 species representing 5 genera were estimated by FCM (Table 2.1). The DNA content estimates varied nearly thirteen-fold, ranging from 150 Mbp in *Spirodela polyrhiza* to 1,881 Mbp in *Wolffia arrhiza*. We superimposed the estimated 1C-value on a phylogenetic tree for *Lemnoideae* based on combination of morphological, flavonoid, allozyme and DNA sequence analysis [6] and found there is a continuous increase of DNA content in order of *Spirodela*, *Landoltia, Lemna*, *Wolffiella* and *Wolffia*, which correlates well with the morphological reduction

within the family (Figure 2.1 and 2.2).

In the genus *Spirodela*, we measured genome size for 34 accessions and found that the 1C DNA content only varies from 150 to 167 Mbp (Figure 2.1 and Table 2.1). The analysis of variance (ANOVA: single factor test) revealed that there was not a significant difference in *Spirodela polyrhiza* genome sizes ($P > 0.05$). Similarly, the 1C DNA content for 19 accessions of *Landoltia punctata* from 372 to 397 Mbp did not show significant variation (Figure 2.1). In the genus *Wolffiella,* the genome sizes range from 623 Mbp to 973 Mbp (Figure 2.1), which is almost as 4-6 times large as *Spirodela polyrhiza*. Like *Spirodela polyrhiza* and *Landoltia punctata*, there are no obvious intraspecific genome size variations in *Wolffiella hyalina* and *Wolffiella lingulata*. In the genus *Wolffia*, we measured 11 species and found they have the largest genome sizes on average among the duckweed family (Figure 2.1). 5.3-fold difference was observed from *Wolffia australiana* (357 Mbp) to *Wolffia arrhiza* (1,881 Mbp).

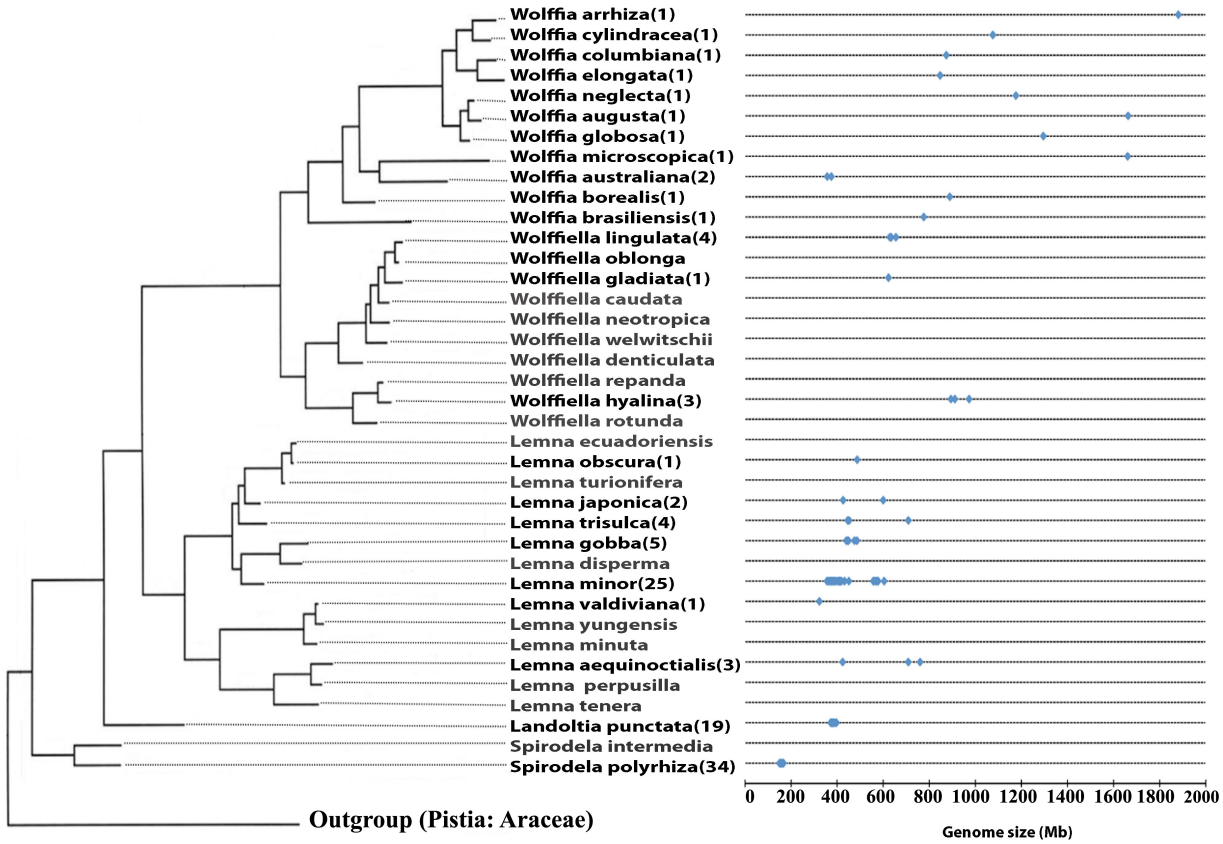


Figure 2.1 Genome size variation across the duckweeds.

Estimated 1C-value superimposed on a phylogenetic tree for *Lemnoideae* based on combination of morphological, flavonoid, allozyme and DNA sequence analysis [6]. The species in black were what we tested and the species in the grey were the ones we did not examine in this experiment. In the bracket is the number of different accessions we tested.

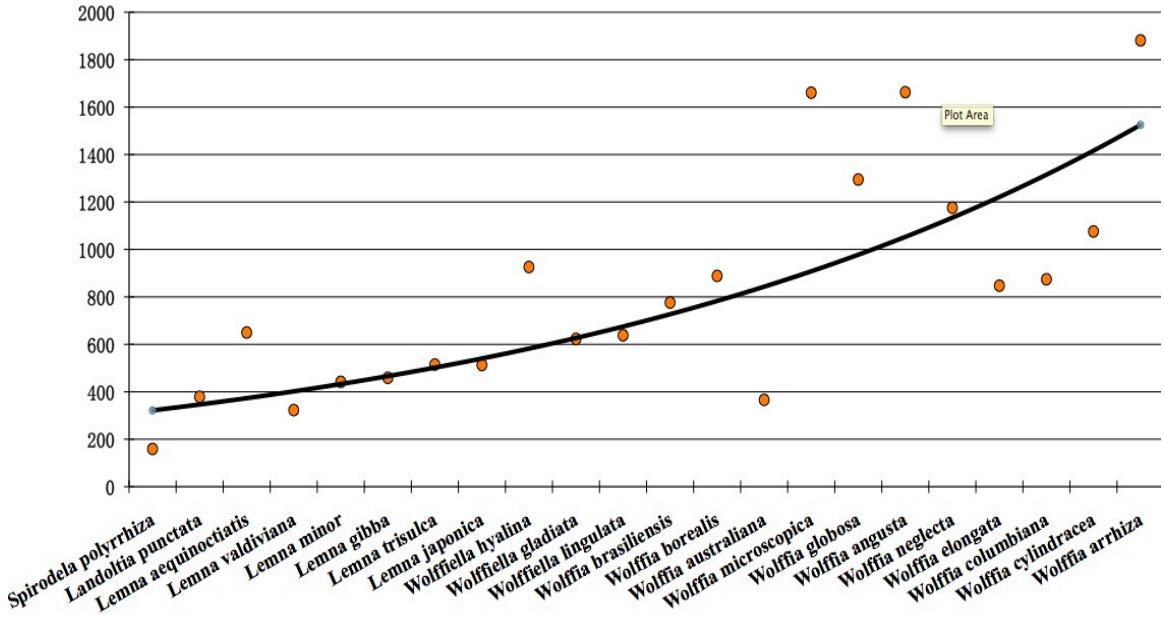
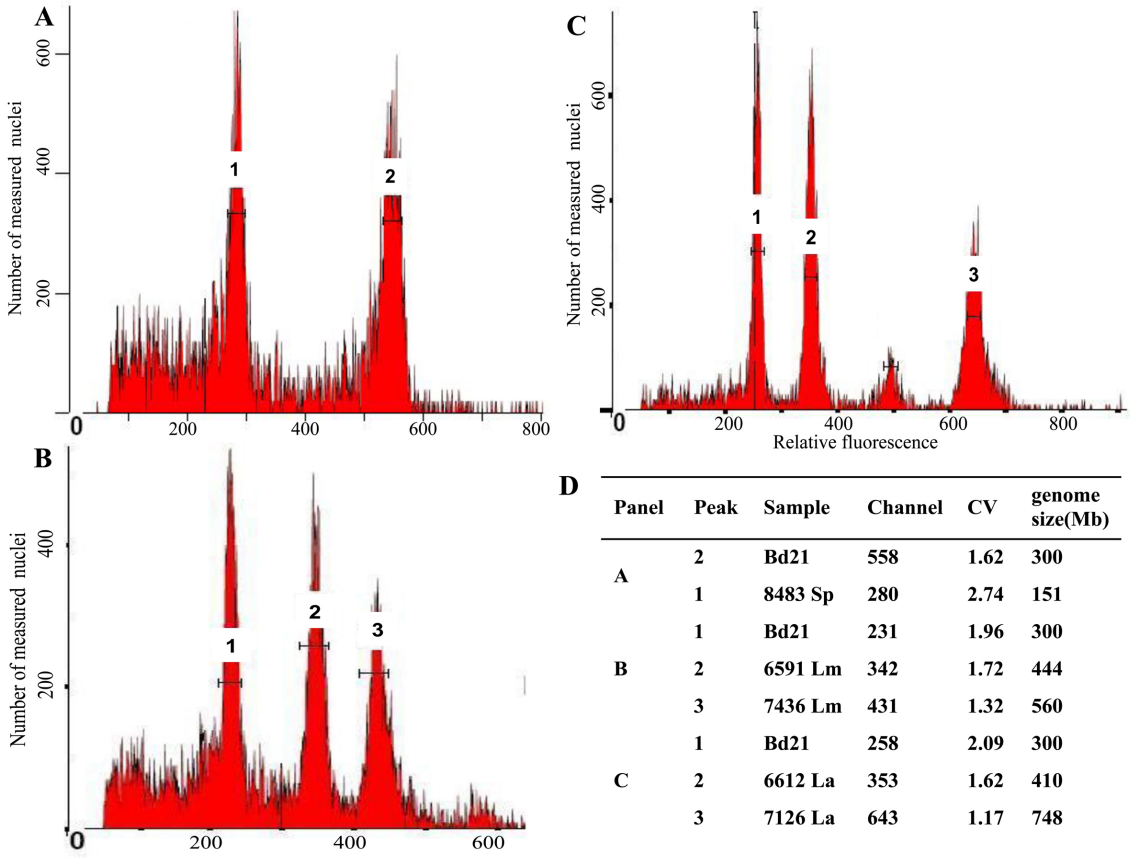


Figure 2.2 Average genome sizes of duckweed species.

Duckweed species are arranged on the x-axis from lower to higher evolutionary status, deduced from primitive and derived morphological traits [2].

In the genus *Lemna*, 7 species were investigated. There is a large amount of genome size variation in this genus. *Lemna valdiviana* has the smallest genome size (323 Mbp), while *Lemna aequinoctialis* has the biggest (760 Mbp). Surprisingly, intraspecific genome-size fluctuations are also impressive. For *Lemna minor,* 26 accessions have genome sizes ranging from 356 to 604 Mbp with up to 69.6% of the intraspecific DNA content variance. We confirmed the intraspecific difference of them by randomly choosing 2 *Lemna minor* with simultaneous measurement of both accessions (26.0% difference between 6591 Lm and 7436 Lm, Figure 2.3B). Statistical analyses revealed

significant differences among the *Lemna minor* accessions ($P < 0.01$). As well, *Lemna aequinoctialis* (424-760 Mbp, 79.2%) (Figure 2.3C), *Lemna trisulca* (446-709 Mbp, 59.0%), and *Lemna japonica* (426-600 Mbp, 40.8%) all show intraspecific difference, indicating a drastically uneven evolution of intraspecific genome expansion in *Lemna*.



Panel	Peak	Sample	Channel	CV	genome size(Mb)
A	2	Bd21	558	1.62	300
	1	8483 Sp	280	2.74	151
B	1	Bd21	231	1.96	300
	2	6591 Lm	342	1.72	444
	3	7436 Lm	431	1.32	560
C	1	Bd21	258	2.09	300
	2	6612 La	353	1.62	410
	3	7126 La	643	1.17	748

Figure 2.3 Flow cytometry (FCM) histograms.

(A) Histogram showing relative DNA content of *Spirodela polyrhiza* (1, 151 Mbp) and internal standard *Brachypodium distachyon* Bd21 (2, 300 Mbp) based on relative PI fluorescent intensity (channel number). Linear PI fluorescence intensity of G1 nuclei was

used for the calculation of DNA content (3500 particles were counted); (B) Difference in relative DNA content of two simultaneously measured *Lemna minor* accessions (2, Lm6591, 444 Mbp; 3, Lm7436, 560 Mbp) with internal standard Bd21 (1); 5000 particles were counted. (C) Difference in relative DNA content of two simultaneously measured *Lemna aequinoctialis* accessions (2, La6612, 410 Mbp; 3, La7126, 748 Mbp) with internal standard Bd21 (1); 5000 particles were counted.

1C-value and latitude, longitude and altitude

To investigate whether there is a correlation between genome-size variations and the geographic distribution in the duckweed, we compared genome size estimates with the latitude, longitude and altitude of recorded collection. However, genome size variation was not correlated with latitude by Pearson coefficient (r-value: *Spirodela* =-0.05, *Landoltia* =0.17, *Lemna*=-0.07, *Wolffiella* =-0.17, *Wolffia* =0.34) (Figure 2.4B), nor with longitude (r-value: *Spirodela* =0.17, *Landoltia* =0.04, *Lemna* =0.26, *Wolffia*=-0.41) (Figure 2.4C) except *Wolffiella* with a high r-value -0.86 possibly due to limited accessions (n=8). No correlation was found between C-values and altitude, either (r-value: *Spirodela* =0.13, *Landoltia* =-0.25, *Lemna* =-0.33, *Wolffiella* =-0.41, *Wolffia* =0.13) (Figure 2.4D). It is interesting we found that most of *Spirodela*, *Landoltia*, *Wolffiella* and *Wolffia* were collected from a similar geographic range between 0° to 45° and preferred to localize above 600 m to 1200 m of altitude. In contrast, most of *Lemna* species were collected between 30° to 60° and preferred to distribute below 600 m.

However, this most likely represents a sampling bias, and could also explain the absence of a relationship between genome size and the environment in duckweed.

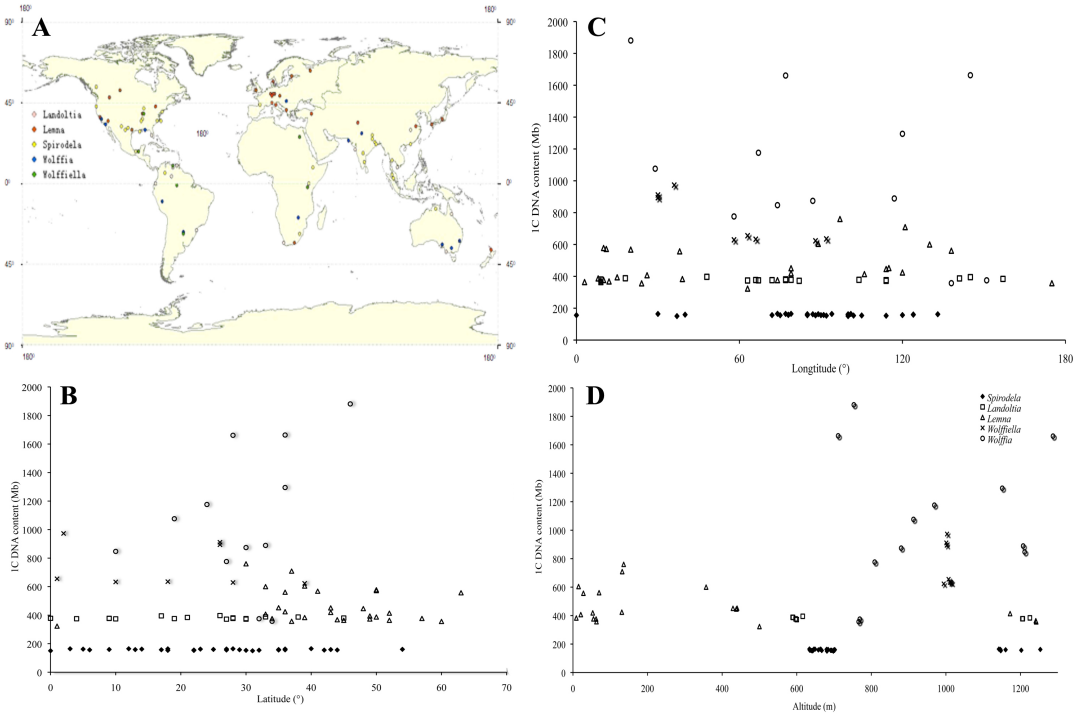


Figure 2.4 1C DNA content with geographical coordinates and altitude.

(A) Geographical origin of the duckweed samples analyzed; (B) Latitude and 1C DNA content; (C) Longitude and 1C DNA content; (D) Altitude and 1C DNA content.

Name	Genus	Species	1C DNA	pg	Chr	Lat	Hem	Lon	Hem	Alt	Country
8683	*Spirodela*	*polyrhiza*	150±6	0.15	ns	0	N	37	E	698	Kenya
8118	*Spirodela*	*polyrhiza*	150±6	0.15	40	31	N	100	W	681	USA
7205	*Spirodela*	*polyrhiza*	153±9	0.16	40	22	N	114	E	643	Hong Kong
7652	*Spirodela*	*polyrhiza*	153±12	0.16	30	18	N	92	W	1147	Mexico
7120	*Spirodela*	*polyrhiza*	154±6	0.16	40	30	N	102	W	640	USA
7160	*Spirodela*	*polyrhiza*	154±7	0.16	40	32	N	105	W	641	USA
7687	*Spirodela*	*polyrhiza*	155±12	0.16	40	42	N	88	W	667	USA
8787	*Spirodela*	*polyrhiza*	156±8	0.16	ns	27	N	85	E	701	Nepal

9295	*Spirodela*	*polyrhiza*	156±11	0.16		22	N	88	E		India
8483	*Spirodela*	*polyrhiza*	156±8	0.16	40	35	N	75	W	694	USA
8403	*Spirodela*	*polyrhiza*	156±10	0.16	40	44	N	0	W	691	France
8409	*Spirodela*	*polyrhiza*	156±8	0.16	40	35	N	90	W	692	USA
9203	*Spirodela*	*polyrhiza*	157±4	0.16	ns	6	N	72	W	1202	Colombia
6613	*Spirodela*	*polyrhiza*	157±6	0.16	40	36	N	120	W	635	USA
7003	*Spirodela*	*polyrhiza*	157±7	0.16	40	29	N	91	W	637	USA
9305	*Spirodela*	*polyrhiza*	157±9	0.16		17	N	78	E		India
7206	*Spirodela*	*polyrhiza*	159±12	0.16	40	13	N	100	E	644	Thailand
6731	*Spirodela*	*polyrhiza*	159±9	0.16	40	43	N	124	W	636	USA
7498	*Spirodela*	*polyrhiza*	159±2	0.16	40	35	N	78	W	658	USA
8756	*Spirodela*	*polyrhiza*	159±1	0.16	ns	9	N	40	E	700	Ethiopia
8442	*Spirodela*	*polyrhiza*	160±14	0.16	40	25	N	85	E	1160	India
8790	*Spirodela*	*polyrhiza*	160±9	0.16	ns	54	N	124	W	702	Canada
8229	*Spirodela*	*polyrhiza*	162±2	0.17	40	5	N	100	E	684	Malaysia
7212	*Spirodela*	*polyrhiza*	162±5	0.17	40	23	N	87	E	1253	India
7551	*Spirodela*	*polyrhiza*	162±4	0.17	40	14	S	133	E	1146	Northern Territory
7674	*Spirodela*	*polyrhiza*	163±2	0.17	40	27	N	85	E	665	Nepal
7960	*Spirodela*	*polyrhiza*	163±3	0.17	40	36	N	89	W	680	USA
SJ	*Spirodela*	*polyrhiza*	164±8	0.17							Europe
7657	*Spirodela*	*polyrhiza*	164±3	0.17	30	18	N	94	W	663	Mexico
7364	*Spirodela*	*polyrhiza*	164±1	0.17	30	28	S	30	E	649	South Africa
7222	*Spirodela*	*polyrhiza*	164±8	0.17	40	3	N	101	E	646	Malaysia
9290	*Spirodela*	*polyrhiza*	164±1	0.17		28	N	77	E		India
7379	*Spirodela*	*polyrhiza*	165±2	0.17	40	12	N	79	E	1142	India
6581	*Spirodela*	*polyrhiza*	165±10	0.17	40	40	N	74	W	634	USA
7487	*Landoltia*	*punctata*	372±14	0.38	40	27	N	82	W	600	USA
9279	*Landoltia*	*punctata*	372±2	0.38		30	N	114	E		China
9348	*Landoltia*	*punctata*	372±8	0.38							Brazil
9278	*Landoltia*	*punctata*	373±18	0.38		30	N	114	E		China
9393	*Landoltia*	*punctata*	374±15	0.38		10	N	63	W		Venezuela
9376	*Landoltia*	*punctata*	375±15	0.38		4	N	67	W		Venezuela
9323	*Landoltia*	*punctata*	376±5	0.38		19	N	72	E		India
9328	*Landoltia*	*punctata*	376±10	0.38		30	N	114	E		China
7449	*Landoltia*	*punctata*	377±6	0.39	40	28	N	77	E	597	India
9387	*Landoltia*	*punctata*	377±7	0.39		9	N	66	W		Venezuela
9234	*Landoltia*	*punctata*	378±6	0.39	ns	0	S	79	W	1205	Ecuador
9245	*Landoltia*	*punctata*	378±7	0.39	ns	9	N	104	E	1206	Vietnam
9354	*Landoltia*	*punctata*	380±16	0.39		45	N	9	E		Switzerland
9289	*Landoltia*	*punctata*	381±15	0.39		28	N	77	E		India

9264	*Landoltia*	*punctata*	384±2	0.39		21	N	157	W	1225	USA
7260	*Landoltia*	*punctata*	387±20	0.40	50	38	S	141	E	590	Victoria
7248	*Landoltia*	*punctata*	388±24	0.40	40	33	S	18	E	589	South Africa
8721	*Landoltia*	*punctata*	395±18	0.40	ns	17	S	145	E	616	Queensland
9332	*Landoltia*	*punctata*	397±26	0.41		26	S	48	W		?
6612	*Lemna*	*aequinoctialis*	424±7	0.43	40	36	N	120	W	131	USA
6746	*Lemna*	*aequinoctialis*	709±13	0.72	80	37	N	121	W	132	USA
7126	*Lemna*	*aequinoctialis*	760±27	0.78	60	30	N	97	W	136	USA
7288	*Lemna*	*valdiviana*	323±12	0.33	40	1	S	63	W	500	Brazil
9253	*Lemna*	*minor*	356±22	0.36		60	N	24	E	1242	Finland
8731	*Lemna*	*minor*	357±29	0.37	ns	37	S	175	E	63	New Zealand
9223	*Lemna*	*minor*	364±22	0.37		52	N	3	W	1241	United Kingdom
9345	*Lemna*	*minor*	364±3	0.37		45	N	9	E		Switzerland
9415	*Lemna*	*minor*	368±20	0.38		44	N	12	E		Italy
9441	*Lemna*	*minor*	373±21	0.38		49	N	9	E		Germany
8676	*Lemna*	*minor*	376±8	0.38	ns	34	N	74	E	62	India
8623	*Lemna*	*minor*	377±25	0.39	40	57	N	10	E	55	Denmark
7018	*Lemna*	*minor*	383±26	0.39	40	39	N	39	E	9	Turkey
9417	*Lemna*	*minor*	387±34	0.40		50	N	8	E		Germany
utcc491	*Lemna*	*minor*	392±25	0.40							
9438	*Lemna*	*minor*	394±33	0.40		49	N	15	E		Czech Republic
utcc490	*Lemna*	*minor*	403±23	0.41							
7210	*Lemna*	*minor*	407±36	0.42	40	33	S	26	E	21	South Africa
utcc492	*Lemna*	*minor*	411±25	0.42							
7123	*Lemna*	*minor*	414±10	0.42	42	52	N	106	W	1172	Canada
8434	*Lemna*	*minor*	419±3	0.43	40	43	N	79	W	53	Canada
utcc270	*Lemna*	*minor*	446±25	0.44							
6591	*Lemna*	*minor*	451±13	0.46	42						USA
7436	*Lemna*	*minor*	557±20	0.57	40	63	N	38	E	28	USSR
9016	*Lemna*	*minor*	561±24	0.57	ns	36	N	138	E	70	Japan
9436a	*Lemna*	*minor*	568±12	0.58		41	N	20	E		Albania
9436b	*Lemna*	*minor*	591±15	0.58		41	N	20	E		Albania
9439	*Lemna*	*minor*	572±26	0.58		50	N	11	E		Germany
9440	*Lemna*	*minor*	578±32	0.59		50	N	10	E		Germany
7136	*Lemna*	*minor*	604±30	0.62	40	39	N	89	W	15	USA
BIOL	*Lemna*	*gibba G-3*	440±11	0.45							

EX											
JS6F7-11	*Lemna*	*gibba G-3*	447±3	0.46							
JSPL	*Lemna*	*gibba G-3*	449±4	0.46							
utcc310	*Lemna*	*gibba G-3*	475±12	0.49							
Japan	*Lemna*	*gibba G-3*	486±6	0.50							
7943	*Lemna*	*trisulca*	446±9	0.46	40	48	N	114	W	438	USA
7579	*Lemna*	*trisulca*	451±7	0.46	80	43	N	79	W	429	Canada
8137	*Lemna*	*trisulca*	452±18	0.46	60	35	N	115	W	440	USA
UTCC 399	*Lemna*	*trisulca*	709±3	0.72							
8339	*Lemna*	*japonica*	426±9	0.44							
7182	*Lemna*	*japonica*	600±23	0.61	50	33	N	130	E	357	Japan
CA/SD	*Lemna*	*obscura*	487±8	0.50							USA
7378	*Wolffiella*	*hyalina*	894±30	0.91	40	26	N	30	E	1002	Egypt
7376	*Wolffiella*	*hyalina*	911±17	0.93	40	26	N	30	E	1001	Egypt
8640	*Wolffiella*	*hyalina*	973±35	0.99	40	2	S	36	E	1003	Tanzania
8350	*Wolffiella*	*gladiata*	623±12	0.64	40	39	N	88	W	994	USA
7725	*Wolffiella*	*lingulata*	629±8	0.64	20	28	S	58	W	1015	Argentina
7464	*Wolffiella*	*lingulata*	633±19	0.65	20	10	N	66	W	1011	Venezuela
7655	*Wolffiella*	*lingulata*	635±25	0.65	40	18	N	92	W	1013	Mexico
7289	*Wolffiella*	*lingulata*	655±13	0.67	50	1	S	63	W	1007	Brazil
8743	*Wolffia*	*brasiliensis*	776±52	0.79	ns	27	S	58	W	809	Argentina
9123	*Wolffia*	*borealis*	889±64	0.91	ns	33	N	117	W	1207	USA
7733	*Wolffia*	*australiana*	357±25	0.37	20	34	S	138	E	766	South Australia
8730	*Wolffia*	*australiana*	375±8	0.38	ns	32	S	151	E	769	New South Wales
9276	*Wolffia*	*microscopica*	1661±12	1.70		28	N	77	E	1287	India
8152	*Wolffia*	*globosa*	1295±42	1.32	60	36	N	120	W	1151	USA
7476	*Wolffia*	*angusta*	1663±34	1.70	40	36	S	145	E	711	Victoria
9149	*Wolffia*	*neglecta*	1176±40	1.20	ns	24	N	67	E	969	Pakistan
9188	*Wolffia*	*elongata*	847±42	0.87	ns	10	S	74	W	1211	Colombia
7972	*Wolffia*	*columbiana*	874±69	0.89	40	30	N	87	W	880	USA
9080	*Wolffia*	*cylindracea*	1076±86	1.10	ns	19	S	29	E	913	Zimbabwe
8872	*Wolffia*	*arrhiza*	1881±83	1.92	ns	46	N	20	E	753	Hungary

Table 2.1 Genome size across duckweeds.

1C-value expressed in megabase (Mbp) with standard deviation and picograms (1pg = 978 Mbp), somatic chromosome number (2n), latitude, longitude, altitude and location (when recorded) were deposited in this table.

2.4 Discussion

Genome evolution in duckweeds

In the phylogeny of *Lemnoideae* there is a strong relationship observed between genome size evolution and morphological progression. We found that the ancestral genus *Spirodela* has the smallest genome size, while the most advanced genus *Wolffia* contains biggest genome size (Figure 2.2), which correlates with the morphological reduction rather than organism complexity within the family. This result is consistent with Geber's finding, which showed there was a relationship between DNA content and degree of primitivity [7].

Genome doubling has been a pervasive force in plant evolution, which has occurred repeatedly [8]. Even the smaller genome of *Arabidopsis thaliana* has been impacted by genome duplication [9]. Cytological variation by counting the chromosomes was extensively investigated within duckweed. Landolt concluded that polyploidy (2n=20, 30, 40, 50, 60 and 80) is the main intra-populational variation [10], which means polyploidization was very active and occurred in the duckweeds for multiple rounds in the past. After polyploidization, transposable element mobility, insertions, deletion, and

epigenome restructuring contribute to the successful development of a new species and also genome size changes [11]. Changes in genome structure could lead to differential gene loss, extensive changes in gene expression [12], and have immediate effects on the phenotype and fitness of an individual [13]. It is likely polyploidy might drive the divergence during duckweed evolution.

Geographic distribution and genome size variation

It was suggested that variation in DNA content has adaptive significance and is correlated with the environmental traits of species [14]. The environmental conditions of plants are to a large extent determined by latitude, longitude and altitude. Previous studies have indicated a positive correlation between genome size and latitude (associated with the length of sun light with the growing season and the temperature), and also altitude (associated with the temperatures) among plant species. For example, the increase of DNA content corresponded with the increasing latitude found in the *Pinaceae* family [15] and with increasing altitude observed in *Zea mays* [16]. Duckweeds are distributed broadly around the world (Figure 2.4A). Our result shows there is no significant overall correlation of genome size with latitude, longitude and altitude (Figure 2.4). The same result was found in *Vicia faba* [17], *Sesleria albicans* [18], and *Asteraceae* [19]. A summary revealed that these relationships were not straightforward and not clear. Five studies (*Picea sitchensis, Berberis, Poaceae* and *Fabaceae*, Tropical vs. temperate grasses, 329 tropical vs. 527 temperate plants) found positive, seven

(*Arachis duranensis, Festuca arundinacea,* North American cultivars of *Zea mays,* 162 British plants, 23 Arctic plants, 22 North American *Zea mays*, 11 North American *Zea mays*) found negative, and five (*Allium cepa, Dactylis glomerata, Helianthus*) found non-significant correlations between genome size and latitude. Additionally, nine were positive, eight were negative, and six were not statistically significant between genome size and altitude [20]. But the different environmental distribution of the *Lemna* genus (30° to 60° of latitude and below 600m of altitude) with the other four duckweed genera (0° to 45° of latitude and 600m to 1200m of altitude) might explain the large intraspecific genome size variation.

Intraspecific variation in genome size

Intraspecific genome consistency has been reported in *Allium cepa* [21], *Glycine max* [22] and *Capsicum* and *Gossypium* [23]. We also found a similar result for *Spirodela polyrhiza*, *Landoltia punctata, Wolffiella hyalina, Wolffiella lingulata,* and *Wolffia australiana*, which don't have statistical intraspecific differences in genome size. However, more samples are needed for Wolffiella and Wolffia species to further confirm their genome stability. One explanation is that these species have a mechanism to maintain genome size constancy, e.g., by intraspecific stabilizing selection on genome size [24]. On the other hand, we found obvious intraspecific variation in *Lemna minor, Lemna aequinoctialis*, *Lemna trisulca* and *Lemna japonica.* Some artifacts of intraspecific variation in genome size have been noted, such as environmentally induced

variations, secondary compounds and fluorescence staining inhibitor, and erroneously determined species [4, 25]. However, our experiments are not complicated by these factors. We developed an easy bar-coding method to correctly identify duckweed species, which allowed us to correct any misnamed duckweed in the collection [26]. As cytosolic components may change in response to changes in the environment, we grew the duckweed plants under identical conditions. We used internal standardization such as *Brachypodium distachyon*, *Arabidopsis thaliana* and *Physcomitrella patens* that were prepared simultaneously and under the same experimental conditions as the duckweed accessions. Both duckweed and the internal standard have very little secondary compounds, which may affect genome size estimates. Additionally, we performed biological replicate on different days to eliminate instrument bias. In addition, intraspecific differences were independently confirmed by simultaneously measuring two accessions of the same species by FCM (Figure 2.3B and C).

The intraspecific variation may result from different numbers of repeated sequences, including satellite DNA [27], transposable elements [28] and ribosomal genes [29]. Large-scale polymorphism of heterochromatic repeats exist in the DNA of *Arabidopsis thaliana* and could account for about 50% of the variance among the *Arabidopsis thaliana* accession [30]. In addition, the amount of rDNA accounts for the differences in genome size between closely related lines of *Linum usitatissimum* (flax) [29]. The activity of transposable elements (TE) potentially multiply 20~100 times (~0.1-

1 Mbp) in a single generation [31]. For example, the BARE-1 TE is positively correlated with genome size within wild barley (*Hordeum spontaneum*) in response to sharp microclimatic divergence [28]. Deletions and insertions (INDELs) are most likely not candidates for genome size differences in duckweed. In *Drosophila melanogaster*, genome loss is only less than 1 bp per generation [32], indicating a small contribution to genome-size variation. However, in the fast growing duckweeds, which only need 2~5 days for each generation, one could imagine it is more likely that TE have higher rate than other flowering plants to influence genome size within and between species.

This is the first extensive analysis of genome sizes in duckweeds and examination of genome size variations across a range of taxonomic levels. We showed that duckweeds, in general, have remarkable smaller genome size compared with other flowering plants. The smallest genome size of *Spirodela polyrhiza*, combined with its sterile and controllable culture, fast growing, and promising application in research, suggest that this species may be good candidates for ongoing whole-genome sequencing projects and a model experimental tool. The 157 Mbp *Spirodela polyrhiza* genome is being sequenced by the DOE-JGI community-sequencing program (CSP), which will address challenges in alternative energy, bioremediation, and global carbon cycling. Also, the availability of a DNA C-values database of duckweeds and a consensus higher-level phylogenetic tree has opened the way for exploring the general processes underlying the evolution of genomes. Obvious intraspecific variation in duckweeds will

also provide nice material to study the mechanism of within-species and between-species variation in genome size. However, the main force driving the intraspecific variance and how the genome size affects the phenotype still requires more research.

2.5 Materials and methods

Plant materials

115 accessions of 23 duckweed species representing all 5 genera were measured in this study. Elias Landolt collected most of the duckweed accessions described in this work over the past 50 years (Landolt Duckweed Collection) [10]. Accessions were either obtained directly from Elias Landolt, BIOLEX (North Carolina, USA) or The University of Toronto Culture Collection of Algae and Cyanobacteria (UTCC). Currently the Landolt Duckweed Collection has been moved to Rutgers University. Plants were grown aseptically for 2 weeks with 1/2 full concentration of Schenk and Hildebrandt Basal Salt mixture (Sigma, USA) liquid culture medium under short day growth condition (8h light and 16h darkness with constant temperature 23°C). We bar-coded all the determined and undetermined species by identification of polymorphisms of chloroplast *atpF-atpH* noncoding spacer [26].

Isolation and staining of nuclei

To estimate nuclear DNA contents with flow cytometry (FCM), sample tissue

nuclei were stained with propidium iodide (PI) [33]. Briefly, 10 mg of fresh duckweed tissue and the same amount of the internal standard were chopped simultaneously with new razor blades and isolation buffer in a plastic Petri dish [34]. Isolates were filtered through a 30-µm nylon mesh into an Eppendorf tube. The suspensions of nuclei were stained with 50 µg ml^{-1} PI mixed with 50 µg ml^{-1} RNase (R4875,Sigma). The samples were incubated on ice for a few minutes before estimation by FCM.

Analysis of nuclear DNA content by FCM

PI-stained nuclei were analyzed for DNA content with a Coulter Cytomics FC500 Flow Cytometer (Beckman Coulter, Inc., Miami, FL). In all experiments, the fluorescence of at least 3000 G1-phase nuclei was measured. DNA content of each target sample was calculated by comparing its mean nuclear fluorescence with that of an internal standard (Figure 2.3A). We utilized internal controls that closely match the duckweed genome sizes being measured to ensure accuracy. The internal standard is a *Brachypodium distachyon* line, (Bd21, 300 Mbp) [5], *Arabidopsis thaliana* Columbia., (At, 147 Mbp) [35] and *Physcomitrella patens* ssp patens, (Pp, 480 Mbp) [36]. The numbers in bracket were generated by our flow cytometry equipment and our methods. Therefore, the validated genome sizes are not exactly the same but very close to cited references. Both duckweed and internal standards have very little secondary compounds, which will interfere with quantitative DNA staining. The absolute DNA content of a sample is calculated based on the values of the G1 peak means:

Sample 1C DNA content = [(sample G1 peak mean)/(standard G1 peak mean)]×standard 1C DNA content (Mbp). At least three independent biological replicates for each sample were analyzed on different days to estimate the mean DNA content. The transformation factor from pg to Mbp is: 1pg =978 Mbp [37].

Statistical analysis

Data on intraspecies variation of genome size were analysed by ANOVA: single factor test. To test whether genome size variation was correlated with geographic location or altitude of populations, the Spearman correlation coefficient (r) was used.

Acknowledgement

This work has been published in the Special Issues of Journal of Botany in 2011 by authors of Wang W, Kerstetter R, Michael T. with the title of "Evolution of genome size in duckweeds (*Lemnaceae*)".

2.6 References

1. Lister R, Gregory BD, Ecker JR: **Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond**. *Current Opinion in Plant Biology* 2009, **12**(2):107-118.
2. Landolt E: **The family of Lemnaceae - a monographic study, Vol 2**, vol. 1: Veroffentlichungen des Geobotanischen Institutes ETH, Stiftung Rubel, Zurich; 1986.
3. Geber G: **Zur Karyosystematik der Lemnaceae**: University of Vienna, Vienna; 1989.
4. Doležel J, Bartos J: **Plant DNA Flow Cytometry and Estimation of Nuclear Genome Size**. *Annals of Botany* 2005, **95**(1):99-110.

5. Bennett MD, Leitch IJ: **Nuclear DNA Amounts in Angiosperms: Progress, Problems and Prospects**. *Annals of Botany* 2005, **95**(1):45-90.
6. Les DH, Crawford DJ, Landolt E, Gabel JD, Kimball RT: **Phylogeny and Systematics of Lemnaceae, the Duckweed Family**. *Systematic Botany* 2002, **27**(2):221-240.
7. Landolt E: **Biosystematic investigations in the family of duckweeds (Lemnaceae)**. In., vol. 1: Veroffentlichungen des Geobotanischen Institutes ETH, Stiftung Rubel, Zurich; 1980: 30-97.
8. Hegarty MJ, Hiscock SJ: **Genomic Clues to the Evolutionary Success of Polyploid Plants**. *Current Biology* 2008, **18**(10):R435-R444.
9. Seoighe C, Gehring C: **Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome**. *Trends in Genetics* 2004, **20**(10):461-464.
10. Landolt E: **The family of Lemnaceae - a monographic study, Vol 1**, vol. 1: Veroffentlichungen des Geobotanischen Institutes ETH, Stiftung Rubel, Zurich; 1986.
11. Leitch AR, Leitch IJ: **Genomic Plasticity and the Diversity of Polyploid Plants**. *Science* 2008, **320**(5875):481-483.
12. Adams KL, Wendel JF: **Polyploidy and genome evolution in plants**. *Current Opinion in Plant Biology* 2005, **8**(2):135-141.
13. Otto SP: **The evolutionary consequences of polyploidy**. *Cell* 2007, **131**(3):452-462.
14. Bottini MCJ, Greizerstein EJ, Aulicino MB, Poggio L: **Relationships among Genome Size, Environmental Conditions and Geographical Distribution in Natural Populations of NW Patagonian Species of Berberis L. (Berberidaceae)**. *Annals of Botany* 2000, **86**(3):565-573.
15. Ohri D, Khoshoo TN: **Genome size in gymnosperms**. *Plant Systematics and Evolution* 1986, **153**(1):119-132.
16. Rayburn AL, Auger JA: **Genome size variation in Zea mays ssp. mays adapted to different altitudes**. *Theoretical and Applied Genetics* 1990, **79**(4):470-474.
17. Ceccarelli M, Minelli S, Maggini F, Cionini PG: **Genome size variation in Vicia faba**. *Heredity* 1995, **74**(2):180-187.
18. Lysak MA, Rostkova A, Dixon JM, Rossi G, Dolezel J: **Limited Genome Size Variation in Sesleria albicans**. *Annals of Botany* 2000, **86**(2):399-403.
19. Chrtek J, Jr, Zahradnicek J, Krak K, Fehrer J: **Genome size in Hieracium subgenus Hieracium (Asteraceae) is strongly correlated with major phylogenetic groups**. *Annals of Botany* 2009, **104**(1):161-178.
20. Knight CA, Molinari NA, Petrov DA: **The large genome constraint hypothesis: evolution, ecology and phenotype**. *Annals of Botany* 2005, **95**(1):177-190.
21. Bennett MD, Johnston S, Hodnett GL, Price HJ: **Allium cepa L. Cultivars from Four Continents Compared by Flow Cytometry show Nuclear DNA**

Constancy. *Annals of Botany* 2000, **85**(3):351-357.

22. Greilhuber J, Obermayer R: **Genome size and maturity group in Glycine max (soybean)**. *Heredity* 1997, **78**(5):547-551.
23. Hendrix B, Stewart J: **Estimation of the Nuclear DNA Content of Gossypium Species**. *Annals of Botany* 2005, **95**(5):789-797.
24. Šmarda P, Horová L, Bureš P, Hralová I, Marková M: **Stabilizing selection on genome size in a population of Festuca pallens under conditions of intensive intraspecific competition**. *New Phytologist* 2010, **187**(4):1195-1204.
25. Greilhuber J: **Intraspecific Variation in Genome Size in Angiosperms: Identifying its Existence**. *Annals of Botany* 2005, **95**(1):91-98.
26. Wang W, Wu Y, Yan Y, Ermakova M, Kerstetter R, Messing J: **DNA barcoding of the Lemnaceae, a family of aquatic monocots**. *BMC Plant Biology* 2010, **10**:205.
27. Bosco G, Campbell P, Leiva-Neto JT, Markow TA: **Analysis of Drosophila Species Genome Size and Satellite DNA Content Reveals Significant Differences Among Strains as Well as Between Species**. *Genetics* 2007, **177**(3):1277-1290.
28. Kalendar R, Tanskanen J, Immonen S, Nevo E, Schulman AH: **Genome evolution of wild barley (Hordeum spontaneum) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence**. *The Proceedings of the National Academy of Sciences* 2000, **97**(12):6603-6607.
29. Cullis CA: **Mechanisms and control of rapid genomic changes in flax**. *Annals of Botany* 2005, **95**(1):201-206.
30. Davison J, Tyagi A, Comai L: **Large-scale polymorphism of heterochromatic repeats in the DNA of Arabidopsis thaliana**. *BMC Plant Biology* 2007, **7**(1):44.
31. Petrov DA: **Evolution of genome size: new approaches to an old problem**. *Trends in Genetics* 2001, **17**(1):23-28.
32. Petrov DA, Hartl DL: **High rate of DNA loss in the Drosophila melanogaster and Drosophila virilis species groups**. *Molecular Biology and Evolution* 1998, **15**(3):293-302.
33. Doležel J, Greilhuber J, Suda J: **Estimation of nuclear DNA content in plants using flow cytometry**. *Nature Protocols* 2007, **2**(9):2233-2244.
34. Galbraith DW, Harkins KR, Maddox JM, Ayres NM, Sharma DP, Firoozabady E: **Rapid Flow Cytometric Analysis of the Cell Cycle in Intact Plant Tissues**. *Science* 1983, **220**(4601):1049-1051.
35. Arumuganathan K, Earle ED: **Nuclear DNA content of some important plant species**. *Plant Molecular Biology Reporter* 1991, **9**:208-218.
36. Schween G, Gorr G, Hohe A, Reski R: **Unique Tissue-Specific Cell Cycle in Physcomitrella**. *Plant Biology* 2003, **5**(1):50-58.
37. Doležel J, Bartoš J, Voglmayr H, Greilhuber J: **Nuclear DNA content and genome size of trout and human**. *Cytometry* 2003, **51A**(2):127-128.

CHAPTER 3 CHLOROPLAST GENOME

3.1 Abstract

Chloroplast genomes provide a wealth of information for evolutionary and population genetic studies. Chloroplasts play a particularly important role in the adaption for aquatic plants because they float on water and their major surface is exposed continuously to sunlight. The subfamily of *Lemnoideae* represents such a collection of aquatic species that because of photosynthesis represents one of the fastest growing plant species on earth.

We sequenced the chloroplast genomes from three different genera of *Lemnoideae*, *Spirodela polyrhiza*, *Wolffiella lingulata* and *Wolffia australiana* by high-throughput DNA sequencing of genomic DNA using the SOLiD platform. Unfractionated total DNA contains high copies of plastid DNA so that sequences from the nucleus and mitochondria can easily be filtered computationally. Remaining sequence reads were assembled into contiguous sequences (contigs) using SOLiD software tools. Contigs were mapped to a reference genome of *Lemna minor* and gaps, selected by PCR, were sequenced on the ABI3730xl platform.

This combinatorial approach yielded whole genomic contiguous sequences in a cost-effective manner. Over 1,000-time coverage of chloroplast from total DNA were reached by SOLiD platform in an individual spot on a quadrant slide without purification. Comparative analysis indicated that the chloroplast genome was conserved in gene number and organization with respect to the reference genome of *L. minor*. However, higher nucleotide substitution, abundant deletion and insertion occurred in non-coding

regions of these genomes, indicating a greater genomic dynamics than expected from the comparison of other related species in *Pooideae*. Noticeably, there was no transition bias over transversion in *Lemnoideae*. The data should have immediate applications in evolutionary biology and plant taxonomy with increased resolution and statistical power.

3.2 Introduction

Each plant cell has three genomes, separated in three subcellular compartments, the nucleus, the chloroplasts, and the mitochondria. Chloroplasts are key organelles of green plants for photosynthesis. They are also responsible for storage of starch, and synthesizing chlorophyll, nucleic acids, and 50% of soluble protein in leaves. Chloroplasts are highly conserved in terms of their structure, genome size (from 120 to 217 Kb) and its gene content (~130 genes) [1]. Chloroplasts contain multiple copies of a circular, double-stranded DNA molecule. For instance, leaf cells of tobacco and pea typically have ~100 chloroplasts and up to 10,000 DNA copies [2]. Total genomic DNA could have as many as 5,000 times the copies of chloroplast DNA relative to nuclear gene copies as tested in monocots and dicots [3]. In addition to its important biological roles, chloroplast genome sequences are widely used in evolutionary studies, comparative genomics [4], and biotechnology [5].

Lemnoideae (duckweeds) are a subfamily of the *Araceae* of aquatic flowering monocot plants [6]. However, their minute size and simple morphologically characteristics made them extremely difficult in systematic analysis and species identification. *atpF-atpH* barcode markers has been proposed to serve as species-level identification [7], while distinguishing the populations of a same species (ecotype) from

different geographical locations is still a problem. The distinct physiological attributes make ecotype-level discrimination intrigue [8]. For example, the ecotypes of Spirodela polyrhiza have a very broad of turion yield from 0.22 to 5.9 by one vegetative frond. Their finding also indicated that their phenotypic difference is inherited by the result of DNA mutations [9]. Therefore, in an effort to screen and isolate the right ecotypes with high starch, protein or rapid growth, or with heavy metal tolerance for the application of animal food, biofuel, wastewater treatment, it is prerequisite to identify ecotypes besides species [10]. It has so far proved to be difficult to genotype ecotypes only by looking for the very limited DNA markers. More informative polymorphism is needed to delimitate the border of ecotypes inside of the same species. Thus, the full plastid genome becomes the best option due to their highly conserved sequence but still increased resolution and informative sequence variation. Together with the fast improvement of next-generation sequencing technology, it is feasible to get multiple plastid genomes simultaneously with the multiplex bar-coded library system [11].

Compared with the traditional way of primer walking based on closely related known genomes which is time-consuming and labor-intensive [12], a recent study reported that chloroplast genome sequences were recovered from total DNA including nuclei, chloroplasts, and mitochondria by using an Illumina-based sequencing platform. Still, many gaps could not be bridged because of highly divergent regions [13]. However, here we could demonstrate that it is possible to assemble complete chloroplast genome sequences from total leaf DNA with the SOLiD sequencing platform. To obtain regions from the chloroplast genome that diverged from a reference genome, *de novo* assembly was employed using paired reads. Before assembly, SOLiD reads from mitochondrial and

nuclear DNA, were filtered electronically. Furthermore, we could use the chloroplast genome of the closely related species *L. minor* as a reference that has been sequenced with traditional overlapping long reads [12]. Genome assembly, the comparative and phylogenetic analyses of these genomes are presented here.

3.3 Results

De novo assembly of short sequence reads yields high quality contigs

The sequenced species *S. polyrhiza*, *W. lingulata* and *W. australiana* in this study were selected to comprise the phylogenetic diversity of the subfamily *Lemnoideae* and also to represent their extensively variation of nuclear genome sizes (Table 3.1) [14]. The three genomes were sequenced using mate-paired libraries on the SOLiD^{TM} 3 System. The previously sequenced *L. minor* chloroplast genome was used as reference to retrieve chloroplast reads from the mixture of nuclei, mitochondria and chloroplasts. Considering the identical feature of two inverted repeats, we first assembled 136 Kb of the chloroplast genome from the LSC, IRa, SSC regions. All three genomes were each processed into one single large scaffold of 92 Kb (S.pol), 136Kb (W.lin), and 134 Kb (W.aus), respectively. Assembly of SOLiD reads resulted between 39 to 60 contigs and 1 to 3 scaffolds per genome (Table 3.2). With the second largest scaffold of 40 Kb for S.pol, the length of all the added contigs already reached a size expected for a chloroplast genome excluding the IRb region. However, alignment of these assemblies with the reference genome suggested between one to three misassembled scaffolds that needed to be corrected. Most contigs were interrupted by mononucleotide repeats and low complexity sequences.

Species	Nuclear Genome Size[b] (Mbp)	Chloroplast Genome Size (bp)	Inverted Repeats Size (bp)	Genbank #
S. polyrhiza 7498	160	168788	31755	JN160603
L. minor (Reference)[a]	356-604	165955	31223	DQ400350
W. lingulata 7289	655	169337	31683	JN160604
W. australiana 7733	357	168704	31930	JN160605

Table 3.1 Species used for comparative genomic analysis.

[a]Reference chloroplast genome [12]; [b]Nuclear genome sizes [14].

By comparison of assembly of chloroplast genome with or without selection of pure chloroplast reads, we found that both ways were accessible to get the complete genome. *De novo* assembly from total reads generated 60 - 82 contigs with 2333 – 4062 bp of N50 contig length, while assembly from pure chloroplast reads gave us better results: 18% to 35% less contig number but up to one time longer N50 of contig length (Table 3.2). Technically, 13 - 29 more PCR reactions need to close the gaps from total reads assembly than from pure chloroplast reads.

Using the ends of contigs separated by Ns, primers were designed for PCR amplification. Because of the alignment with the reference genome, the correct ordering of contigs could be confirmed by the fact that PCR amplification occurred. Furthermore, when PCR products were sequenced by the CE ABI 3730XL platform, overlapping sequences could be used to close gaps and validate the order of contigs. Accumulative overlaps for the three genomes totaled 48 Kb. When short read assemblies were compared with CE long read sequences, the cumulative differences amounted to just 0.041%, reflecting a high consensus between the two sequencing methods. We also could test the short read assembler by mapping *de novo* assemblies back to the complete

genome. Although only 2.5~12.9% of the reads were successfully aligned, keeping in mind the DNA mixture from plant tissue, this was sufficient to give a mean coverage between 1,070 to 5,474 times (Table 3.2 and Figure 3.1). The IRa and IRb regions had lower coverage due to random placement of repetitive read pairs when mapping. For nuclear genome sequences, we found 12 to 42-fold coverage by ignoring mitochondrial DNA reads (Table 3.2). Based on these assessments, there were approximately 100 chloroplast genome copies for every nuclear genome copy.

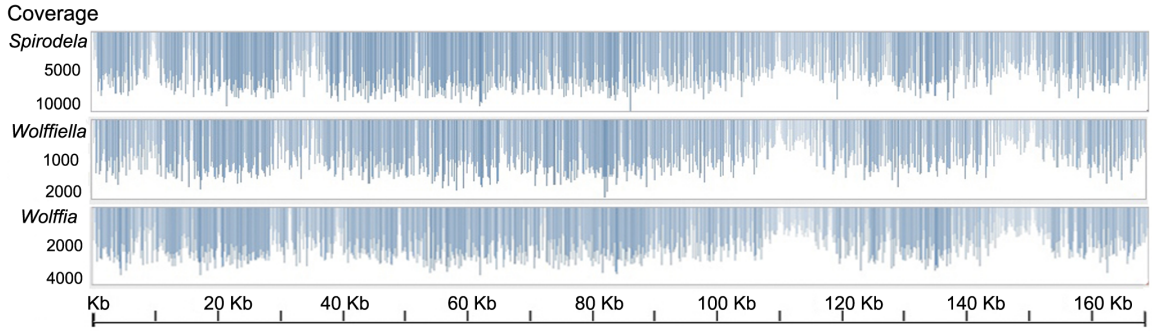


Figure 3.1 Coverage of *Lemnoideae* chloroplast genomes.

Depth of coverage was plotted along the genome coordinates. Blue peaks show the coverage.

Species	Read processing (with selection)	Scaffolds #	N50 scaffolds (bp)	Contigs #	N50 contigs (bp)	Sum contig length (bp)
Spirodela polyrhiza 7498	yes	3	92558	60	4246	136597

	no	6	36267	73	4062	132134
Wolffiella lingulata 7289	yes	1	136457	53	4708	139523
	no	8	25221	82	2333	133615
Wolffia australiana 7733	yes	2	134892	39	8677	137183
	no	3	98687	60	3743	132446

Species	**Read processing**	**Coverage cut-off[a]**	**Total reads (X10^6)**	**Aligned reads (%)**	**Chloroplast coverage**	**Nuclear coverage**
Spirodela polyrhiza 7498	yes	11	153	12.9	5474	42
	no	30				
Wolffiella lingulata 7289	yes	11	155	2.5	1070	12
	no	6				
Wolffia australiana 7733	yes	11	111	6.2	1912	15
	no	11				

Table 3.2 *De novo* assembly statistics.

[a]Coverage cut-off: minimum coverage required to form a contig.

Sequence comparison and phylogeny among *Lemnoideae* chloroplast genomes

The chloroplast genomes of duckweeds appeared to be within a short range of 165,955 bp to 169,337 bp in length (Table 3.1). All of them include a pair of inverted repeats of around 31 Kb separated by SSC and LSC. Large single copy (LSC) and Small Single Copy (SSC) regions were close to 90 Kb and 10 Kb long, respectively. *S. polyrhiza*, *W. lingulata* and *W. australiana* contain the same gene number and order as the reference genome *L. minor* . The representative map of *S. Polyrhiza* was shown here (Figure 3.2).

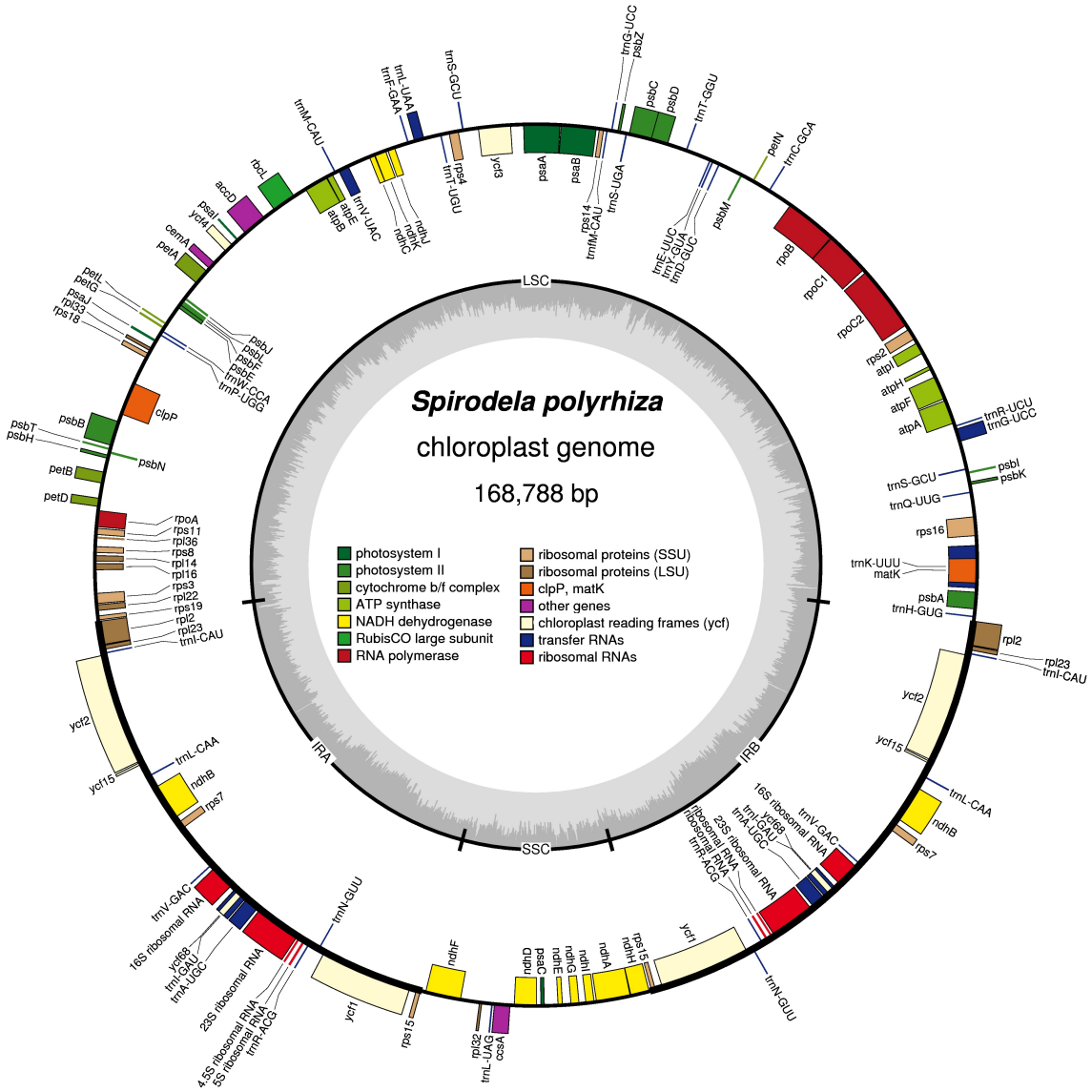


Figure 3.2 The chloroplast genome map of Spirodela.

The conservation of the overall structure of the chloroplast genomes allowed us to align the sequences of four duckweed species at the genome-wide level. Comparison of the sequences revealed multiple hotspots of high sequence length polymorphism (Figure 3.3). The IRs showed lower sequence divergence than the single-copy regions. The majority of highly divergent regions were in non-coding regions as illustrated in a

mVISTA alignment plot. The region between *rpoB* and *psbD* from position 28 Kb to 36 Kb is one of the most polymorphic regions. For example, *W. australiana* has a 425-bp deletion in the 29 Kb *rpoB*-tRNA-Cys region. *S. polyrhiza* has a 505-bp deletion compared with 100-bp deletions in *W. lingulata*, while a 353-bp insertion occurred at 31 Kb of the intergenic *petN-psbM* region of *W. australiana*. Both *W. lingulata* and *W. australiana* have a 460-bp deletion in the 32 Kb *psbM*-tRNA-Asp region. Moreover, some INDELs existed in introns, such as a 123-bp insertion in *atpF* of *Spirodela* at 13 Kb, and 114-bp deletion in *ndhA* for *W. lingulata* and 105-bp for *W. australiana* at the 132 Kb region (Figure 3.3).

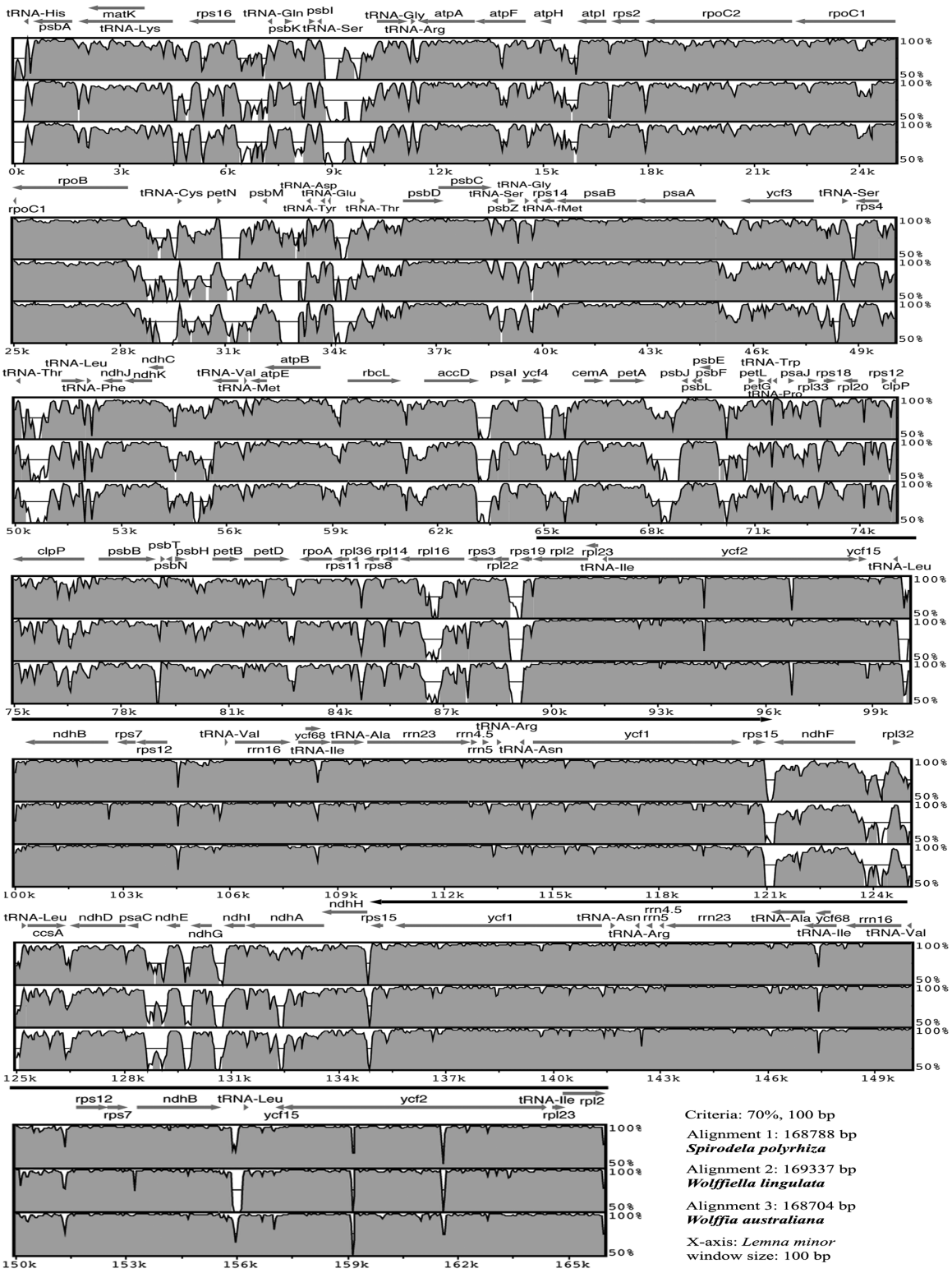


Figure 3.3 Alignment of *Lemnoideae* chloroplast genomes.

The sequence of *L. minor* chloroplast genome was compared to those of *S. polyrhiza* (top), *W. lingulata* (middle), *W. australiana* (bottom). Sequences were aligned in mVISTA and the annotation shown above the alignment corresponds to the *L. minor* genome. Grey arrows above the alignment indicate genes and their orientation. Thick black lines show the position of the IRs. The grey peaks determine the percent identity between two sequences of *L. minor* as the reference and our sequenced genomes.

Maximum parsimony produced a single fully resolved tree with strong node support (Figure 3.4). Our phylogenetic results showed *Wolffiella* and *Wolffia* were more closely related than the others. Furthermore, our analysis strongly supported that *Spirodela* was at the basal position of the taxon, followed by *Lemna* and *Wolffiella*, whereas *Wolffia* was the most derived (Figure 3.4).

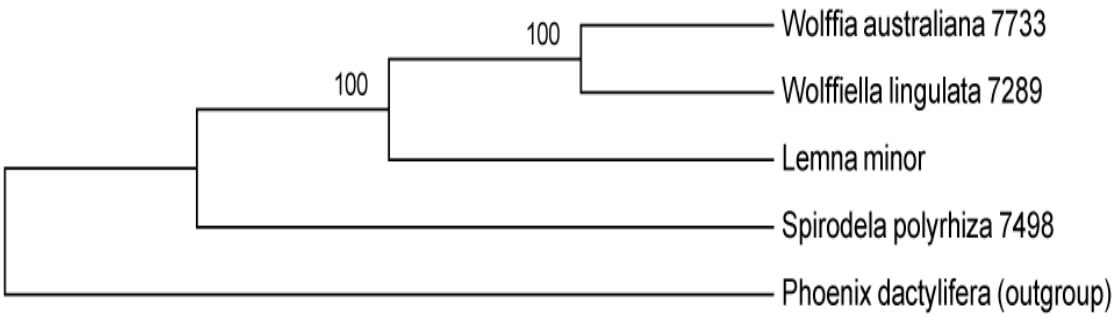


Figure 3.4 Complete chloroplast genome phylogeny of *Lemnoideae*.

The phylogram was drawn by Maximum Parsimony with 1000 replicates of bootstrap test. The tree was rooted by *Phoenix dactylifera* as an outgroup. Support from bootstrap value was shown at the nodes. The GenBank accessions used for the analyses are JN160603 (*S. polyrhiza*), DQ400350 (*L. minor*), JN160604 (*W. lingulata*), JN160605 (*W. australiana*) and GU811709 (*P. dactylifera*). The whole genome sequences were aligned by Multi-LAGAN and MEGA 5 was used to draw the tree.

Evolution of *Lemnoideae* and *Pooideae*, with chloroplast genomes in different orders

To further evaluate the pace of evolutionary divergence, we compared chloroplast genomes from different monocot orders by quantifying nucleotide substitution rates and INDELs ratios. The subfamily of *Pooideae* within the *Poaceae* belongs to the order of the Poales, whereas the *Lemnoideae* belong to the order of the Alismatales. When such a comparison is made, duckweeds have a higher rate of substitution than species of the *Pooideae* at the whole genome level and in protein-coding regions. Moreover, INDELs were very prominent in duckweed genomes with ratios of 0.061 to 0.095, whereas it is much lower as expected in conservative coding regions from 0.006 to 0.012. When we compared duckweeds with species of the *Pooideae*, duckweeds had twice as many INDELs in their chloroplast genomes than the *Pooideae's* species based on the same level of intra-tribe or inter-tribe comparisons (Table 3.3). Based on INDELs length in genome and coding regions (Table 3.3), we could conclude that most INDELs were located in non-coding regions. Interestingly, we found that transversions were higher than transitions in the subfamily of *Lemnoideae* with R-values from 0.6 to 0.7 of the total genome. The same result was discovered in protein coding regions except between *S. polyrhiza* and *L. minor* (R=1.1). However, these values were completely the opposite in the species of the subfamily of *Pooideae* with R-values from 1.2 to 1.7, where transitions were more numerous than transversions (Table 3.3).

Comparative Type	Alignment Region	Pair Alignment	Alignment Length	Substitution Rate[a]	R= si[b]/sv[c]	INDELs Length	INDELs Ratio[d]
intra-tribe	whole genome	S.pol+L.min	141014	0.05	0.7	10262	0.073
intra-tribe	whole genome	W.lin+W.aus	141506	0.04	0.6	8635	0.061
inter-tribe	whole genome	S.pol+W.lin	143722	0.07	0.6	12757	0.089
inter-tribe	whole genome	S.pol+W.aus	142828	0.07	0.6	11849	0.083
inter-tribe	whole genome	L.min+W.lin	142965	0.07	0.6	13543	0.095
inter-tribe	whole genome	L.min+W.aus	141968	0.07	0.6	12429	0.088

intra-tribe	whole genome	wheat+barley	115940	0.02	1.2	4365	0.038
inter-tribe	whole genome	wheat+B.dis	117055	0.04	1.2	6615	0.057
inter-tribe	whole genome	barley+B.dis	116768	0.04	1.3	6196	0.053
intra-tribe	81 Protein genes	S.pol+L.min	69247	0.03	1.1	420	0.006
intra-tribe	81 Protein genes	W.lin+W.aus	69503	0.03	0.8	633	0.009
inter-tribe	81 Protein genes	S.pol+W.lin	69539	0.04	0.9	819	0.012
inter-tribe	81 Protein genes	S.pol+W.aus	69459	0.04	0.9	682	0.01
inter-tribe	81 Protein genes	L.min+W.lin	69521	0.04	0.9	831	0.012
inter-tribe	81 Protein genes	L.min+W.aus	69468	0.04	0.9	748	0.011
intra-tribe	71 Protein genes	wheat+barley	58607	0.01	1.5	290	0.005
inter-tribe	71 Protein genes	wheat+B.dis	58658	0.03	1.7	1045	0.018
inter-tribe	71 Protein genes	barley+B.dis	58647	0.03	1.7	1034	0.018

Table 3.3 Pairwise sequence divergence of *Lemnoideae* and *Pooideae*.

[a]Substitution Rates = substitution/alignment length; [b]si (Transitional Pairs) = AG+CT; [c]sv (Transversional Pairs) = TA+TG+CA+CG; [d]INDELs Ratio = INDELs length/alignment length. AG means A is mutated to G and others follow the same rules. S.pol = *S. polyrhiza*, L.min = *L. minor*, W.lin = *W. lingulata*, W.aus = *W. australiana,* B.dis = *B. distachon*

3.4 Discussion

Next generation sequencing platforms have mainly been used for re-sequencing, SNP analysis, and expression profiling because it has been difficult to develop *de novo* assembly tools for short sequence reads [15]. Whereas re-sequencing or sequencing of related genomes can be very productive for SNP detection and for map-based cloning of mutant alleles, short-read assemblies often fail to detect large INDELs and variable regions in new genomes because technically there is no reference for them. *De novo* assemblies of short reads could cover all insertions, deletions, and rearrangements that would otherwise be incorrectly assembled based on alignments with a reference genome [11]. The pipeline of the SOLiD™ System *de novo* Accessory Tools 2.0, however, has

been well adapted to assemble high-coverage SOLiD reads of microbial genomes [16]. Because chloroplasts are even smaller than bacterial genomes, more in the order of large viruses, they represent an exception where such method can be applied. Moreover, we could use paired reads from the same DNA fragment to anchor one end to a contig and the other to a gap that could overlap with other unanchored ends. For this purpose, we used a module Assembly Assistant for SOLiD™ to maximally fill gaps in scaffolds by sufficiently utilizing benefits of these paired ends (http://solidsoftwaretools.com/gf/project/denovo/). Indeed, we got good assemblies by using high quality reads and minimizing non-target DNA from read mixtures. However, interference for contig building arose mainly from long mononucleotide repeats and low complexity sequence. Final mapping of SOLiD reads back to the complete chloroplast genome yielded only 2.5~12.9% alignment due to 1,000 times smaller genome size than nuclear genome. After comparison of assembly from pure chloroplast reads with that from total reads, we conclude that it is better to select the chloroplast reads before assembly if a reference is available. If without any reference, then the minimum coverage required to form a contig (coverage cut-off) for Velvet needs to be determined which allows only higher coverage of chloroplast reads rather than ones from much lower coverage of nuclei and mitochondria genome to go into assembly. Exploration of different filters, however, could be used to mask chloroplast sequences to assemble either nuclear or mitochondrial genomic DNA in parallel from the same data set, given a deep enough genome coverage.

It is generally assumed that there is a universal transition bias over transversion, probably as a consequence of the fundamental biochemical basis of mutations [17]. This

rule appears to hold quite well in many vertebrate species [18] and it also works very well in the *Pooideae* subfamily as we have calculated here. Surprisingly, this is not the case for the *Lemnoideae* subfamily, where a transition bias is absent. Although there is an exemption of transition bias in coding regions of *Spirodela* and *Lemna*, which could be explained by a selection of nonsynonymous substitutions. If all type of substitutions were to be equal, a 1:2 ratio of transition/transversion would be expected because of two possibilities of transitions (AG+CT) and four of transversions (AT+AC+GT+GC). Excluding nucleotide mutations in coding regions from whole genomes of duckweed chloroplasts, the number of R-values for non-coding region would be very close to 0.5. In such a case, there would be no significant difference between transition and transversion rates. However, in a study of grasshopper pseudogenes a transition/transversion bias was not universal and both substitution rates reached a 1:1 ratio [19]. Interestingly, transversions could also occur more frequently than transitions in chloroplasts of green algae [20].

Despite the overall high conservation of genome content across different duckweed species, our results demonstrate that substitution rates, insertion and deletion events are more frequent in duckweed chloroplast genomes than in species of the *Pooideae*, especially in non-coding regions (Table 3.3, Figure 3.3). Recent studies also support the observation that *Lemnoideae* have a higher rate of chloroplast sequence evolution relative to *Pistia* and related *Araceae* [21].

Nucleotide substitutions and INDEL mutations are generated during DNA replication or are due to DNA damage [22, 23]. Although the enzymes responsible for the

maintenance of chloroplast replication and DNA repair are highly faithful, under certain conditions chloroplasts may have to tolerate some level of oxidative damage that occurs spontaneously due to an abundance of reactive oxygen species from the water-splitting activity of the photosystem [20]. Because duckweeds float on water surface, are fully exposed to sunlight, and produce biomass at such a fast rate, their plastid genomes probably transmit and accumulate mutations more frequently than other plants. Once the genome of *Spirodela* has been sequenced, it will be interesting to analyze its nuclear genes that are involved in DNA replication and repair of the plastid genome and how they have evolved compared to terrestrial slow growing plants.

So far, all phylogeny constructions of *Lemnoideae* have used selected genes or partial regions as markers. However, with sequenced chloroplast genomes of four species in this subfamily and the powerful program to align them, it is possible for the first time to perform whole chloroplast genome phylogenetic analysis. The topology of nodes, all with 100% bootstrap values, conforms to the accepted phylogeny based on extensive analysis from morphology and DNA sequence markers. However, there were two nodes that were problematic with only 42% and 53% bootstrap values in *Wolffia* [24]. Therefore, our results contradict the hypothesis that *Wolffia* arose from a merger of *Wolffiella* and *Lemna*, which was based on the *trnL-trnF* marker only [21]. Clearly, the addition of more informative sites from whole genome sequences will improve resolution and confidence in phylogenetic analyses.

In summary, our data gave evidence that next-generation platforms have the capacity to sequence the chloroplast genome at over 1,000 times coverage in an

individual spot on a quadrant slide without purification (Table 3.2). In order to gain an improved understanding of genome evolution in members of the duckweed subfamily, we generated chloroplast genomes for three species from different genera using *L. minor* as a reference. Our analysis further suggests that (i) gene content is very conserved in duckweeds; (ii) Fast nucleotide substitution and abundant INDELs played a key role in the evolution of chloroplast genomes of duckweeds; (iii) duckweed chloroplast genome sequences are very promising to become an elusive single-locus plant barcode for systematic analysis. This information will be critical for the development of chloroplast transformation for improving duckweed biomass.

3.5 Materials and methods

DNA isolation and SOLiD DNA sequencing

Duckweed was grown from a cluster of 3-5 fronds produced by a single mother frond. Total DNA was extracted from whole plant tissue by the CTAB method [25]. Sequencing runs were done on a SOLiD™ 3 Analyzer (Applied Biosystems, Foster City, CA) at the Waksman Genomics Core Facility of Rutgers University. Mate-paired libraries with approximately 1.5 Kb inserts were constructed from 20 μg of genomic DNA following the manufacturer's instructions (SOLiD sample preparation protocol for Mate-Paired library sequencing), and deposited on one spot of a quadrant slide. Fifty base reads were obtained from each of the F3 and R3 tags, with more than 100 million reads obtained for each of the genomes.

Sequence data analysis pipeline

To assemble the chloroplast genomes using SOLiD reads and close the remaining gaps with long reads from capillary electrophoresis (CE) sequencers, we used the following steps (Figure 3.5). Because all chloroplast genomes contain two identical inverted repeats (IRs), we first assembled genomes without IRb's using 136 Kb of *L. minor* as a reference, including LSC, SSC, IRa, but added them later on for the full-length molecules.

1) Data filtering: SOLiD mate-paired short reads were preprocessed by Mean Filter of a Perl script [26]; i.e., reads were truncated to 40 bp and average quality of reads were set to exceed the threshold QV score of 20. Because coverage is very high, only successful mate-pair reads went into the next step. 2) Selection of chloroplast-related reads: The filtered mate-pair colorspace reads from each of the three samples were aligned to the chloroplast genome of *L. minor* [12] (GenBank accession number: DQ400350) using the BWA short-read alignment component with default parameters [27]. At least one end of the paired-end reads was anchored to the chloroplast genome of *L. minor* before interrogating the second end to map to a linked sequence or to a gap. 3) 1st run of genome assembly: *De novo* assembly was performed with identified chloroplast-related reads using the SOLiD™ System *de novo* Accessory Tools 2.0 (http://solidsoftwaretools.com/gf/project/denovo/), together with the Velvet assembly engine [28]. The tools are designed to simplify and optimize parameters for ease of usage and best performance. They sample an optimal sub-set of reads and automatically estimate optimal parameters for each step. Velvet parameters generated from the tools

were deposited in Table 3.2 with hash length 19 and coverage cut-off 11. Assembly assistant module in the tools took the input from Velvet and produced scaffolds with 120 mate-pair confirmations to make confident scaffolding in the end. 4) 2nd run of genome assembly: After the first run, all scaffolds were concatenated into pseudomolecules. In order to maximize chloroplast-related reads, the artificial molecule functioned as a new reference and step 2 and 3 were then reiterated. 5) Correction of scaffold building: The biggest scaffolds of each genome were aligned with the most closely related reference genome of *L. minor* using BLAST2 (http://blast.ncbi.nlm.nih.gov/). Indeed in a few instances, non-contiguous genomic regions were found in juxtaposed positions at gap positions. At these gaps scaffolds were broken and contigs reordered in collinearity with the reference genome. Smaller contigs were manually ordered based on the reference genome. All scaffolds were then concatenated into a single full-length molecule, where each gap in the sequence was marked with one N. 6) Gap closure: Gaps were small enough so that flanking primer pairs could be chosen (Table S1) to isolate missing sequences by PCR and apply CE sequencing methods (ABI 3730XL) for closure [7]; 7) Assembly validation: Because PCR amplification of gaps required correct ordering of contigs into scaffolds, the long CE reads provided validation of overlapping sequences and the correct ordering of short read assemblies. Accumulative overlaps and discrepancies between alignments of sequences from both methods were summarized using DNASTAR (http://www.dnastar.com/), which could reflect sequence errors in the SOLiD platform. 8) GenBank deposition: The fully sequenced genomes of three species were annotated by DOGMA [29], checked manually and have been deposited to GenBank as a whole genome shotgun project.

Considering its universal application of this pipeline, we would like to know how efficient *de novo* assembly of chloroplast genome from total reads could be without such good reference-based selection. In this way, we couldn't use default set-up parameters for the pipeline that is only feasible for uniform coverage of a genome, because the precomputed parameters will be shifted from sub-set reads which was actually a mixture of three genomes with different coverage. Instead of step 2, 3 and 4 after data filtering, we determined the optimized parameters shown in Table 3.2 by extensively initial testing and then manually provided to the SOLiD™ System *de novo* Accessory Tools 2.0. All other assembly steps were the same as the one with selected reads.

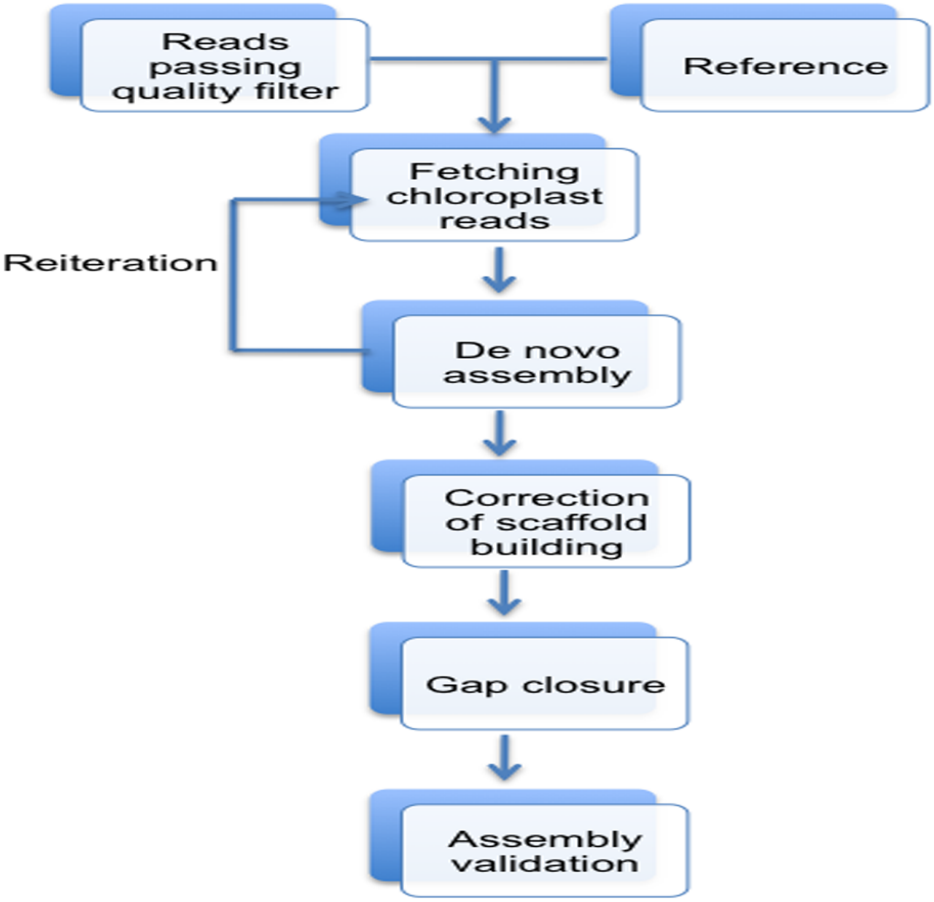


Figure 3.5 Pipeline of chloroplast genome assembly.

Whole genome alignments, comparison, and phylogenetic analysis

Lemnoideae chloroplasts, *S. polyrhiza* 7498 (S.pol), *L. minor* (L.min), *W. lingulata* 7289 (W.lin), *W. australiana* 7733 (W.aus) were aligned by a program of global multiple alignment of finished sequences (Multi-LAGAN) [30] and annotation for the reference genome of *L. minor* [12] was used to construct sequence conservation plots in the program mVISTA [31].

The 81 protein coding nucleotide sequences from duckweeds were retrieved after annotation by DOGMA, concatenated as one full-length molecule and pair-wisely aligned with each other by Multi-LAGAN. MEGA 5 was used to detect transitions, transversions, and INDELs (insertion/deletion) for all genomes except the IRb regions and protein coding sequences. A similar analysis of 71 common genes was done for chloroplast genomes of species in the subfamily of the *Pooideae*, i.e., wheat (AB042240), barley (EF115541) and *Brachypodium* (EU325680). They were chosen because wheat and barley belong to the same tribe of *Triticeae*, while *Brachypodium* belongs to the different tribe of *Brachypodieae* within the same subfamily, which taxonomically is at the same level as the *Lemnoideae,* where *Spirodela* and *Lemna* belong to the same tribe, but *Wolffiella* and *Wolffia* to a different one [6, 24].

To examine if the genome-wide phylogenetic analyses are consistent with those of morphological, flavonoid, and allozyme markers, as well as selected DNA sequences [24], we employed Maximum Parsimony to reconstruct *Lemnoideae* phylogeny with whole chloroplast sequences by using MEGA 5 [32]. *Phoenix dactylifera* is in the same

class of *Liliopsida* as *Lemnoideae* and functions as an outgroup here [33]. However, one of the two inverted repeat regions (IRb) was excluded from phylogenetic analyses.

Acknowledgement

This work has been published in PLoS ONE of 2011 by authors of Wang W, Messing J with the title of "High-Throughput Sequencing of Three Lemnoideae (Duckweeds) Chloroplast Genomes from Total DNA".

3.6 References

1. Palmer JD: **Comparative Organization of Chloroplast Genomes**. *Annu Rev Genet* 1985, **19**(1):325-354.
2. Shaver J, Oldenburg D, Bendich A: **Changes in chloroplast DNA during development in tobacco, Medicago truncatula, pea, and maize**. *Planta* 2006, **224**(1):72-82-82.
3. Lutz K, Wang W, Zdepski A, Michael T: **Isolation and analysis of high quality nuclear DNA with reduced organellar DNA for plant genome sequencing and resequencing**. *BMC Biotechnology* 2011, **11**(1):54.
4. Bortiri E, Coleman-Derr D, Lazo G, Anderson O, Gu Y: **The complete chloroplast genome sequence of Brachypodium distachyon: sequence comparison and phylogenetic analysis of eight grass plastomes**. *BMC Research Notes* 2008, **1**(1):61.
5. Bock R: **Plastid biotechnology: prospects for herbicide and insect resistance, metabolic engineering and molecular farming**. *Curr Opin Biotechnol* 2007, **18**:100-106.
6. Cabrera LI, Salazar GA, Chase MW, Mayo SJ, Bogner J, Davila P: **Phylogenetic relationships of aroids and duckweeds (Araceae) inferred from coding and noncoding plastid DNA**. *Am J Bot* 2008, **95**(9):1153-1165.
7. Wang W, Wu Y, Yan Y, Ermakova M, Kerstetter R, Messing J: **DNA barcoding of the Lemnaceae, a family of aquatic monocots**. *BMC Plant Biology* 2010, **10**:205.
8. Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C *et al*: **Whole-genome sequencing of multiple Arabidopsis thaliana populations**. *Nat Genet* 2011, **43**(10):956-963.

9. Kuehdorf K, Jetschke G, Ballani L, Appenroth K: **The clonal dependence of turion formation in the duckweed Spirodela polyrhiza-an ecogeographical approach.** *Physiol Plant* 2013, **10**.
10. Appenroth K, Borisjuk N, Lam E: **Telling duckweed apart: genotyping technologies for the Lemnaceae**. *Chin J Appl Environ Biol* 2013, **19**(1):1-10.
11. Cronn R, Liston A, Parks M, Gernandt DS, Shen R, Mockler T: **Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology**. *Nucleic Acids Res* 2008, **36**(19):e122-e122.
12. Mardanov A, Ravin N, Kuznetsov B, Samigullin T, Antonov A, Kolganova T, Skyabin K: **Complete sequence of the duckweed (Lemna minor) chloroplast genome: structural organization and phylogenetic relationships to other angiosperms**. *Journal of Molecular Evolution* 2008, **66**(6):555-564.
13. Nock CJ, Waters DLE, Edwards MA, Bowen SG, Rice N, Cordeiro GM, Henry RJ: **Chloroplast genome sequences from total DNA for plant identification**. *Plant Biotech J* 2010, **9**(3):328-333.
14. Wang W, Kerstetter R, Michael T: **Evolution of genome size in duckweeds (Lemnaceae)**. *Journal of Botany* 2011(Special Issues).
15. Paszkiewicz K, Studholme DJ: **De novo assembly of short sequence reads**. *Brief Bioinform* 2010, **11**(5):457-472.
16. den Bakker H, Cummings C, Ferreira V, Vatta P, Orsi R, Degoricija L, Barker M, Petrauskene O, Furtado M, Wiedmann M: **Comparative genomics of the bacterial genus Listeria: Genome evolution is characterized by limited gene acquisition and limited gene loss**. *BMC Genomics* 2010, **11**(1):688.
17. Wakeley J: **The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance**. *Trends Ecol Evol* 1996, **11**(4):158-162.
18. Nachman MW, Crowell SL: **Estimate of the mutation rate per nucleotide in humans**. *Genetics* 2000, **156**(1):297-304.
19. Keller I, Bensasson D, Nichols RA: **Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes**. *PLoS Genet* 2007, **3**(2):e22.
20. Guhamajumdar M, Sears BB: **Chloroplast DNA base substitutions: an experimental assessment**. *Mol Genet Genomics* 2005, **273**(2):177-183.
21. Rothwell GW, Van Atta MR, Ballard HE, Stockey RA: **Molecular phylogenetic relationships among Lemnaceae and Araceae using the chloroplast trnL-trnF intergenic spacer**. *Mol Phylogenet Evol* 2004, **30**(2):378-385.
22. Friedberg EC: **DNA damage and repair**. *Nature* 2003, **421**(6921):436-440.
23. Magee AM, Aspinall S, Rice DW, Cusack BP, Semon M, Perry AS, Stefanovic S, Milbourne D, Barth S, Palmer JD *et al*: **Localized hypermutation and associated gene losses in legume chloroplast genomes**. *Genome Res* 2010, **20**(12):1700-1710.
24. Les DH, Crawford DJ, Landolt E, Gabel JD, Kimball RT: **Phylogeny and Systematics of Lemnaceae, the Duckweed Family**. *Systematic Botany* 2002, **27**(2):221-240.
25. Murray MG, Thompson WF: **Rapid isolation of high molecular weight plant DNA**. *Nucleic Acids Res* 1980, **8**(19):4321-4325.

26. Sasson A, Michael TP: **Filtering error from SOLiD Output**. *Bioinformatics* 2010, **26**(6):849-850.
27. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.
28. Zerbino DR, Birney E: **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs**. *Genome Res* 2008, **18**(5):821-829.
29. Wyman SK, Jansen RK, Boore JL: **Automatic annotation of organellar genomes with DOGMA**. *Bioinformatics* 2004, **20**(17):3252-3255.
30. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Program NCS, Green ED, Sidow A, Batzoglou S: **LAGAN and Multi-LAGAN: Efficient Tools for Large-Scale Multiple Alignment of Genomic DNA**. *Genome Res* 2003, **13**(4):721-731.
31. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I: **VISTA: computational tools for comparative genomics**. *Nucleic Acids Res* 2004, **32**(suppl 2):W273-W279.
32. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods**. *Mol Biol Evol* 2011.
33. Yang M, Zhang X, Liu G, Yin Y, Chen K, Yun Q, Zhao D, Al-Mssallem IS, Yu J: **The Complete Chloroplast Genome Sequence of Date Palm (Phoenix dactylifera L.)**. *PLoS ONE* 2010, **5**(9):e12762.

CHAPTER 4 MITOCHONDRIAL GENOMICS

4.1 Abstract

Spirodela polyrhiza is a species of the order Alismatales, which represent the basal lineage of monocots with more ancestral features than the Poales. Its complete sequence of the mitochondrial (mt) genome could provide clues for the understanding of the evolution of mt genomes in plant.

Spirodela polyrhiza mt genome was sequenced from total genomic DNA without physical separation of chloroplast and nuclear DNA using the SOLiD platform. Using a genome copy number sensitive assembly algorithm, the mt genome was successfully assembled. Gap closure and accuracy was determined with PCR products sequenced with the dideoxy method.

This is the most compact monocot mitochondrial genome with 228,493 bp. A total of 57 genes encode 35 known proteins, 3 ribosomal RNAs, and 19 tRNAs that recognize 15 amino acids. There are about 600 RNA editing sites predicted and three lineage specific protein-coding-gene losses. The mitochondrial genes, pseudogenes, and other hypothetical genes (ORFs) cover 71,783 bp (31.0%) of the genome. Imported plastid DNA accounts for an additional 9,295 bp (4.1%) of the mitochondrial DNA. Absence of transposable element sequences suggests that very few nuclear sequences have migrated into Spirodela mtDNA. Phylogenetic analysis of conserved protein-coding genes suggests that Spirodela shares the common ancestor with other monocots, but there is no obvious synteny between Spirodela and rice mtDNAs. After eliminating genes, introns, ORFs, and plastid-derived DNA, nearly four-fifths of the Spirodela

mitochondrial genome is of unknown origin and function. Although it contains a similar chloroplast DNA content and range of RNA editing as other monocots, it is void of nuclear insertions, active gene loss, and comprises large regions of sequences of unknown origin in non-coding regions. Moreover, the lack of synteny with known mitochondrial genomic sequences shed new light on the early evolution of monocot mitochondrial genomes.

4.2 Introduction

Usually, a plant cell contains three genomes: plastid, mitochondrial, and nuclear. In a typical *Arabidopsis* leaf cell, there are about 100 copies of mitochondrial DNA (mtDNA), about 1,000 copies of chloroplast DNA (cpDNA), and two copies of nuclear DNA (ncDNA) [1].

The mitochondrial genome plays fundamental roles in development and metabolism as the major ATP production center via oxidative phosphorylation [2]. The mitochondrial genetic system in flowering plants exhibit multiple characteristics that distinguish them from other eukaryotes: large genome size with dispersed genes, an incomplete set of tRNAs, trans-splicing, and frequent uptake of plastid DNA or of foreign DNA fragments by horizontal and intracellular gene transfer [2-6]. Plant mtDNAs are a major resource for evolutionary studies, because coding regions evolve slowly, in contrast to the flexible non-coding DNA. Therefore, the structural evolution and plasticity of plant mtDNAs make them powerful model for exploring the forces that affect their divergence and recombination.

With the emergence of second-generation sequencing technologies, the number of completed plant mitochondrial genomes deposited in the GenBank database (http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid=33090&opt=organelle) has increased until August of 2012 to 69. Most are from Chlorophyta (17 of green algae) and seed plants (26 of eudicotyledons). So far, among 11 sequenced monocot mt genomes, 10 are from the Poales, which have been extensively studied and only one, *Phoenix*, a palm, from the order of Arecale has been sequenced [7]. Obviously, complete mt sequence data will be needed not only from closely but also distant related taxa to give us a broader perspective of mt genome organization and evolution.

Spirodela polyrhiza, with great potential for industrial and environmental applications, is a small, fast growing aquatic plant in the Araceae family of the Alismatales order [8, 9]. There are 14 families, 166 genera, and about 4,500 species in this order. The early diverging phylogenetic position of Alismatales offers a broader view at features of monocot mt genomes. Plant mitochondria could also open a strategy for transgenes with high expression level and biological containment because of their maternal inheritance [10]. Here, we demonstrate the *de novo* assembly of a complete mt genome sequence from total leaf DNA using the SOLiD sequencing platform and a genome copy number-sensitive algorithm that can filter chloroplast and nuclear sequences. Indeed, comparative analysis of this genome provides us with unique features and new insights of this class of plants that differ from other monocots.

4.3 Results and discussion

The *de novo* assembly of SOLiD reads

The optimal parameter of the SOLiD™ System *de novo* Accessory Tools 2.0 for the assembly of the Spirodela mitochondrial genome has a hash length of 25 and coverage cut-off of 45. Under these conditions, assembly of SOLiD reads from total leaf DNA resulted in 15 scaffolds and 88 contigs, of which three scaffolds were mitochondrial (173,697, 47,896, 1824 bp) (Table 4.1). As expected, the other scaffolds were mainly copies of ribosomal RNA genes and retroelements of the nuclear genome because their copy number was comparable to the copies of mitochondrial genomes per leaf cell. To validate the assemblies, gaps were amplified with PCR for dideoxy sequencing with the CE ABI 3730xl system. With this information the order of the three scaffolds were resolved. Furthermore, after the SOLiD short read assembly was aligned with the CE long read sequences, only 0.036% discrepancy was found within 19 Kb sequence of overlaps, demonstrating high consistency between the two platforms. When we mapped the total reads back to the complete mtDNA, a total of 467-fold coverage was calculated. Considering the 5,474-fold chloroplast coverage, we found 41-fold coverage of nuclear genome sequences (Table 4.1). This level of coverage from assembled sequences was consistent with the expected representation of the three genomes in total leaf DNA, yielding chloroplast, mitochondria, and nuclei with the approximate ratio of 100:10:1.

Statistical list	**Number**
Number of scaffolds	15
N50 scaffolds (bp)	173697
Number of contigs	88
N50 contigs (bp)	6528
Sum contig length (bp)	240987

Hash length	25
Expected coverage	90
Coverage cut-off[a]	45
Total reads (X10^6)	153
Aligned reads (%)	1.4
Average chloroplast coverage[b]	5474
Average mitochondrial coverage	467
Average nuclear coverage	41

Table 4.1 ***de novo*** **assembly statistics.**

[a]Coverage cut-off: minimum coverage required to form a contig.

[b]Average chloroplast coverage was cited from its genome assembly [11].

Here, we applied a layered approach of sequencing organelle genomes without fractionation from total leaf DNA. Thanks to an assembly algorithm of sequence reads that is sensitive to the differential copy number of organelle and nuclear genomes, we did not physically need to fractionate plastid, mitochondrial, and nuclear DNA for deep sequencing. Therefore, we first assembled the complete Spirodela chloroplast genome from ABI SOLiD and gap-closure 3730xl reads, which permitted us to mask all plastid DNA reads before assembling mitochondrial DNA, which is in access of nuclear DNA but not as abundant as plastid DNA [11, 12]. Furthermore, we can take advantage of the ratios of these genomes to limit the value of coverage cut-off with identical dataset of SOLiD reads, which is taken in consideration for the assembly algorithm to distinguish between plastid, mitochondrial, and nuclear genome sequence reads [13, 14]. Assemblies were validated like in the case of chloroplast DNA by PCR and gap sequencing of long

reads with the traditional ABI 3730xl sequencing system. Following this protocol, we obtained a complete mitochondrial genome from an aquatic plant in a very cost-efficient way, which can serve as a reference for future mt genomics.

Features of the Spirodela **mitochondrial genome**

The mitochondrial genome was assembled into a 228,493 bp master circle (Figure 4.1), which makes it the smallest genome of all sequenced monocots, much smaller than the 715,001 bp of *Phoenix dactylifera* [7], 490,520 bp of *Oryza sativa* [15], or 569,630 bp of *Zea mays* mitochondria [16]. Because Spirodela diverged at a very early stage in the monocot lineage, it suggests that either the common ancestor of monocots had a relatively compact genome, with a series of independent expansions by accumulation of chloroplast and nuclear sequences or proliferation of pairs of repeats, leading to the large genomes in rice and maize [5, 15, 16], or a number of size contractions happened in Spirodela from the large genome of their ancestor. The GC content in the mtDNA was 45.7%, slightly higher than 43.8% of *Oryza* and 43.9% of *Zea* [15, 16]. The coding sequences covered 31% of the mitochondrial genome compared with 57.4% of the chloroplast genome [11] (Table 4.2). There were 57 functional genes and 4 pseudogenes in total, encoding 35 proteins, 19 tRNAs and 3 rRNAs (Table 4.3). Therefore, it gave rise to a density of 4.0 Kb per gene. Noticeably, eight genes (*ccmFc*, *cox2*, *nad1*, *nad2*, *nad4*, *nad5*, *nad7*, *rps3*) had 15 *cis*-spliced group II introns, whereas *nad1*, *nad2* and *nad5* were disrupted by 6 *trans*-splicing sites (Table 4.2). Previous studies suggested that *trans*-splicing had evolved before the emergence of hornworts [17]. In general, the numbers

and locations of introns in the Spirodela mtDNA were rather well conserved in other sequenced monocot genomes.

Feature	Value
Genome size (bp)	228,493
GC content (%)	45.7
Coding sequences (%)[a]	31.4
Protein coding gene #	35
ORFs #	39
cis-/trans-intron #	15/5
tRNA gene #	19
rRNA gene #	3
Chloroplast-derived (%)	4.1
Gene density (bp)	4009

Table 4.2 Features for mitochondrial genome.

[a]coding sequences include identified mitochondrial genes, pseudogenes, ORFs and cis-spliced introns.

Protein genes and transcript editing

The content of key protein coding genes in Spirodela mtDNA is highly conserved with other angiosperms [16, 18-20]. There were nine subunits of the oxidative phosphorylation complex I (*nad1*, 2, 3, 4, 4L, 5, 6, 7 and 9); one subunit of complex II (*sdh4*); one subunit of complex III (*cob*); three subunits of complex IV (*cox1*, *cox2* and *cox3*); five subunits of complex V (*atp1*, 4, 6, 8 and 9); and four subunits of a complex involved in cytochrome c biogenesis (*ccmB*, *ccmC*, *ccmFn* and *ccmFc*). Other genes encoding maturase (*matR*) and transport membrane protein (*mttB*) were also present in Spirodela mtDNA. As in maize [16], the matR gene in Spirodela also resided in the intron 4 of *nad1*, which is trans-spliced after transcription. In Spirodela, there were ten functional ribosomal genes and two pseudogenes of *rps14* and *rps19* with early stop

codons, whereas rice had a functional *rps19* and a non-functional *rps14* [15] and both were missing in maize (Table 4.3) [16]. All annotated genes and coordinates were listed in Table 4.3 and shown in a graphical map (Figure 4.1).

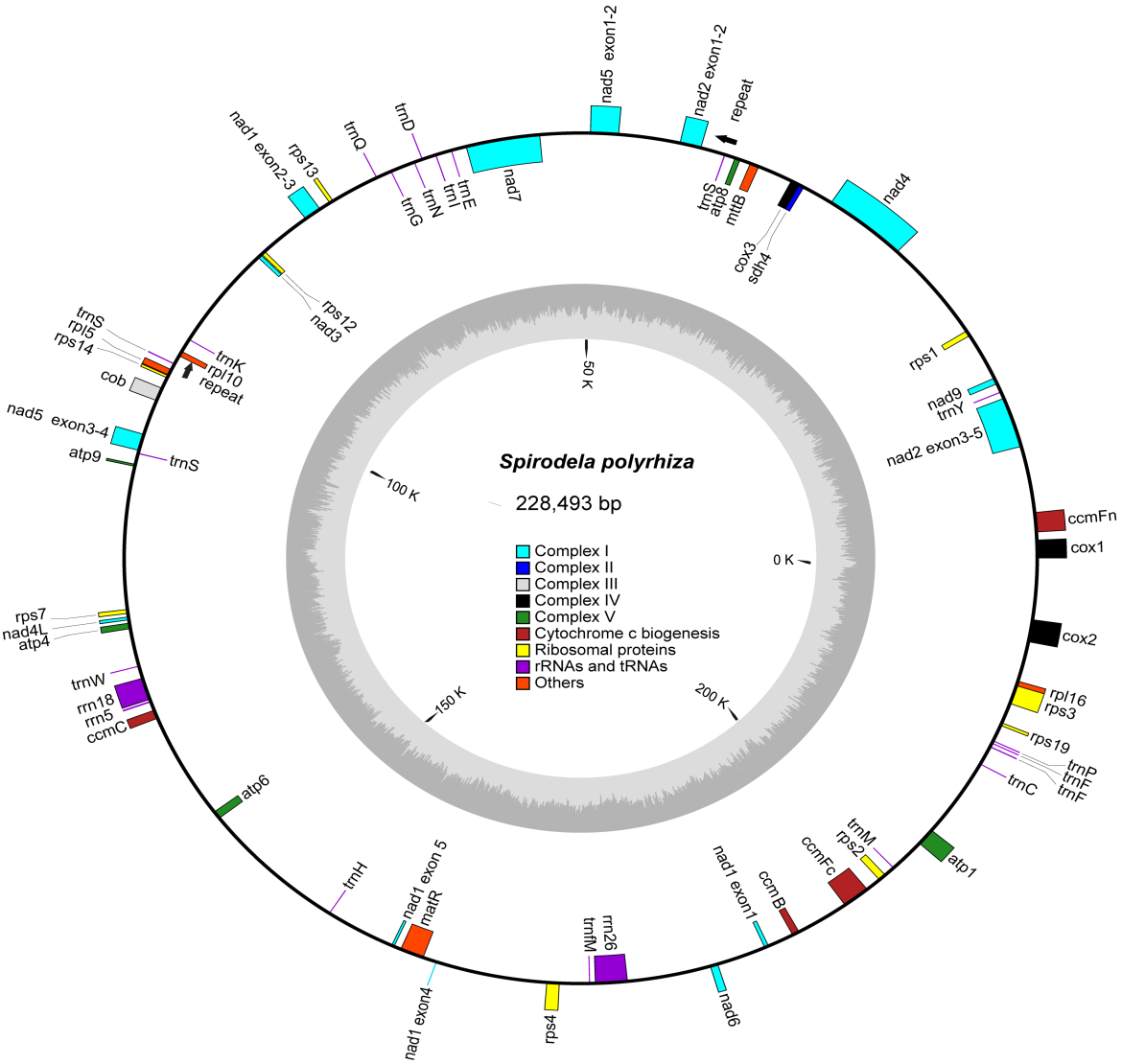


Figure 4.1. The gene map of mitochondrial genome.

Genes indicated as closed boxes on the outside of the circle are transcribed clockwise, whereas those on the inside were transcribed counter-clockwise. Pseudogenes were indicated with the prefix ‘‘Ψ’’. The biggest repeat pair was also marked by arrows. The genome coordinate and GC content are shown in the inner circle.

	Gene	Exon	Start	Stop	Size (bp)	Strand
Complex I	**nad1**				965	
		exon 1	186808	186437	372	-
		exon 2	79193	79273	81	+
		exon 3	80456	80650	195	+
		exon 4	159441	159383	59	+
		exon 5	156079	155822	258	-
	nad2				1431	
		exon 1	47421	47573	153	+
		exon 2	48827	49177	351	+
		exon 3	14038	13883	156	-
		exon 4	11804	11223	582	-
		exon 5	9843	9655	189	-
	nad3		85690	85334	357	-
	nad4				1482	
		exon 1	29243	29701	459	+
		exon 2	31055	31567	513	+
		exon 3	33854	34273	420	+
		exon 4	35940	36029	90	+
	nad4L		119250	119552	303	+
	nad5				1989	
		exon 1	54073	54300	228	+
		exon 2	55155	56369	1215	+
		exon 3	103474	103869	396	+
		exon 4	104795	104944	150	+
	nad6		181813	182472	660	+
	nad7				1128	
		exon 1	66367	66212	156	-
		exon 2	63773	63309	465	-
		exon 3	62308	62063	246	-
		exon 4	60630	60370	261	-
	nad9		15929	15357	573	-
Complex II	**sdh4**		38982	38530	453	-
Complex III	**cob**		99277	100458	1182	+
Complex IV	**cox1**		1	1584	1584	+
	cox2				729	
		exon 1	221286	221675	390	+
		exon 2	222911	223249	339	+
	cox3		39707	38910	798	-
Complex V	**atp1**		201750	203273	1524	+
	atp4		119785	120315	531	+
	atp6		138180	137461	720	-
	atp8		44606	44136	471	-
	atp9		106123	106347	225	+

Cytochrome c	**ccmB**		189517	188906	612	-
	ccmC		127508	128290	783	+
	ccmFn		2225	3940	1716	+
	ccmFc				1317	
		exon 1	196067	195309	759	-
		exon 2	194360	193803	558	-
Ribosomal	**rps1**		20532	20014	519	-
	rps2		197954	197295	660	-
	rps3				1677	
		exon 1	214037	214111	75	+
		exon 2	215854	217455	1602	+
	rps4		168641	169690	1050	+
	rps7		118609	118998	390	+
	rps12		85289	84912	378	-
	rps13		77950	78300	351	+
	Ψrps14		98082	98340	259	+
	Ψrps19		213786	214074	289	+
	rpl5		97522	98076	555	+
	rpl10		96226	95753	474	-
	rpl16		217319	217858	540	+
Other	**matR**		158573	156612	1962	-
	mttB		43302	42532	771	-
tRNA	**trnN-GTT-cp**		70722	70651	72	-
	trnD-GTC		69966	70039	74	+
	trnC-GCA		210090	210160	71	+
	trnQ-TTG		73922	73993	72	+
	trnE-TTC		67604	67533	72	-
	trnG-GCC		72733	72662	72	-
	trnH-GTG-cp		150206	150133	74	-
	ΨtrnH-GTG		191157	191091	67	-
	trnI-CAT		68910	68830	81	-
	trnK-TTT		94640	94568	73	-
	trnM-CAT-cp		198869	198797	73	-
	trnfM-CAT		172112	172039	74	-
	trnF-GAA		212105	212178	74	+
	trnP-TGG		212325	212399	75	+
	trnS-GGA-cp		105233	105147	87	-
	trnS-GCT		211761	211848	88	+
	trnS-TGA		45484	45398	87	-
	trnS-TGA		96955	97041	87	+
	trnW-CCA		123540	123613	74	+
	trnY-GTA		14641	14559	83	-
rRNA	**rrn5**		126685	126803	119	+
	rrn26		175130	172512	2619	-

	Ψrrn26		175418	175140	279	-
	rrn18		124621	126562	1942	+
Putative ORF	**orf115a**		19127	18780	348	-
	orf129		22415	22804	390	+
	orf107a		26004	26327	324	+
	orf107b		26490	26813	324	+
	orf116a		27724	28074	351	+
	orf107c		46987	46664	324	-
	orf100a		56718	57020	303	+
	orf116b		78800	79150	351	+
	orf307		84450	83527	924	-
	orf99a		94364	94065	300	-
	orf107d		103115	103438	324	+
	orf113		107042	107383	342	+
	orf112		110499	110837	339	+
	orf114		111130	111474	345	+
	orf257		111825	112598	774	+
	orf115b		113772	113425	348	-
	orf111		114469	114804	336	+
	orf105		116075	116392	318	+
	orf143a		120699	121130	432	+
	orf100b		121123	120821	303	-
	orf125		128579	128956	378	+
	orf172		132119	132637	519	+
	orf99b		134984	134685	300	-
	orf139		136005	136424	420	+
	orf100c		139752	139450	303	-
	orf126		145095	145475	381	+
	orf121		146128	146493	366	+
	orf117a		147196	147549	354	+
	orf101		153486	153791	306	+
	orf106		154610	154290	321	-
	orf117b		154804	154451	354	-
	orf150		159198	159650	453	+
	orf143b		166461	166030	432	-
	orf161		171133	170648	486	-
	orf120		179396	179758	363	+
	orf99c		193656	193357	300	-
	orf130a		198383	197991	393	-
	orf130b		207211	207603	393	+
	orf102		209373	209681	309	+

Table 4.3 Gene content for mitochondrial genome.
Gene content includes protein-coding genes, tRNA, rRNA and putative ORFs. "Ψ" means pseudo gene and "cp-" means chloroplast-derived gene.

Post-transcriptional editing occurs in nearly all plant mitochondria, which results in altered amino acid sequences of the translated protein by converting specific Cs into Us in their transcripts. We used the program of the predictive RNA editor of plant mitochondrial genomes (PREP-mt) to predict the location of RNA editing sites, which are based on well-known principles that plant organelles maintain the conservation of protein sequences across many species by editing mRNA [21]. By setting the cut-off value to 0.6 within the 35 protein-coding genes of Spirodela mtDNA 600 sites were predicted as C-to-U RNA editing sites (Table 4.4). To validate the accuracy of this prediction, we compared RNA transcripts from *atp9*, *nad9*, *cox3* and *rps12* by RT-PCR with the corresponding genomic sequences yielding a confirmation for 90.8% of the predicted sites. Considering a level of about 10% artificial predictions, we estimate about 540 RNA editing sites, a number that lies between the 441 of protein-coding genes of *Oryza* [15] and 1,084 of *Cycas* [19].

	Gene	**Cut-off**		
		=1	**>=0.8**	**>=0.6**
Complex I	**nad1**	17	4	4
	nad2	22	2	7
	nad3	11	6	1
	nad4	39	9	7
	nad4L	6	4	1
	nad5	13	11	4
	nad6	14	1	3
	nad7	25	7	1
	nad9	6	3	3
Complex II	**sdh4**	0	1	1
Complex III	**cob**	14	2	1
Complex IV	**cox1**	19	0	3
	cox2	13	2	0
	cox3	4	9	3
Complex V	**atp1**	1	5	0
	atp4	6	2	1

	atp6	15	2	3
	atp8	3	0	2
	atp9	7	0	0
Cytochrome c	**ccmB**	17	13	10
	ccmC	19	4	14
	ccmFn	11	4	2
	ccmFc	23	5	7
Ribosomal proteins	**rps1**	1	0	4
	rps2	5	0	2
	rps3	4	3	2
	rps4	8	6	1
	rps7	0	1	1
	rps12	4	5	1
	rps13	2	3	1
	rpl5	3	2	2
	rpl10	2	6	0
	rpl16	3	1	3
Other proteins	**matR**	9	2	7
	mttB	7	10	10
Total		353	488	600

Table 4.4 redicted RNA editing numbers in each protein-coding gene for Spirodela mtDNA.

The cutoff value for each predicted site was the percentage of matches in alignment to the corresponding amino acid across species.

It is generally accepted that RNA editing is essential for functional protein expression as it is required to modify amino acids to maintain appropriate structure and function [22], or to generate new start or stop codons [23]. Indeed, the abundance of RNA editing sites in Spirodela mtDNA might have increased genome complexity and pace of divergence. We summarized the number of potentially modified codons of Spirodela mtDNA in Table 4.5. Three edited codons (TCA (S) => TTA (L); TCT (S) => TTT (F); CCA (P) => CTA (L)) were found most frequently, whereas three editing

events from two codons (CAA (Q) => TAA (X); CAG (Q) => TAG (X)) resulted in stop codons (Table 4.5). Even though three new stop codons are located close at the C-end of proteins (ccmC, rps1 and rpl16), it is not clear whether these small truncations affect their functions or not, which would require experimental evidence.

RNA editing	Cut-off		
	=1	>=0.8	>=0.6
ACA (T) => ATA (I)	1	3	2
ACG (T) => ATG (M)	4	1	0
ACT (T) => ATT (I)	2	2	1
CAA (Q) => TAA (X)	2	0	0
CAC (H) => TAC (Y)	6	1	1
CAG (Q) => TAG (X)	1	0	0
CAT (H) => TAT (Y)	15	5	2
CCA (P) => CTA (L)	30	14	8
CCA (P) => TCA (S)	8	1	2
CCC (P) => CTC (L)	6	4	3
CCC (P) => TCC (S)	8	2	4
CCC (P) => TTC (F)	0	2	4
CCG (P) => CTG (L)	23	6	3
CCG (P) => TCG (S)	4	5	0
CCT (P) => CTT (L)	16	8	6
CCT (P) => TCT (S)	11	6	4
CCT (P) => TTT (F)	4	4	0
CGC (R) => TGC (C)	9	2	2
CGG (R) => TGG (W)	31	2	6
CGT (R) => TGT (C)	17	6	11
CTC (L) => TTC (F)	5	2	5
CTT (L) => TTT (F)	6	4	1
GCA (A) => GTA (V)	0	0	1
GCC (A) => GTC (V)	0	0	1
GCG (A) => GTG (V)	3	0	2
GCT (A) => GTT (V)	2	0	2
TCA (S) => TTA (L)	54	19	13
TCC (S) => TTC (F)	21	7	11
TCG (S) => TTG (L)	25	18	9
TCT (S) => TTT (F)	39	11	8
Total	353	488	600

Table 4.5 Type and number of codon modification in predicted RNA editing sites.

The rRNA and tRNA genes

Spirodela mtDNA contains 3 ribosomal RNA genes (*rrn5*, *rrn18*, *rrn26*) and one pseudogene of *rrn26*. The 19 putatively expressed tRNA genes are specific for 15 amino acids (Table 4.3). Four of them (*trnN*-GTT, *trnH*-GTG, *trnM*-CAT and *trnS*-GGA) are probably chloroplast-derived because of high sequence similarity. They are also predicted as chloroplast origin in maize, rice, sugar beet and *Arabidopsis* except *trnS*-GGA in maize [16]. Therefore, they were not recently acquired from chloroplast, but more likely an event of horizontal transfer in a common ancestor. One *trnH*-GTG is considered to be a non-functional pseudogene. Functional tRNA genes for the amino acids Ala, Arg, Leu, Thr and Val are absent. Because all 20 amino acids are required for protein synthesis, and all 64 codons are used in the Spirodela mt genome based on a codon-usage scan [24], the missing tRNAs are presumably encoded by the nuclear genome and imported from the cytosol into the mitochondria [25] [26].

ORFs and intergenic sequences

Only ORFs encoded by a hypothetical gene with more than 300 bp in length between start and stop codons and no match with a known mt coding sequence were counted. Based on this cut-off, we found 39 mitochondrial ORFs, most of which were not cp migrations and specific to Spirodela (Table 4.3). We named ORFs using their amino acid numbers. When the same length of ORFs happened, a lower case letter (a, b, c, etc) was added. Given the large amount of intergenic DNA in Spirodela mtDNA, it is not surprising to find an abundance of additional ORFs in its genome. Rarely, ORFs showed conservation to any other plants so that putative ORFs were considered to be spurious

prediction [27]. However, orf100a had an ortholog of a NADH-ubiquinone oxidoreductase chain in *Nicotiana tabacum* (GenBank: YP_717128) and orf257 had sequence similarity to DNA polymerase (GenBank: YP_003875487) found in plant mt plasmids [4]. Some studies found that unidentified ORFs had transcripts in rapeseed [28] or to be actively transcribed in sugar beet [29], but further studies are needed to determine whether they encode functional proteins.

A striking feature of Spirodela mtDNA was that 81% of the intergenic regions were species-specific and showed no sequence similarity to any other known sequence. It seemed that anonymous sequences in intergenic DNA were quite common. For instance, unidentifiable sequences comprised 70% of *Beta vulgaris* mtDNA [30]. Although they split about 50 million years ago, 76% of rice mtDNA sequences appeared to be highly divergent from maize in intergenic regions [16]. The repetitive DNAs [31], mt plastidal migrations [32] and viral DNA insertions [33] could contribute to the expansion of intergenic regions, but still comprised a rather small fraction in most seed plant mt genomes. On the other hand, it was quite common that multipartite mt genomes could be generated through large repeat pairs with high frequency [27]. Indeed, 29 potential candidates of repeat pairs with more than 50 bp were found in Spirodela mtDNA by using REPuter [34] (Table 4.6). However, we could not detect repeat-specific contigs from the assembly that could be explained of isomeric and subgenomic molecules derived from a master circle after recombination. Probably, the high rate of non-coding sequence turnover in Spirodela mtDNA was mainly generated through the process of micro-homologous recombination or non-homologous end joining, later on of active rearrangement and continuous reshuffling. Still, the high proportion of enigmatic non-

coding regions in mtDNA is quite extensive. To understand where all these enigmatic sequences might come from and why they appeared to be so common would require additional sequences from closely related species.

Repeat length	Start position	Match direction	Repeat length	Start position	E-value
1535	45070	P	1174	96196	0.00E+01
236	50135	F	236	104843	8.53E-130
137	124339	P	137	175911	1.99E-70
128	13019	F	128	64008	4.87E-65
134	85318	P	134	120516	3.28E-64
129	82501	P	129	222228	2.36E-63
110	13037	F	110	64026	8.71E-57
104	82526	P	104	222228	3.57E-53
113	92268	F	113	199599	8.61E-52
89	107193	F	89	221998	1.18E-37
80	76738	F	80	221186	2.41E-36
79	92302	F	79	199633	1.11E-33
63	71140	F	63	88350	1.73E-28
62	25586	P	62	162923	6.90E-28
62	54713	F	62	214604	6.90E-28
62	119173	P	62	198625	6.90E-28
69	150140	F	69	191090	5.96E-26
57	34898	F	57	223562	7.07E-25
65	115922	P	65	219551	1.27E-23
63	39865	F	63	160103	1.85E-22
52	125889	P	52	173932	7.24E-22
53	81634	P	53	172361	2.88E-20
51	48719	F	51	80393	4.43E-19
54	107228	F	54	222033	5.83E-19
50	77888	F	50	120636	1.28E-16
50	106264	F	50	119017	1.28E-16
50	150160	F	50	191110	1.28E-16
50	124431	P	50	175906	6.13E-15
50	131626	P	50	163414	6.13E-15

Table 4.6 Predicted repeat pairs.

“F” and “P” means forward and palindromic matches.

Phylogenetic analysis and gene loss in angiosperm mitochondrial genomes

After re-examining mitochondrial genome annotations from seven species, a selection of 19 conserved genes (*nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5*, *nad6*, *nad7*, *nad9*, *cob*, *cox1*, *cox2*, *cox3*, *atp1*, *atp4*, *atp6*, *atp8*, *atp9* and *rps3*) was concatenated to permit alignment analysis of 19,824 sites in eight genomes (dicot: *Arabidopsis*, *Nicotiana* and *Boea*; monocot: Spirodela, *Phoenix*, *Oryza* and *Zea*; outgroup: *Cycas*). The gene tree topology from multiple loci (Figure 4.2) was largely congruent with the known phylogenetic relationships inferred from analysis of rbcL. There were two subclades of monocots and dicots within the angiosperm [35]. Previous studies of fossil records [36], morphology and molecular analysis [37] also supported that Alismatales (Spirodela) was a basal monocot followed by Arecales (*Phoenix*), whereas the Poales (rice and maize) resided in the most developed positions.

The loss of protein coding and tRNA genes in seven genomes relative to the outgroup was examined based on the phylogenetic tree. Generally, most losses were limited in their phylogenetic depth to a single family and must have occurred recently (Figure 4.2). Three ribosomal protein genes *rps10*, *rps11* and *rpl2* were missing in Spirodela mtDNA. Frequent gene losses of ribosomal protein genes also occurred in other species. At a closer look, *rps2* seemed to have been lost early in the evolution of dicots, whereas *rps2* was present in *Cycas*, *Marchantia*, and other monocots [38]. The *rps11* gene was missing in dicots (*Arabidopsis*, *Nicotiana* and *Boea*) and also in some monocots (Spirodela, *Oryza* and *Zea*). The corresponding mt *rps2* and *rps11* genes have

been transferred to the nucleus in *Arabidopsis*, soybean, and tomato, suggesting that gene loss followed functional transfer to the nucleus [6, 38]. The unparallel loss of *rps11* and *rpl2* in Spirodela compared with other monocots suggested that the loss of many genes might have occurred independently in various lineages during speciation of angiosperms. The *sdh3* gene was absent and the *sdh4* gene was present in both Spirodela and *Phoenix*, whereas neither was retained in rice and maize (Figure 4.2). A previous study showed that *sdh4* losses were concentrated in the monocots and no losses were detected in basal angiosperms by Southern blot survey of 280 angiosperm genera, which further showed most of the losses were limited in phylogenetic depth to a single family [39].

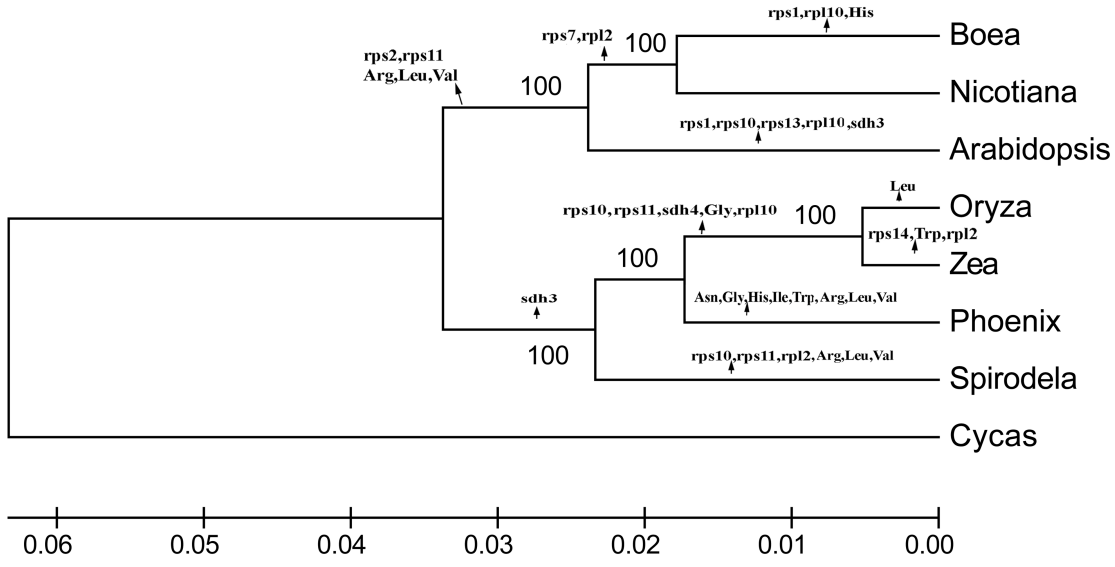


Figure 4.2 Phylogenetic tree based on 19 conserved genes.

The ML calculation was run by MEGA5 with 1,000 bootstrap replicates. All the gene losses were mapped on the tree branches. *Cycas* was included in the analysis as an outgroup. The signs of Amino Acid (Ala, Arg, Leu, Thr, His, Trp, Ile, Gly, Leu and Val) mean corresponding functional tRNA genes were absent in their mtDNAs.

Our data lend support to previous studies that most gene losses occurred with mt ribosomal protein genes and rarely with respiratory genes, which was well documented with a Southern blot survey of gene distribution in 281 diverse angiosperms [6]. When a gene was missing from mtDNA of a given species, it was generally assumed that the original copy had been transferred to the nucleus. Therefore, our results strongly suggested that intracellular gene transfer of ribosomal protein and tRNA genes from mitochondria to the nuclear genome was a frequent process, which in return allowed the nucleus to control the organelle by encoding organelle-destined proteins [25, 26]. Still, functional copies of these putative transferred genes will have to be confirmed after the whole nuclear genome sequence will be available. The finding of many intermediate stages of the *cox2* gene transfer in legumes had shown that physical movement of mtDNA to the nuclear genome was an ongoing process [40].

Chloroplast DNA insertions

The Spirodela mtDNA contained multiple cp-originated insertions, ranging in size from 69 to 1,048 bp. These sequences added up to 9,295 bp of the total amount of transferred cpDNA (Table 4.7), accounting for 4.07% of the mtDNA. A total of 4,436 bp was derived from the inverted repeats of the chloroplast genome, whereas 4,859 bp was transferred from single copy regions of cpDNA. The similarity level of each insertion to the chloroplast genome varied between 75% and 100%. Moreover, the migrated plastid fragments had 732 substitutions, 28 insertions, and 49 deletions within 9,295 bp. They also contained fragments of plastid genes, such as *psbA*, *petB*, *psbC* and *ycf1* (Table 4.7). All of the protein-coding genes of plastid origin in Spirodela mtDNA were likely to be

non-functional as a result of truncations and mutations, whereas four tRNAs of plastidal origin appeared to be intact. Indeed, chloroplast-derived sequences were very common in plant mt genomes, such as 6% in rice [15], 4% in maize [16] and 1% in *Arabidopsis* [18]. Surprisingly, 42.4% of the chloroplast genome of *Vitis* has been incorporated into its mt genome [41]. And a large segment of 113 Kb from chloroplast sequences was captured by the *Cucurbita* mt genome [5].

% identity	cpDNA	cpDNA end	Annotation	mtDNA	mtDNA end
95.92	228	324	trnH-GUG	150132	150228
97.1	237	305	trnH-GUG	191091	191159
90.27	533	1564	psbA	150343	151372
90	4476	4645	trnK-UUU	162465	162632
74.73	37978	38421	psbD	188802	188408
81.09	38572	39513	psbC	188311	187390
86.96	49498	49584	-	105747	105658
83.17	49595	50177	trnS-GCU	105672	105093
92.23	50705	50807	rps4	50125	50023
89.24	54083	54431	ndhJ	226358	226697
75.16	55330	55911	ndhC	16666	16132
94.59	58264	58337	trnM-CAU	198870	198797
85.67	72721	73072	trnP-UGG	123630	123282
81.01	73632	74204	rpl33	123301	122745
93.21	80853	81073	psbN	225145	225359
89.95	81266	82267	petB	225358	226349
90.79	84750	84977	rpoA	90484	90704
93.39	91223	91461	rpl2	150090	149853
100	108036	108139	rrn16	171055	171158
74.86	108322	109185	rrn16	124907	125765
100	110069	110156	ycf68	37345	37432
85.71	112278	112374	-	175279	175183
95.05	116140	116341	trnN-GUU	70647	70848
84.99	118231	118785	-	70899	71440

Table 4.7 The cpDNA-derived regions in Spirodela mtDNA.

Integrated nuclear DNA

It is believed that transposable elements in mitochondria are nuclear-derived and are therefore common in mt intergenic regions [30, 42]. For instance, 4% of *Arabidopsis* mtDNA was probably derived from transposons of nuclear origins [18]. Four fragments of transposable elements were found in maize mtDNA [16] and nineteen were identified in rice [15]. However, we could not find any transposons in the Spirodela mt genome when we searched against the Repbase repetitive element database [43]. This suggests that either very few nuclear sequences have migrated into Spirodela mtDNA or Spirodela mitochondria select against transposable elements.

Comparison of genome synteny

A significant degree of synteny was found within mitochondrial genomes of liverworts, mosses, and chlorophytes at the base of land plants, including a set of gene clusters (more than two genes together), such as the ribosomal protein cluster, *ccm* gene cluster, and two regions containing the *nad* and *cox* genes [44]. It was clear that the sequences of protein-coding genes were highly conserved, but the relative order of genes was greatly rearranged between Spirodela and rice (Figure 4.3). Many ribosomal proteins were independently lost in both Spirodela and rice (Figure 4.2); therefore, synteny between the remaining genes became harder to detect. The ancestral *cob-nad1-cox3-cox2-nad6-atp6-rps7-rps12-nad2-nad4-nad5* gene order of basal land plants has been lost due to various recombination and rearrangement events in angiosperm mtDNA evolution. [4, 33, 45].

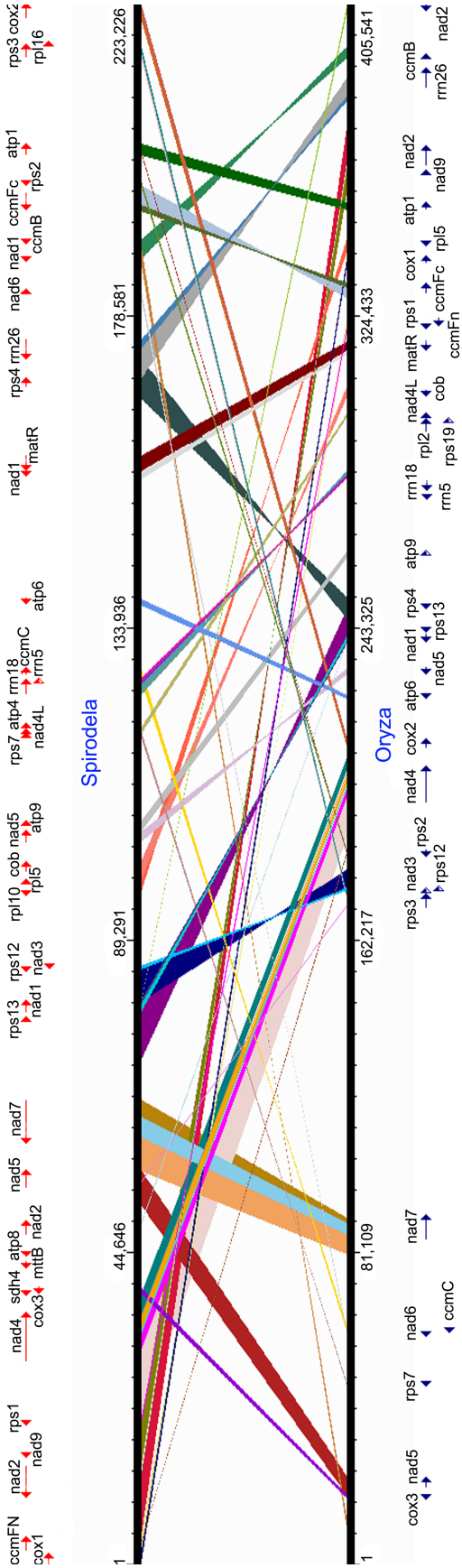

Spirodela
Oryza
1
44,646
89,291
133,936
178,581
223,226
1
81,109
162,217
243,325
324,433
405,541
ccmFN
cox1
nad2
nad9
rps1
cox3
nad5
rps7
nad4
sdh4
mttB
atp8
nad2
nad6
ccmC
nad5
nad7
rps13
nad1
rps12
nad3
rpl10
rpl5
cob
nad5
atp9
rps7
nad4L
atp4
rrn18
rrn5
ccmC
atp6
nad1
matR
rps4
rrn26
nad6
nad1
ccmB
ccmFc
rps2
atp1
rps3
rpl16
cox2
rps3
nad3
rps12
rps2
nad4
atp6
nad5
nad1
rps4
rps13
atp9
rrn18
rrn5
rpl2
rps19
nad4L
cob
matR
ccmFn
rps1
cox1
ccmFc
rpl5
atp1
nad9
nad2
ccmB
rrn26
nad2

Figure 4.3 Comparison of synteny in Spirodela and *Oryza*.

The annotated protein-coding genes were indicated for Spirodela and *Oryza*. Major conserved regions were bridged by lines. The visualized genome synteny was performed by GSV: a web-based genome synteny viewer [46].

In summary, our data provides further evidence that SOLiD platforms can assemble both chloroplast and mitochondrial genomes with regular coverage without any organellar purification (Table 4.1) [11]. Our analysis of the mt genome of Spirodela, having the smallest size among sequenced monocots, elucidates the evolutionary change among monocot mt genomes. Although the critical genes for the electron transport chain in Spirodela mtDNA are well conserved, different types of ribosomal protein genes are missing in comparison to other monocots. The number of RNA editing in protein coding genes is within a typical range as other plants. Still, no known transposable elements can be found in its genome, suggesting a rather rare migration from the nucleus to the mitochondria. Sequence-based phylogenetic analysis clearly supports the hypothesis that Spirodela is at the very basal lineage of monocots. Comparative analyses of mitochondrial genes between Spirodela and rice have shown that the relative order of genes is greatly rearranged over a very short evolutionary time. In this regard, additional complete mitochondrial sequences from closely related species will be needed to fortify the distinct evolution of plant mitochondrial genomes.

4.4 Materials and methods

DNA Isolation and SOLiD DNA sequencing

The methods for DNA extraction and DNA sequencing by the SOLiD platform followed a protocol as previously published [11]. Briefly, total genomic DNA was extracted from the clonally grown whole plant tissue of *Spirodela polyrhiza*. A mate-paired library was made with 1.5 Kb insertions and read length was 50 bp. Since nucleic, mitochondrial and chloroplast sequence all exist in reads from total DNA preparation, copy number between three genomes was significantly different [13, 14], so that it was feasible to *de novo* assembly both chloroplast and mitochondria genomes using the same dataset but with different coverage cut-off numbers as described previously [11].

Genome assembly, finishing and validation

The coverage cut-off was fully utilized to only allow the target organellar genome to be assembled due to obvious differentiation of copy number for three genomes in total reads [13]. Furthermore, low-level contaminating sequences from foreign DNA (mainly nuclear DNA) were discarded by this approach. Quality control and other details were described recently [11]. Before we assembled the mitochondria genome using mate-paired reads, we masked chloroplast reads to reduce effects due to plastid sequence predominance. The detailed pipeline was shown below (Figure 4.4).

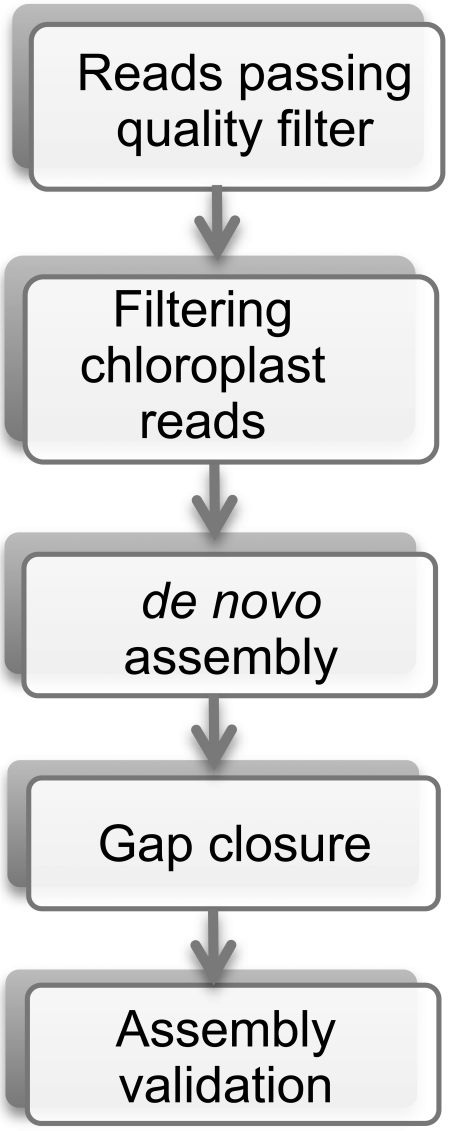


Figure 4.4 Pipeline of mitochondrial genome assembly.

1) Filtering chloroplast reads: we mapped total high quality reads to existing chloroplast genome (GenBank # JN160603) by BWA short-read alignment component with default parameters [47]. Only unmapped reads were used in the next step. 2) *de novo* assembly: the assembly was executed using the SOLiD^{TM} System *de novo* Accessory Tools 2.0 (http://solidsoftwaretools.com/gf/project/denovo/) in conjunction with the

Velvet assembly engine [48]. 3) Gap closure: since chloroplast reads were pre-removed before mitochondrial assembly, theoretically, any location with chloroplast insertion in mtDNA would create a gap. Using flanking primers bridging 57 gaps, the missing sequences were amplified and sequenced with the ABI 3730xl system, yielding a complete contiguous mtDNA sequence. To validate the circularity of the Spirodela mtDNA, PCR products were sequenced with pairs of primers bridging gaps and overlapping with the assembled linear scaffold. 4) Most gaps were small enough for single CE (capillary electrophoresis) sequence reads and overlapping sequences served as a measure for the accuracy of the SOLiD assembly and error rate. Therefore, PCR amplification and CE sequence provided validation of the order of contigs and also revealed sequencing discrepancies between these two platforms.

Genome annotation and sequence analysis

The main pipeline for mitochondrial genome annotation was adapted from other sources [5]. Databases for protein-coding genes, rRNA and tRNA genes were compiled from all previously sequenced seed plant mitochondrial genomes. BLASTX and tRNAscan-SE were the mainly used programs [5]. The boundaries for each gene were manually curated. The sequin file including sequence and annotation was submitted to NCBI GenBank as JQ804980. The graphical gene map was processed by OrganellarGenomeDRAW program [49]. The codon usages for all protein coding genes in Spirodela and *Oryza* were calculated by using the Sequence Manipulation Suite [24].

Cp-derived tRNAs were identified by aligning all tRNA in annotated cpDNA to mtDNA with 80% of identity, an e-value of 1e-10 and a 50% coverage threshold. All

remaining sequences were further scanned by EMBOSS getorf for open reading frames (ORFs) with more than 300 bp [50].

Putative RNA editing sites in protein-coding genes were identified by the PREP-mt Web-based program based on the evolutionary principle that editing increases protein conservation among species (http://prep.unl.edu/) [21]. The optimized cut-off value 0.6 was set in order to achieve the maximal accurate prediction. RNA editing sites from four genes were validated by RT-PCR with gene-specific primers.

Sequences transferred to mtDNA were found by BLASTN search of mtDNA against the Spirodela chloroplast genome with 80% of identity, e-value of 1e-10 and 50 bp of length threshold. Repeat sequence analysis was predicted by using REPuter web-based interface, including forward, palindromic, reverse and complemented repeats with a cut-off value of 50 bp [34]. The mitochondrial genome was screened by repeatmasker under cross_match search engine (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker) for interspersed repeats and low complexity DNA sequences [51].

Phylogenetic analysis

We aligned 19 homologous protein-coding gene sequences (*nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5*, *nad6*, *nad7*, *nad9*, *cob*, *cox1*, *cox2*, *cox3*, *atp1*, *atp4*, *atp6*, *atp8*, *atp9* and *rps3*) from the Spirodela mitochondrial genome and other seven plant organisms (*Cycad*, NC_010303; *Phoenix*, NC_016740; Spirodela, JQ804980; *Oryza*, NC_011033; *Zea*, NC_007982; *Boea*, NC_016741; *Nicotiana*, NC_006581; *Arabidopsis*, NC_001284)

and constructed a phylogenetic tree. Annotations were revalidated and sequences were concatenated into a single continuous sequence from 18,537 to 19,041 bp to initiate alignment by MEGA5 [52]. The phylogeny of the mitochondrial genome was estimated by maximum likelihood (ML) with 1,000 Bootstrap of replicates. *Cycas* was used as the outgroup.

Comparison of global genome structure

The conserved regions for protein-coding and rRNA genes were identified between Spirodela and *Oryza* sequences by BLASTN. The synteny together with the annotation file were uploaded to a web-based genome synteny viewer GSV [46].

Acknowledgement

This work has been published in PLoS ONE of 2012 by authors of Wang W, Wu Y, Messing J with the title of "The mitochondrial genome of an aquatic plant, *Spirodela polyrhiza*".

4.5 References

1. Logan DC: **The mitochondrial compartment**. *J Exp Bot* 2006, **57**(6):1225-1243.
2. Mackenzie S, McIntosh L: **Higher plant mitochondria**. *Plant Cell* 1999, **11**(4):571-586.
3. Keeling PJ, Palmer JD: **Horizontal gene transfer in eukaryotic evolution**. *Nat Rev Genet* 2008, **9**(8):605-618.
4. Sloan DB, Alverson AJ, Storchova H, Palmer JD, Taylor DR: **Extensive loss of translational genes in the structurally dynamic mitochondrial genome of the angiosperm Silene latifolia**. *BMC Evol Biol* 2010, **10**:274.
5. Alverson AJ, Wei X, Rice DW, Stern DB, Barry K, Palmer JD: **Insights into the evolution of mitochondrial genome size from complete sequences of Citrullus lanatus and Cucurbita pepo (Cucurbitaceae)**. *Mol Biol Evol* 2010, **27**(6):1436-1448.
6. Palmer JD, Adams KL, Cho Y, Parkinson CL, Qiu YL, Song K: **Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates**. *Proc Natl Acad Sci U S A* 2000, **97**(13):6960-6966.
7. Fang Y, Wu H, Yang M: **A Complete Sequence and Transcriptomic Analysis of Date Palm (Phoenix dactylifera L.) Mitochondrial Genome**. *NCBI GenBank NC_0167401* 2012.
8. Cabrera LI, Salazar GA, Chase MW, Mayo SJ, Bogner J, Davila P: **Phylogenetic relationships of aroids and duckweeds (Araceae) inferred from coding and noncoding plastid DNA**. *Am J Bot* 2008, **95**(9):1153-1165.
9. Stomp A-M, El-Gewely MR: **The duckweeds: A valuable plant for biomanufacturing**. In: *Biotechnology Annual Review*. vol. Volume 11: Elsevier; 2005: 69-99.
10. Ljaz S: **Plant mitochondrial genome: "A sweet and safe home'' for transgene**. *African Journal of Biotechnology* 2010, **9**(54):4.
11. Wang W, Messing J: **High-Throughput Sequencing of Three Lemnoideae (Duckweeds) Chloroplast Genomes from Total DNA**. *PLoS ONE* 2011, **6**(9):e24670.
12. Les DH, Crawford DJ, Landolt E, Gabel JD, Kimball RT: **Phylogeny and Systematics of Lemnaceae, the Duckweed Family**. *Systematic Botany* 2002, **27**(2):221-240.
13. Zhang T, Zhang X, Hu S, Yu J: **An efficient procedure for plant organellar genome assembly, based on whole genome data from the 454 GS FLX sequencing platform**. *Plant Methods* 2011, **7**:38.
14. Lutz K, Wang W, Zdepski A, Michael T: **Isolation and analysis of high quality nuclear DNA with reduced organellar DNA for plant genome sequencing and resequencing**. *BMC Biotechnology* 2011, **11**(1):54.

15. Notsu Y, Masood S, Nishikawa T, Kubo N, Akiduki G, Nakazono M, Hirai A, Kadowaki K: **The complete sequence of the rice (Oryza sativa L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants**. *Mol Genet Genomics* 2002, **268**(4):434-445.
16. Clifton SW, Minx P, Fauron CM, Gibson M, Allen JO, Sun H, Thompson M, Barbazuk WB, Kanuganti S, Tayloe C *et al*: **Sequence and comparative analysis of the maize NB mitochondrial genome**. *Plant Physiol* 2004, **136**(3):3486-3503.
17. Malek O, Knoop V: **Trans-splicing group II introns in plant mitochondria: the complete set of cis-arranged homologs in ferns, fern allies, and a hornwort**. *RNA* 1998, **4**(12):1599-1609.
18. Unseld M, Marienfeld JR, Brandt P, Brennicke A: **The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366,924 nucleotides**. *Nat Genet* 1997, **15**(1):57-61.
19. Chaw SM, Shih AC, Wang D, Wu YW, Liu SM, Chou TY: **The mitochondrial genome of the gymnosperm Cycas taitungensis contains a novel family of short interspersed elements, Bpu sequences, and abundant RNA editing sites**. *Mol Biol Evol* 2008, **25**(3):603-615.
20. Zhang T, Fang Y, Wang X, Deng X, Zhang X, Hu S, Yu J: **The complete chloroplast and mitochondrial genome sequences of Boea hygrometrica: insights into the evolution of plant organellar genomes**. *PLoS One* 2012, **7**(1):e30531.
21. Mower JP: **The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments**. *Nucleic Acids Res* 2009, **37**(suppl 2):W253-259.
22. Giege P, Brennicke A: **RNA editing in Arabidopsis mitochondria effects 441 C to U changes in ORFs**. *Proc Natl Acad Sci* 1999, **96**(26):15324-15329.
23. Takenaka M, Verbitskiy D, van der Merwe JA, Zehrmann A, Brennicke A: **The process of RNA editing in plant mitochondria**. *Mitochondrion* 2008, **8**(1):35-46.
24. Stothard P: **The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences**. *Biotechniques* 2000, **28**(6):1102, 1104.
25. Woodson JD, Chory J: **Coordination of gene expression between organellar and nuclear genomes**. *Nat Rev Genet* 2008, **9**(5):383-395.
26. Schneider A: **Mitochondrial tRNA import and its consequences for mitochondrial translation**. *Annu Rev Biochem* 2011, **80**:1033-1053.
27. Mower JP, Sloan DB, Alverson AJ: **Plant Mitochondrial Genome Diversity: The Genomics Revolution.** In: *Plant Genome Diversity.* Edited by Wendel JF, Greilhuber J, Dolezel J, Leitch IJ, vol. Volume 1: Springer Vienna; 2012: 123-144.
28. Handa H: **The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (Brassica napus L.): comparative analysis of the mitochondrial genomes of rapeseed and Arabidopsis thaliana**. *Nucleic Acids Res* 2003, **31**(20):5907-5916.
29. Satoh M, Kubo T, Nishizawa S, Estiati A, Itchoda N, Mikami T: **The cytoplasmic male-sterile type and normal type mitochondrial genomes of**

sugar beet share the same complement of genes of known function but differ in the content of expressed ORFs. *Mol Genet Genomics* 2004, **272**(3):247-256.

30. Satoh M, Kubo T, Mikami T: **The Owen mitochondrial genome in sugar beet (Beta vulgaris L.): possible mechanisms of extensive rearrangements and the origin of the mitotype-unique regions**. *Theor Appl Genet* 2006, **113**(3):477-484.
31. Lilly JW, Havey MJ: **Small, repetitive DNAs contribute significantly to the expanded mitochondrial genome of cucumber**. *Genetics* 2001, **159**(1):317-328.
32. McDermott P, Connolly V, Kavanagh TA: **The mitochondrial genome of a cytoplasmic male sterile line of perennial ryegrass (Lolium perenne L.) contains an integrated linear plasmid-like element**. *Theor Appl Genet* 2008, **117**(3):459-470.
33. Alverson AJ, Zhuo S, Rice DW, Sloan DB, Palmer JD: **The mitochondrial genome of the legume Vigna radiata and the analysis of recombination across short mitochondrial repeats**. *PLoS One* 2011, **6**(1):e16404.
34. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R: **REPuter: the manifold applications of repeat analysis on a genomic scale**. *Nucleic Acids Res* 2001, **29**(22):4633-4642.
35. Janssen T, Bremer K: **The age of major monocot groups inferred from 800+ rbcL sequences**. *Botanical Journal of the Linnean Society* 2004(146):4.
36. Stockey RA: **The fossil record of basal monocots**. 2006, **22**:16.
37. Borsch T, Hilu KW, Quandt D, Wilde V, Neinhuis C, Barthlott W: **Noncoding plastid trnT-trnF sequences reveal a well resolved phylogeny of basal angiosperms**. *J Evol Biol* 2003, **16**(4):558-576.
38. Perrotta G, Grienenberger JM, Gualberto JM: **Plant mitochondrial rps2 genes code for proteins with a C-terminal extension that is processed**. *Plant Mol Biol* 2002, **50**(3):523-533.
39. Adams KL, Rosenblueth M, Qiu YL, Palmer JD: **Multiple losses and transfers to the nucleus of two mitochondrial succinate dehydrogenase genes during angiosperm evolution**. *Genetics* 2001, **158**(3):1289-1300.
40. Adams KL, Song K, Roessler PG, Nugent JM, Doyle JL, Doyle JJ, Palmer JD: **Intracellular gene transfer in action: Dual transcription and multiple silencings of nuclear and mitochondrial cox2 genes in legumes**. *Proc Natl Acad Sci* 1999, **96**(24):13863-13868.
41. Goremykin VV, Salamini F, Velasco R, Viola R: **Mitochondrial DNA of Vitis vinifera and the issue of rampant horizontal gene transfer**. *Mol Biol Evol* 2009, **26**(1):99-110.
42. Knoop V, Unseld M, Marienfeld J, Brandt P, Sunkel S, Ullrich H, Brennicke A: **copia-, gypsy- and LINE-like retrotransposon fragments in the mitochondrial genome of Arabidopsis thaliana**. *Genetics* 1996, **142**(2):579-585.
43. **RepeatMasker Open-3.0. http://www.repeatmasker.org/**
44. Terasawa K, Odahara M, Kabeya Y, Kikugawa T, Sekine Y, Fujiwara M, Sato N: **The mitochondrial genome of the moss Physcomitrella patens sheds new light on mitochondrial evolution in land plants**. *Mol Biol Evol* 2007, **24**(3):699-709.

45. Chang S, Yang T, Du T, Huang Y, Chen J, Yan J, He J, Guan R: **Mitochondrial genome sequencing helps show the evolutionary mechanism of mitochondrial genome formation in Brassica**. *BMC Genomics* 2011, **12**:497.
46. Revanna KV, Chiu CC, Bierschank E, Dong Q: **GSV: a web-based genome synteny viewer for customized data**. *BMC Bioinformatics* 2011, **12**:316.
47. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform**. *Bioinformatics* 2009, **25**(14):1754-1760.
48. Zerbino DR, Birney E: **Velvet: Algorithms for de novo short read assembly using de Bruijn graphs**. *Genome Res* 2008, **18**(5):821-829.
49. Lohse M, Drechsel O, Bock R: **OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes**. *Curr Genet* 2007, **52**(5-6):267-274.
50. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends Genet* 2000, **16**(6):276-277.
51. Bergman CM, Quesneville H: **Discovering and detecting transposable elements in genome sequences**. *Brief Bioinform* 2007, **8**(6):382-392.
52. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0**. *Mol Biol Evol* 2007, **24**(8):1596-1599.

CHAPTER 5 NUCLEAR GENOMICS

5.1 Abstract

The subfamily of the *Lemnoideae* belongs to a different order than other monocotyledonous species that have been sequenced and comprises aquatic plants that grow rapidly on the water surface. Here, we selected *Spirodela polyrhiza* for whole-genome sequencing. We show that Spirodela has a primordial genome with no signs of recent retrotranspositions but signatures of two ancient whole-genome duplications, possibly 95 million years ago (Mya), older than those in Arabidopsis and rice. Its genome has only 19,623 predicted protein-coding genes, 28% less than the dicotyledonous *Arabidopsis thaliana* and 50% less than monocotyledonous rice. We propose that at least in part the neotenous reduction of these aquatic plants is based on readjusted copy numbers of promoters and repressors of the juvenile-to-adult transition. The Spirodela genome along with its unique biology and physiology will stimulate new insights into environmental adaptation, ecology, evolution, plant development, and will be instrumental for future bioenergy applications.

5.2 Introduction

The *Lemnoideae*, commonly known as duckweeds, are the smallest, fastest growing, and morphologically simplest of flowering plants[1]. The plant body is organized as a thalloid or ‘frond’ lacking a stem and in more derived species even roots (Figure 5.1). Based on fossil records, the peculiar plant body architecture of this

subfamily evolved by neotenous reduction from an *Araceae* ancestor and it has been interpreted botanically as juvenile or embryonic tissue[2]. The reduction and simplification of the plant body progresses within the *Lemnoideae* from ancient species like Spirodela towards more derived species like Wolffia. Although reduced flowers are observed in duckweeds, they usually reproduce by vegetative daughter fronds initiated from the mother frond (Supplementary Fig. S1). Doubling time of the fastest growing duckweeds under optimal growth conditions is less than 30 hours, nearly twice as fast as other "fast-growing" flowering plants and more than double that of conventional crops. They are easy to grow and have negligible lignin and high energy content in the form of easily fermentable starch (40 to 70% of biomass). Duckweeds have been used for removal of high levels of contaminants from wastewater[3], production of recombinant proteins for pharmaceutical applications[4, 5], and high impact biofuel feedstock that does not compete for land in food production[6]. From a taxonomic point of view, genomic efforts have largely focused on the taxa of the *Commelinid* monocots such as the grasses from *Poales* and *Musa acuminata*, the wild type diploid progenitor of banana, from *Zingiberales* (Figure 5.1). Here we describe the genome and transcriptome of Greater Duckweed, *Spirodela polyrhiza*, representing the smallest monocot genome to date with a size of 158 Mb, which is similar to the plant model genome of *Arabidopsis thaliana*. Spirodela represents a basal monocotyledonous species from the *Alismatales* and will be an invaluable genomic resource to study the history of the monocotyledonous lineage.

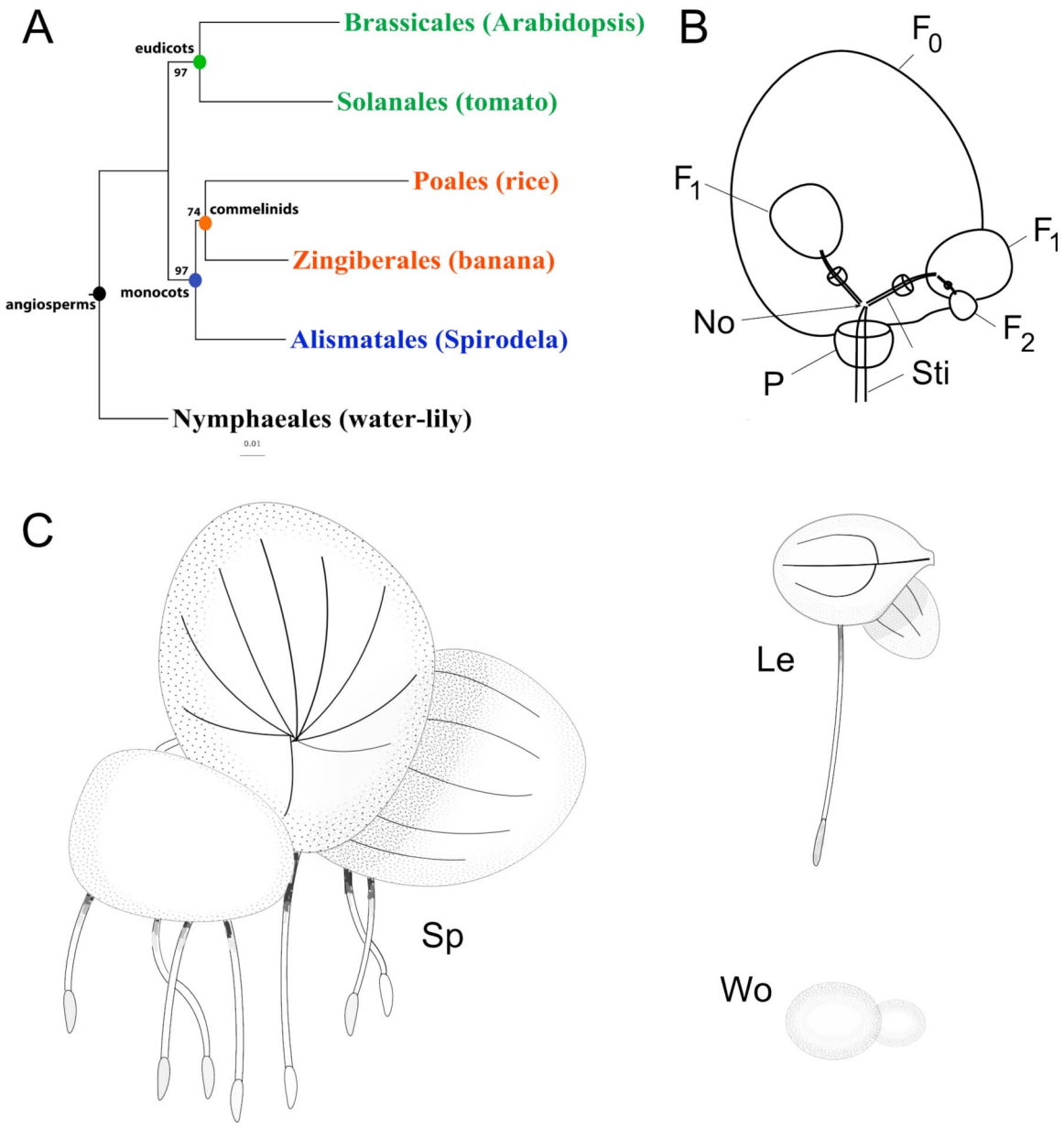


Figure 5.1 Systematics and Biology of the *Lemnoideae*.

(A) shows a phylogenetic tree of plastid-rbcL genes of two dicots - Arabidopsis thaliana (Brassicales, NC_000932.1) and tomato (Solanales, NC_007898.2) -, three monocots *Spirodela polyrhiza* (Alismatales, NC_015891.1); rice (Poales, NC_001320.1); banana (Zingiberales, EU017045.1); and water-lily as outgroup (Nuphar advena, Nymphaeales, , NC_008788). (B) shows a ventral view of Spirodela, illustrating schematically the clonal, vegetative propagation of duckweeds. Daughter fronds (F1) originate from the vegetative node (No) from the mother frond F0 and remain attached to it by the stipule (Sti), which eventually breaks-off,

thereby releasing a new plant cluster. Daughter fronds may already initiate new fronds (F2) themselves before full maturity. Roots are attached at the prophyllum (P). (C) illustrates the progressing reduction from a leaf-like body with several veins and unbranched roots to a thallus-like morphology in the *Lemnoideae*. Sp: *Spirodela polyrhiza*, Le: *Lemna minor*, Wo: *Wolffia arrhiza*.

5.3 Results

Sequence assembly

Genome sizes in the five *Lemnoideae* genera span an order of magnitude from *Spirodela polyrhiza* at 158 Mb to *Wolffia arrhiza* at 1,881 Mb[7]. Due to its small size and basal position in the *Lemnoideae* we sequenced the *Spirodela polyrhiza* accession 7498 by whole-genome shotgun sequencing[8] using ~20X single end, ~1X pair-end Roche/454 next-generation sequencing and ~1X pair-end Sanger sequencing (Supplementary Table S1). Although next-generation sequencing has been used to reduce the cost of sequencing genomes, short-read technologies have been insufficient to assemble chromosome-size molecules with megabase (Mb) genomes, especially for Spirodela, which does not have synteny with other fully sequenced genomes[9]. Therefore, the read-length threshold of Roche/454 with a depth of 21X in combination with long paired-reads from BACs and fosmids appeared to be a significant improvement in the balance of cost and genome sequence quality for new evolutionary references. Of the 158 Mb genome, as measured by flow cytometry (Supplementary Fig. S2), 90% was assembled into contigs, 97% of the contigs assembled in 252 scaffolds, and 94.1% of them in the top 50 largest scaffolds (Supplementary Table S2). Scaffolds were joined into 32 pseudomolecules, using a DNA fingerprinted physical map with anchored sequenced

tagged sites from BAC ends, and one pseudomolecule labeled “0” with all unanchored scaffolds (Figure 5.2, Supplementary Table S3). 10.7% of the genome remained as Ns due to repetitive elements. Based on cytogenetic data, the 32 pseudomolecules corresponded to 20 chromosomes. 0.15% and 0.13% of the nuclear genome are derived from chloroplast and mitochondrial sequence[10, 11] (Supplementary Table S4 and S5). Whereas centromeric repeats have greater species specificity[12], telomere repeats are conserved throughout most flowering plants. In total, we found 20 telomeric repeat fragments in Spirodela ranging from 32 to 258 bp with a signature sequence [$(TTTAGGG)_n$] identical to the telomeres of Arabidopsis[13]. In all cases, clusters of telomeric repeats were found at the ends of pseudomolecules, with 16 at one end and 2 at both ends (Supplementary Fig. S3). These results probably reflect intact chromosomes and confirming the quality of the sequence assembly. We constructed a Spirodela genomic library of bacterial artificial chromosomes (BACs) that was subjected to DNA fingerprinting[14] (Supplementary information), resulting in a physical map. Because the BACs were sequenced at the ends, these ends were used to align the assembled sequence with the physical map, providing us with another proof of an accurate assembly of DNA sequences. We also used BACs that were aligned to the assembly with their sequenced ends to derive their entire sequence from the assembled sequence and used this information to select those that were low in repeat sequences for fluorescence in situ hybridization (FISH) (Supplementary Fig. S4). Cytogenetics could therefore provide us with a quality control for the assembled sequences. Overall the assembly completeness could be verified by three different sized sets of random sampled 454 reads with at least 90% for genic sequences and >=80% for the rest (Supplementary Fig. S5). To further

ensure the quality of sequence assembly, we randomly selected 24 fosmids for conventional sequencing that were then aligned with the assembled 454 sequences and found that the sequencing error rates were 8 in 10,000, providing a 98.22% accuracy (Supplementary Table S6).

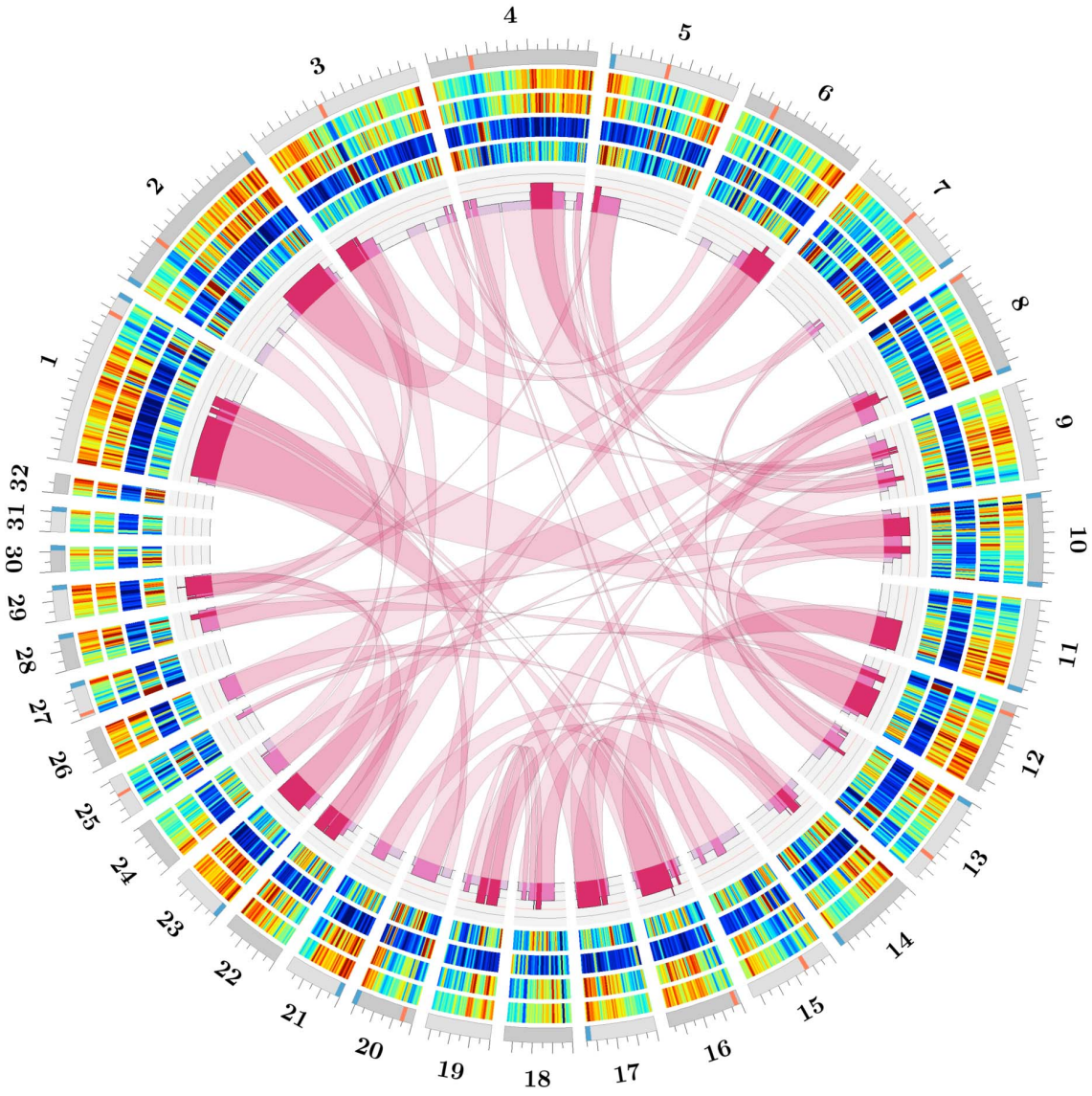


Figure 5.2 Characteristics of the Spirodela genome.

The outer circle shows the 32 pseudomolecules of the Spirodela genome assembly, tick scaling is 500 Kb and blue and red bars depict position of telomeric and centromeric

clusters. Heatmap tracks illustrate from outer to inner circle GC content, gene, repeat and GAGA-repeat densities. Color map ranges are 30-50%, 0-30%, 0-70% and 0-1.5%, respectively. GC and gene content are positively, repeat and gene densities negatively correlated; while GAGA-repeats are present both in gene- and repeat-rich regions. The genome contains two rounds of ancient genome duplications. For each genomic segment, the copy number of paralogous regions is shown as bar chart in the innermost circle, duplication history is illustrated by red ribbons.

Gene number and transcriptome

Once we assembled high-quality 32 chromosome-size pseudomolecules, we were able to determine gene content and order. We identified 19,623 protein-coding genes in Spirodela, making it the fewest predicted genes of any sequenced plants to date with 28% less than Arabidopsis (27,416), and 50% less than rice (39,049) (Supplementary Table S7 and S8). The gene models were supported by 379,502 EST sequences assembled from transcriptome libraries generated from various diurnal time course and stress conditions (Supplementary Table S9). 95.7% of the sequenced transcriptome was mapped to the pseudomolecules, consistent with a complete genome assembly (Supplementary Table S10) Spirodela appeared to have a significantly lower number of tandem gene clusters (948) than rice (2,602), tomato (2,340), and Arabidopsis (1,938), however, surprisingly close to banana (1,048), which had ~1.9 times the gene number of Spirodela (Supplementary Table S11). A total of 413 miRNA loci comprising 93 families were identified in the Spirodela genome assembly by sequence similarity and structural features (Supplementary Table S12). Other small RNAs are described in the Supplementary Information.

Nucleotide composition and repeat elements

In comparison to dicots, increased GC contents were observed in monocotyledonous coding sequences, mainly due to a bias of G and C in the third codon position (Supplementary Fig. S6). Spirodela protein-coding sequences exhibited a pronounced GC3 bias (Supplementary Fig. S7 and S8), which was the highest amongst currently sequenced monocotyledonous genomes. Elevated GC3 contents were found in Spirodela-specific genes as well as genes shared with monocots and dicots and thus seemed to be a general feature of Spirodela-coding sequences (Supplementary Fig. S9). In contrast to the reported distinct bimodal distributions in rice and maize, previously reported GC3 compositions of date palm genes showed a sharp unimodal distribution. Broader distributions resulting from a composite of genes with low and high GC3 content were observed both for Musa and Spirodela genes (Supplementary Fig. S8), suggesting that the distinct multimodal patterns evolved specifically in the grass family and high GC3 biases might have evolved independently and several times in the monocotyledonous lineage.

Spirodela repetitive DNA content is in line with other small sequenced plant genomes at around 13% and significantly less than other sequenced monocotyledonous genomes (Supplementary Fig. S10 and S11, Supplementary Table S13). Spirodela had an exceptional high proportion of microsatellite tandem repeats, 50% versus 3 to 6% in other four reference genomes of Sorghum, rice, Brachypodium and Arabidopsis (Supplementary Fig. S12, S13 and S14). The heat map of the 32 Spirodela pseudomolecules followed the known pattern of anti-correlation between gene and LTR-retrotransposon densities (Supplementary Fig. S15, S16 and S17). In Spirodela, the insertion age distributions of the full-length LTR-retrotransposons revealed distinctly

different patterns with a strong shift towards older elements (Supplementary Fig. S18). The small genome size and atypical LTR age distribution of Spirodela suggested a tight control of transposon activity during recent evolutionary times. Both features might well be connected to the continuous clonal propagation of Spirodela. Transposon transcription is usually activated during seed development and full-length LTR-retrotransposon elements are often removed by meiotic unequal crossing over between solo-LTRs[15], two processes that are limited by the propensity of Spirodela to an asexual lifestyle. Also, in small genomes undergoing genome size reduction, transposition potentially could negatively impact gene activity, possibly requiring tighter regulation[16].

Genome evolution

We identified two consecutive rounds of large-scale or whole-genome duplications (WGDs) in the Spirodela genome based on intra-species comparison of paralogous genes (Figure 5.2, Supplementary Fig. S19 and S20). In contrast to previously sequenced higher plant genomes, Spirodela does not possess recent genome duplications and Ks distributions of syntenic clusters indicated almost concurrent occurrences of both WGDs (Supplementary Fig. S21). Previous studies had shown both strong positive and negative correlations between the estimation of synonymous substitution rates, the applied methodology and GC3 content of paralogs[17]. Likewise, the high GC3 bias in Spirodela resulted in strong deviations of Ks estimates for gene pairs with high (GC3 $\geq$ 75%) GC3 content (Supplementary Fig. S21). Restricting our analysis to low GC3 paralogous pairs, both WGDs dated back around 95 Mya and Spirodela and rice diverged at approximately 130 Mya (Supplementary Fig. S22, S23 and S24). While such numbers cannot be absolute dates, they are meant to illustrate distances between these genomes in

evolutionary terms. Consistent with the older age of both WGDs, syntenic conservation between blocks was sparse (Supplementary Table S14), a minimum of 6% and a maximum of 26% of the genes (mean 11.3%) between two blocks retained a conserved order in global block alignments.

Gene families

The Spirodela genome contained very similar patterns of orthologous gene sets in comparison to four representative species (Arabidopsis, tomato, banana, and rice), sharing a total of 8,255 common gene families despite a significantly reduced gene number (Figure 5.3 and Supplementary Fig. S25). However, Spirodela clusters generally showed the lowest average gene expansion and copy number, indicating preferred gene losses of duplicated genes in Spirodela or - vice versa - gene retentions in the other species (Supplementary Table S15). A notable exception from the overall conserved gene content was 750 families present in all four analyzed species except Spirodela. These families included genes involved in water transport by aquaporins, phenylpropanoid, lignin biosynthesis, and cell wall organization by expansins (Figure 5.3 and Table S16). The loss of these gene families is consistent with the specialized morphology and lifestyle of Spirodela. Overrepresented functional categories of Spirodela-specific genes were enriched for various defense related processes including antimicrobial peptides and adapted immune responses (Supplementary Table S17).

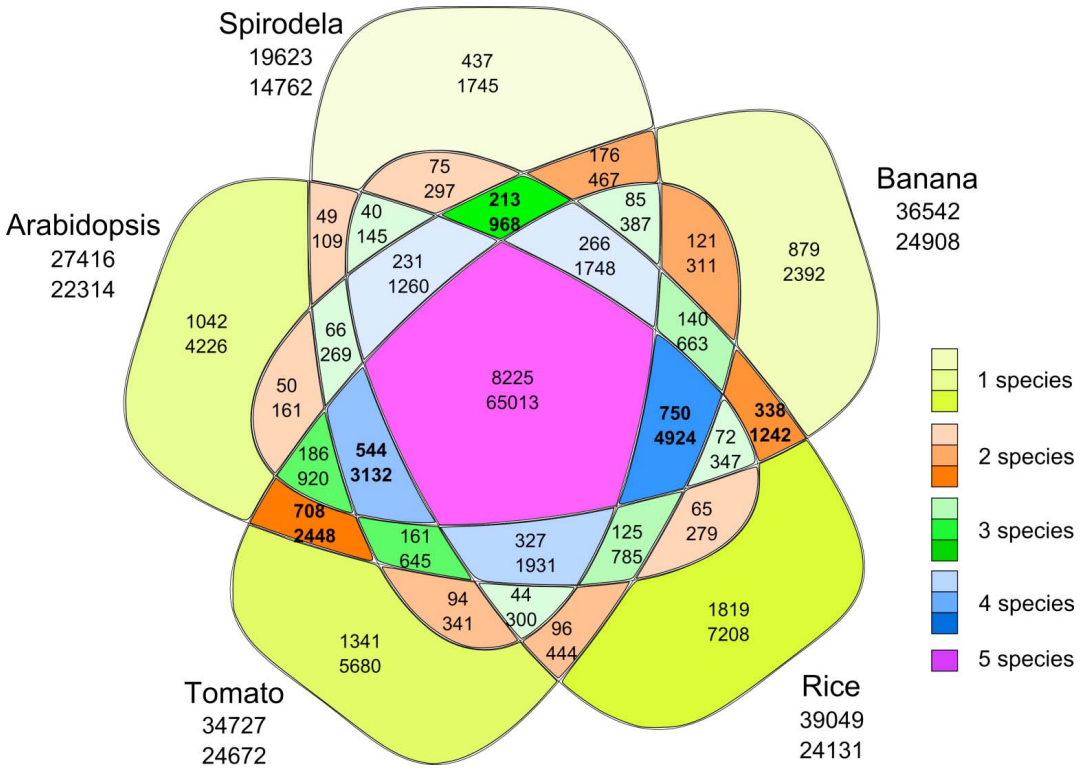


Figure 5.3 OrthoMCL analysis of gene families.

The Venn diagram illustrates shared and distinct cluster classes from an orthoMCL analysis of the plant proteomes of *Arabidopsis thaliana, Solanum lycopersicum, Spirodela polyrhiza, Musa acuminata* and *Oryza sativa indica*. Non-redundant data sets were used in the analysis. Numbers below each species name show number of all genes in species set and number of clustered genes, respectively. For each division in the Venn diagram, the top line shows the number of orthoMCL clusters and the bottom line the number of total genes in these clusters. The different cluster classes are color-coded by the number of species containing genes for the respective class (see legend bar at the right). Divisions for each class are shaded according to their abundance in their class with darker shades indicate larger contributions of the particular division.

Morphogenesis and plant body architecture

Expansins are cell-wall loosening proteins[18] involved in many plant processes including cell growth and expansion, root, and root hair expansion, fruit softening, ripening, and abscission[19]. We analyzed α- and β-expansins by an integrative pipeline

(Supplementary Fig. S26), which showed reduced copy number in Spirodela (Supplementary Fig. S27 and S28). Several clades of α-expansin genes were missing in Spirodela including AtEXP 2, 8, 17, 11, 7, and 18. The latter two expansins have been implicated in root hair initiation with AtEXP7 restoring a short root hair phenotype in rice indicating orthologous functions of this expansin in monocotyledonous species[20, 21]. Monocotyledonous plants have experienced a great expansion of β-expansins, with 10 detected in banana and on average 20 members in the *Poaceae* species. However, we detected only three β-expansins in Spirodela, indicating that the expansion continually progressed along the monocotyledonous diversification or a selective decrease of this gene family in Spirodela.

Due to the high buoyancy of their habitat, aquatic plants like duckweed do not require the up-straight structural support like land plants that would be consistent with a reduction of genes involved in cell wall biosynthesis and lignification. Whereas cell wall biogenesis genes such as CesA, CslA, CslC, and CslD were conserved across rice, Arabidopsis and Spirodela (Supplementary Table S18), there were two unique rice clades of CslF and CslH and two unique Arabidopsis clades of CslB and CslG. Spirodela is missing comparable members of CslB, CslE, CslF, CslG and CslH (Supplementary Fig. S29). We found all corresponding GT31 subfamily members in Spirodela, but the total copy number was 46.2% lower than in rice (Supplementary Fig. S30). The missing five clades in the Csl family and the fewer members in GT31 are consistent with the low content of 4-16% cellulose in Spirodela[22] in comparison to 62% in rice[23], which might indicate that the contraction or lack of amplification of the cellulose biosynthesis gene family in Spirodela reflects its reduced requirement for rigid cell walls.

Lignin, a major component of secondary cell wall, plays an important role for support, water transport, and stress responses in vascular plants. As shown in previous comprehensive phylogenetic analyses, most lignin biosynthesis gene families experienced rapid and recent duplications; the expansion mainly happened after the speciation between monocotyledonous and dicotyledonous species[24]. Whereas Spirodela contained nearly the entire lignin biosynthesis gene families with 9 out of 10 families (CAD, CCoAMT, 4CL, CCR, PAL, C4H, COMT, C3H, F5H but not HCT), gene copy number was significantly reduced compared to other monocotyledonous species like sorghum and rice (Supplementary Table S19). This was consistent with previous genome analysis, where gene copy number constituted an important evolutionary force for specialization and traits. In addition to genes catalyzing primary lignin biosynthesis, families involved in cell wall cross-linking and lignification also showed reduced copy numbers. We identified only seven members of the laccase multicopper enzymes in Spirodela, for which recent studies had provided experimental evidence for their role in lignification[25] (Supplementary Fig. S31). This was consistent with previous analyses of 3.1% lignin in Spirodela[22] in comparison with 18% in rice straw[23].

Ecological adaptation

Spirodela polyrhiza can undergo an environmentally induced developmental switch from protein-rich vegetative leaf-like “fronds” to a starch-rich dormant stage called “turion”[26]. Unlike the linked vegetative fronds via stipule, turions fall from the mother fronds once mature after starch accumulation. They sink to the bottom of a pond and germinate into new fronds by using starch as energy. These functions require genes

for starch biosynthesis including AGPase, SS plus GBSS, BE, and DBE. Spirodela contained very similar gene family compositions as Arabidopsis (Supplementary Table S20). The conservation of starch gene families from phylogenetic analysis for Spirodela, rice, maize, and Arabidopsis argues for their essential functions. The clades of AGPase large subunit and DBE had multiple members, whereas all others contained only one single member for the corresponding subgroup. SpBEIII did not cluster with any clade, but provided a separate branch as a Spirodela-specific BE member, suggesting that it might have evolved into a special function from their common ancestor (Supplementary Fig. S32).

The high growth rates of Spirodela require the efficient usage of nutrients. Nitrogen is generally a major limiting factor of plant growth and a primary component in fertilizers to promote crop growth. However, leaching of fertilizers, increasing amounts of sewage and wastewater from a steadily growing world population results in water pollution. Spirodela has been successfully exploited for wastewater remediation because of its ability to remove nitrogen with high efficiency, particularly in the form of ammonia, from polluted water[5]. Glutamine synthetase (GS) and glutamate synthase (GOGAT) are the core enzymes of the GS/GOGAT cycle in plants, the major biochemical module for ammonium-assimilation. Despite a genome-wide reduction in gene number, copy numbers of these enzymes were retained or even amplified in Spirodela with up to four times more copies of GOGAT in Spirodela compared to Arabidopsis and rice (Figure 5.4 and Supplementary Fig. S33).

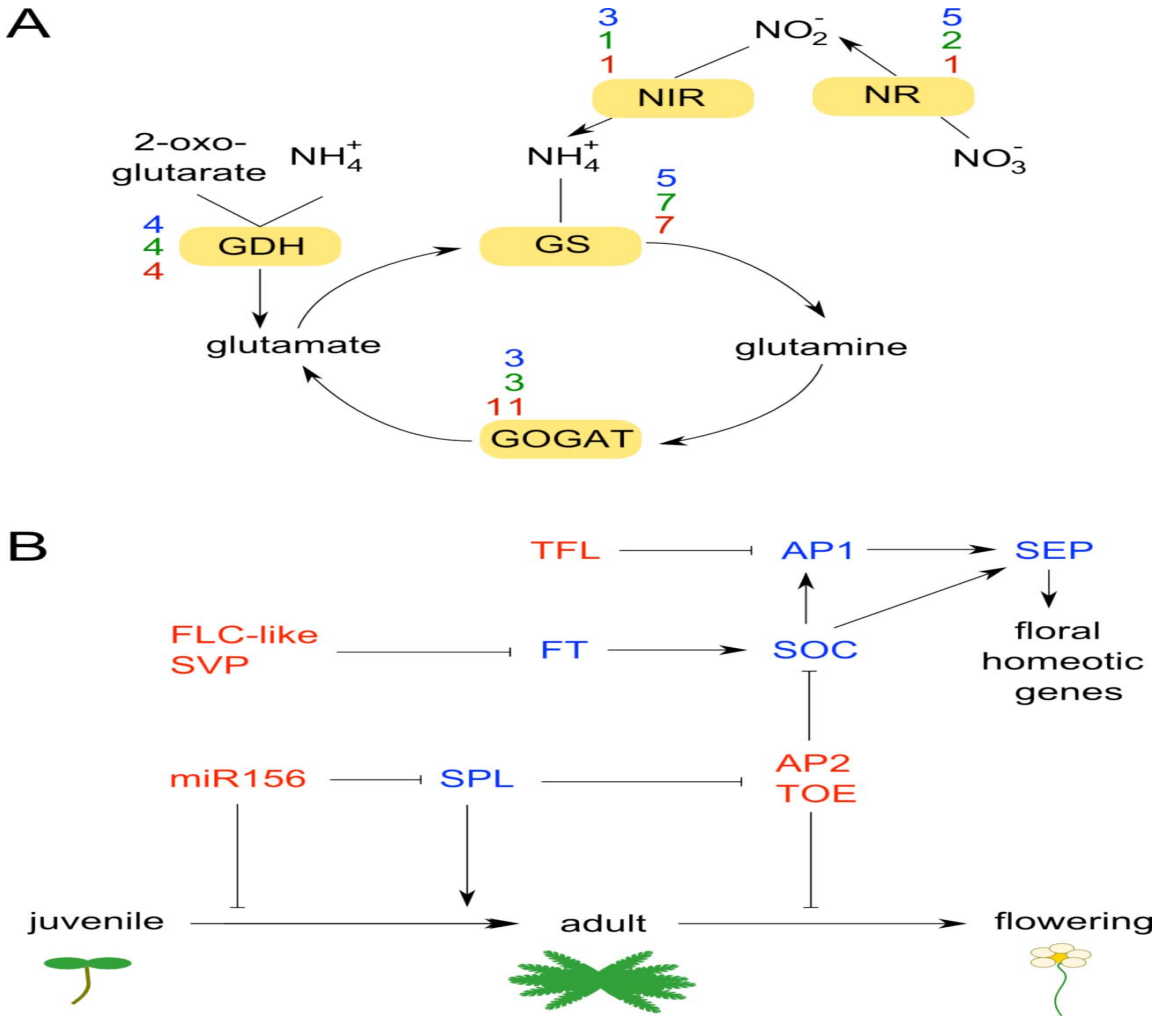


Figure 5.4 Spirodela characteristic pathways.

A shows a scheme of the nitrogen assimilation in higher plants. Spirodela shows a high overrepresentation of enzymes of the GS/GOGAT-cycle, the major module for ammonia-assimilation and consistent with the high ability of Spirodela to remove ammonia from sewage and wastewaters. Copy numbers for each gene are shown in red for Spirodela, blue for rice and green for Arabidopsis. B illustrates a highly simplified scheme of the regulatory network of the juvenile-to-adult phase transition in Arabidopsis. Most genes have several, functionally similar paralogs and are only shown for simplicity reasons by one gene symbol. For example, APETALA1, CAULIFLOWER and FRUITFUL are close paralogs promoting the onset of an inflorescence meristem but are represented only by one gene symbol, AP1. Gene groups having either similar copy numbers or being overrepresented in Spirodela are shown in red; those that have significantly reduced numbers are shown in blue.

Development and reproduction

Flowering plants undergo a series of distinct phase transitions during their life cycle including the progression from a vegetative, or juvenile phase to an adult phase with competency for sexual reproduction (flowering). Neoteny, the prolongation of juvenile traits, is a common phenomenon in the evolution of plant organs. The frond of the *Lemnoideae* has been characterized as embryonic or juvenile tissue, or as a cotyledon-like plant, iteratively bearing new cotyledons (Supplementary Fig. S1).

Whereas Spirodela has an increased copy number of repressors of the transition from juvenile to adult phase in comparison to Arabidopsis and rice, components of the regulatory network enhancing the progression through the adult phase and the onset of an inflorescence meristem were reduced (Figure 5.4, Supplementary Fig. S34, S35, S36, and Supplementary Table S21). In Arabidopsis, the microRNA of miR156 is necessary and sufficient to promote the juvenile phase and inhibit the transition to the adult growth[27]. Copies of miR156 were highly abundant in Spirodela, with 24 loci, or up to 32 loci if highly similar isoforms were included, consistent with the pattern of preferentially retained repressors of the adult phase, whereas Arabidopsis had only 10 and rice had 19 loci. The opposite was true for miRNA169, involved in drought tolerance[28], and miRNA172, involved in the switch from juvenile to adult phase[29], which were reduced from 9 and 5 copies found in Arabidopsis and tomato respectively, to one (Supplementary Table S12). We propose that the predominant vegetative reproduction and low flowering frequency as well as the reduced and simple plant body of Spirodela is at least in part a consequence of the re-engineering of the genetic network that controls transitions to the adult and flowering growth phases. Interestingly, structural reduction

increases in the *Lemnoideae* from the more ancient species like Spirodela towards the more derived members such as *Wolffia* (Figure 5.1). Genomes of duckweeds should therefore present an excellent opportunity to study how different degrees of neoteny translate into molecular changes of developmental networks and gene families.

5.4 Discussion

In higher plants, gene number and genome size seem not correlated. Whereas *Arabidopsis thaliana* has a genome size similar to Spirodela, it contains ~28% more genes; Spirodela has a gene count comparable to *Utricularia gibba* with a genome size of only 82 Mb. The low gene count of Spirodela could in part be due to the structural reduction and juvenile nature reducing the need for and consequently the retention or duplication of genes acting in the adult phase. In addition, Spirodela differs from previously reported angiosperm genomes in its lack of recent WGDs and retrotranspositions. The lower gene number in Spirodela may therefore be simply a consequence of the ongoing non-functionalization and loss of one copy of a duplicated gene pair, a major fate of gene duplication[30]. Because Spirodela provides us with a unique and fascinating biology, its genome sequence will serve us for future evolutionary and comparative genomic studies among angiosperms. Furthermore, developmental biologists can now take new approaches to study neoteny on the molecular basis. In addition to basic question in plant evolution and development, applications of duckweeds in water remediation and as a renewable energy source can now be further optimized. The genome sequence of Spirodela provides the first step to identify, understand, and improve relevant traits for specific target applications.

5.5 Material and Methods

A detailed description of all results and methods with their Fig. (S1-S36) and tables (S1-S21) are provided online as Supplementary Information (SI).

Sequencing

Sequencing reads for the nuclear genome were collected using the whole-genome shotgun sequencing strategy[8] with the Roche 454 XLR next-generation sequencing platform at the Department of Energy Joint Genome Institute in Walnut Creek, California

(http://www.jgi.doe.gov/sequencing/protocols/prots_production.html). Additional BAC and fosmid end sequences and 24 entire fosmids were obtained using standard protocols on ABI3730XL machines at the HudsonAlpha Institute in Huntsville, Alabama (Supplementary information).

Genome assembly and construction of pseudomolecules

The assembly was generated using Newbler version 2.6 with default parameters after trimming poor bases from ends and masking vector sequences. In total, 1071 scaffolds were assembled with an N50 of 3.7 Mb. The largest 252 collectively represent 141.8 Mb. The 32 pseudomolecules were generated with the Spirodela physical map and were validated by FISH and the 24 fosmid-sequences (Supplementary information).

Repeat analysis

The *de novo* searches for complete LTR retrotransposons were carried out with the program LTR-STRUC[31]. Additional complete LTR-retrotransposons were detected by homology searches against the 170 full-length sequences. The insertion age of those full-length LTR-retrotransposons was derived from the divergence between the left and right solo-LTR sequences.

Gene prediction and annotation

Gene models were derived from consensus gene predictions based on *de novo* gene finders, transcript and protein mapping. For all evidence by homology, spliced alignments were generated using GenomeThreader. For *de novo* gene finders (except the self-training GeneMark-ES-GC) a training set was derived from the Spirodela EST assemblies and high quality protein families mapped to scaffold sequences.

Acknowledgements

This work was supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231 and the Selman Waksman Chair in Molecular Genetics. The work was collaborated with the groups of Dr. Mayer, K.F.X., Dr. Schubert, I., Dr. Borodovsky, M. , Dr. Luo, MC and Dr. Schmutz, J.

5.6 References

1. Wang W, Wu Y, Yan Y, Ermakova M, Kerstetter R, Messing J: **DNA barcoding of the Lemnaceae, a family of aquatic monocots**. *BMC Plant Biol* 2010, **10**:205.
2. Bogner J: **The free-floating Aroids (Araceae)-living and fossil**. *Zitteliana* 2009, **48**:113-128.
3. Cheng JJ, Stomp AM: **Growing duckweed to recover nutrients from wastewaters and for production of fuel ethanol and animal feed**. *CLEAN - Soil, Air, Water* 2009, **37**(1):17-26.

4. Li J, Jain M, Vunsh R, Vishnevetsky J, Hanania U, Flaishman M, Perl A, Edelman M: **Callus induction and regeneration in Spirodela and Lemna**. *Plant Cell Rep* 2004, **22**(7):457-464.
5. Stomp AM: **The duckweeds: a valuable plant for biomanufacturing**. *Biotechnol Annu Rev* 2005, **11**:69-99.
6. Hillman W: **The Lemnaceae, or duckweeds**. *The Botanical Review* 1961, **27**(2):221-287.
7. Wang W, Kerstetter RA, Michael TP: **Evolution of Genome Size in Duckweeds (Lemnaceae)**. *Journal of Botany* 2011, **2011**:1-9.
8. Gardner RC, Howarth AJ, Hahn P, Brown-Luedi M, Shepherd RJ, Messing J: **The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing**. *Nucleic Acids Research* 1981, **9**(12):2871-2888.
9. Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y *et al*: **The sequence and de novo assembly of the giant panda genome**. *Nature* 2010, **463**(7279):311-317.
10. Wang W, Messing J: **High-throughput sequencing of three Lemnoideae (duckweeds) chloroplast genomes from total DNA**. *PLoS One* 2011, **6**(9):e24670.
11. Wang W, Wu Y, Messing J: **The mitochondrial genome of an aquatic plant, Spirodela polyrhiza**. *PLoS One* 2012, **7**(10):e46747.
12. Melters DP, Bradnam KR, Young HA, Telis N, May MR, Ruby JG, Sebra R, Peluso P, Eid J, Rank D *et al*: **Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution**. *Genome Biol* 2013, **14**(1):R10.
13. Richards EJ, Ausubel FM: **Isolation of a higher eukaryotic telomere from Arabidopsis thaliana**. *Cell* 1988, **53**(1):127-136.
14. Luo MC, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J: **High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis**. *Genomics* 2003, **82**(3):378-389.
15. Cai X, Xu SS: **Meiosis-driven genome variation in plants**. *Curr Genomics* 2007, **8**(3):151-161.
16. Ibarra-Laclette E, Lyons E, Hernandez-Guzman G, Perez-Torres CA, Carretero-Paulet L, Chang TH, Lan T, Welch AJ, Juarez MJ, Simpson J *et al*: **Architecture and evolution of a minute plant genome**. *Nature* 2013, **498**(7452):94-98.
17. Shi X, Wang X, Li Z, Zhu Q, Tang W, Ge S, Luo J: **Nucleotide substitution pattern in rice paralogues: implication for negative correlation between the synonymous substitution rate and codon usage bias**. *Gene* 2006, **376**(2):199-206.
18. Cosgrove DJ: **Loosening of plant cell walls by expansins**. *Nature* 2000, **407**(6802):321-326.
19. Choi D, Lee Y, Cho HT, Kende H: **Regulation of expansin gene expression affects growth and development in transgenic rice plants**. *The Plant Cell* 2003, **15**(6):1386-1398.

20. ZhiMing Y, Bo K, XiaoWei H, ShaoLei L, YouHuang B, WoNa D, Ming C, Hyung-Taeg C, Ping W: **Root hair-specific expansins modulate root hair elongation in rice**. *The Plant Journal* 2011, **66**(5):725-734.
21. Cho H-T, Cosgrove DJ: **Regulation of root hair initiation and expansin gene expression in Arabidopsis**. *The Plant cell* 2002, **14**(12):3237-3253.
22. Landolt E: **The family of Lemnaceae - a monographic study, Vols. 2**: Veroffentlichungen des Geobotanischen Institutes der Eidgenossischen Technischen Hochschule, Stiftung Rubel.; 1987.
23. Gani A, Naruse I: **Effect of cellulose and lignin content on pyrolysis and combustion characteristics for several types of biomass**. *Renewable Energy* 2006, **2007**:649-661.
24. Xu Z, Zhang D, Hu J, Zhou X, Ye X, Reichel KL, Stewart NR, Syrenne RD, Yang X, Gao P *et al*: **Comparative genome analysis of lignin biosynthesis gene families across the plant kingdom**. *BMC Bioinformatics* 2009, **10 Suppl 11**:S3.
25. Berthet S, Demont-Caulet N, Pollet B, Bidzinski P, Cezard L, Le Bris P, Borrega N, Herve J, Blondet E, Balzergue S *et al*: **Disruption of LACCASE4 and 17 results in tissue-specific alterations to lignification of Arabidopsis thaliana stems**. *The Plant cell* 2011, **23**(3):1124-1137.
26. Wang W, Messing J: **Analysis of ADP-glucose pyrophosphorylase expression during turion formation induced by abscisic acid in Spirodela polyrhiza (greater duckweed)**. *BMC Plant Biol* 2012, **12**:5.
27. Wu G, Park MY, Conway SR, Wang JW, Weigel D, Poethig RS: **The sequential action of miR156 and miR172 regulates developmental timing in Arabidopsis**. *Cell* 2009, **138**(4):750-759.
28. Zhao B, Liang R, Ge L, Li W, Xiao H, Lin H, Ruan K, Jin Y: **Identification of drought-induced microRNAs in rice**. *Biochemical and biophysical research communications* 2007, **354**(2):585-590.
29. Lauter N, Kampani A, Carlson S, Goebel M, Moose SP: **microRNA172 down-regulates glossy15 to promote vegetative phase change in maize**. *Proc Natl Acad Sci USA* 2005, **102**(26):9412-9417.
30. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes**. *Science* 2000, **290**(5494):1151-1155.
31. McCarthy EM, McDonald JF: **LTR_STRUC: a novel search and identification program for LTR retrotransposons**. *Bioinformatics* 2003, **19**(3):362-367.
32. Bremer K: **Early Cretaceous lineages of monocot flowering plants**. *Proceedings of the National Academy of Sciences* 2000, **97**(9):4707-4711.
33. Landolt E: **The family of Lemnaceae - a monographic study, Vols. 1**: Veroffentlichungen des Geobotanischen Institutes der Eidgenossischen Technischen Hochschule, Stiftung Rubel; 1986.

CHAPTER 6 STARCH SYNTHESIS AT TURION FORMATION

6.1 Abstract

Aquatic plants differ in their development from terrestrial plants in their morphology and physiology, but little is known about the molecular basis of the major phases of their life cycle. Interestingly, in place of seeds of terrestrial plants their dormant phase is represented by turions, which circumvents sexual reproduction. However, like seeds turions provide energy storage for starting the next growing season.

To begin a characterization of the transition from the growth to the dormant phase we used abscisic acid (ABA), a plant hormone, to induce controlled turion formation in *Spirodela polyrhiza* and investigated their differentiation from fronds, representing their growth phase, into turions with respect to morphological, ultra-structural characteristics, and starch content. Turions were rich in anthocyanin pigmentation and had a density that submerged them to the bottom of liquid medium. Transmission electron microscopy (TEM) of turions showed in comparison to fronds shrunken vacuoles, smaller intercellular space, and abundant starch granules surrounded by thylakoid membranes. Turions accumulated more than 60% starch in dry mass after two weeks of ABA treatment. To further understand the mechanism of the developmental switch from fronds to turions, we cloned and sequenced the genes of three large-subunit ADP-glucose pyrophosphorylases (*APLs*). All three putative protein and exon sequences were

conserved, but the corresponding genomic sequences were extremely variable mainly due to the invasion of miniature inverted-repeat transposable elements (MITEs) into introns. A molecular three-dimensional model of the SpAPLs was consistent with their regulatory mechanism in the interaction with the substrate (ATP) and allosteric activator (3-PGA) to permit conformational changes of its structure. Gene expression analysis revealed that each gene was associated with distinct temporal expression during turion formation. *APL*2 and *APL*3 were highly expressed in earlier stages of turion development, while *APL*1 expression was reduced throughout turion development.

These results suggest that the differential expression of *APL*s could be used to enhance energy flow from photosynthesis to storage of carbon in aquatic plants, making duckweeds a useful alternative biofuel feedstock.

6.2 Introduction

Duckweed is an aquatic plant seen on water surfaces in many locations in the world. Because it consists mainly of a leaf-like body, called fronds that performs photosynthesis, it is probably the most efficient multicellular biological solar energy converter that we have. There are conditions like temperature shifts due to seasons that can cause a morphological change to a different structure, called turions. Many species of the subfamily *Lemnoideae* can produce this kind of dormant fronds, which are characterized by more starch, smaller vacuoles and air space [1, 2]. This developmental

change is also accompanied by a shift in metabolism. The energy harvested during photosynthesis is shifted to starch biosynthesis, resulting in the accumulation of starch in turions. Because the volume of intercellular air space shrinks and starch increases the density of the tissue, it can sink to the bottom of waters where the organism can survive even if the top of the water freezes. Turions can change back to fronds vegetatively using the starch as an energy source, demonstrating a highly evolved adaptation to the environment. Because fronds have little lignin, which would interfer with the digestion of the carbohydrate fraction of biomass, and turions have high starch content, duckweed might also be suitable as an alternative source of bioenergy. Whereas cellulose is a crystalline, compact and structural compound resistant to biological attack and enzymatic degradation, starch is readily digested. Even though many advances over the past years have been made in the commercialization of cellulosic biomass [3], the cost of producing equal amounts of ethanol from cellulosic biomass is still much higher than production directly from starch [4]. Therefore, growing attention is being devoted to use duckweeds as a source of carbon compounds and convert duckweed biomass into bio-ethanol [5]. Fronds growing in swine wastewater contain 45.8% (dry weight) of starch. Moreover, 50.9% of the original dry biomass can be enzymatically hydrolyzed into a reducing sugar, which contributes to 25.8% fermented ethanol of dry biomass [5].

Recent studies have focused on the influence of various environmental conditions for turion formation or germination [6-10], the sensitivity threshold of ABA for turion

formation [8, 11] and the different structure (air space, vacuole, starch and cell wall) of fronds and turions [2]. On the other hand, information of starch content, granule size, and derivation of starch granules involvement with turion formation, which is critical to explore the potential biofuel of duckweed, is less well understood.

The pathway of starch synthesis is very complex, but ADP-glucose pyrophosphorylase (AGPase) plays a pivotal role in regulating starch levels and in determining patterns of starch deposition in plants. This enzyme comprises two identical large subunits (APLs) and two same small subunits (APSs) in angiosperms, each of which is encoded by distinct genes. Even though the roles of each AGPase subunit in the enzyme are not clear, it is generally proposed that APLs modify the response to allosteric regulators, whereas APSs act as the catalytic part [12]. Recent studies suggest that AGPase are usually in plastidial forms except for a cytosolic one in cereal endosperms [13, 14]. Here, we compared the distinctive attributes between fronds and turions in *S. polyrhiza* and investigated starch production during development upon induction with abscisic acid (ABA), a plant hormone. To gain further insight into the function of the large subunit of AGPase (*APLs*) in starch synthesis as well, we cloned the *Spirodela* genes, analyzed them, and quantified their expression, which will allow in the future targeting expression of transgenes.

6.3 Results

Turion induction with ABA

Spirodela polyrhiza was grown under controlled light conditions as described under Methods. Fronds were harvested and examined under a dissecting microscope. Dividing fronds, representing single leaf-like bodies, were connected, thin, and elliptical (~8mm in length and ~6mm in width). The top of fronds was bright green, whereas the bottom extended a few roots that were submerged into water (Figure 6.1A). Continued growth in the presence of ABA gave rise to turion formation with different morphological features (Figure 6.1B). After 5 days of ABA application a significant shift to starch accumulation took place in collected samples from both wet and dry tissues. Starch accumulation during turion development exhibited a characteristic pattern. There was a progressive increase of starch from 5 to 10 days after ABA application and after 14 days, the starch content became almost stable. The final starch content in turions for wet tissues was 24.4%, which corresponds to 60.1% in dry mass (Figure 6.2A). Turions were also harvested and examined under a dissecting microscope. They appeared thicker and smaller in nearly round shape (~2mm in length and ~3mm in width). Turions were dark green, spotted with many anthocyanin pigments, and retained only rudimentary roots that are not visible by naked eye (Figure 6.1B).

Figure 6.1 Morphological comparison of frond and turion.

A) dorsal and ventral frond; B) dorsal and ventral turion. The turion is formed after 14 days of ABA treatment. Bars= 1 mm.

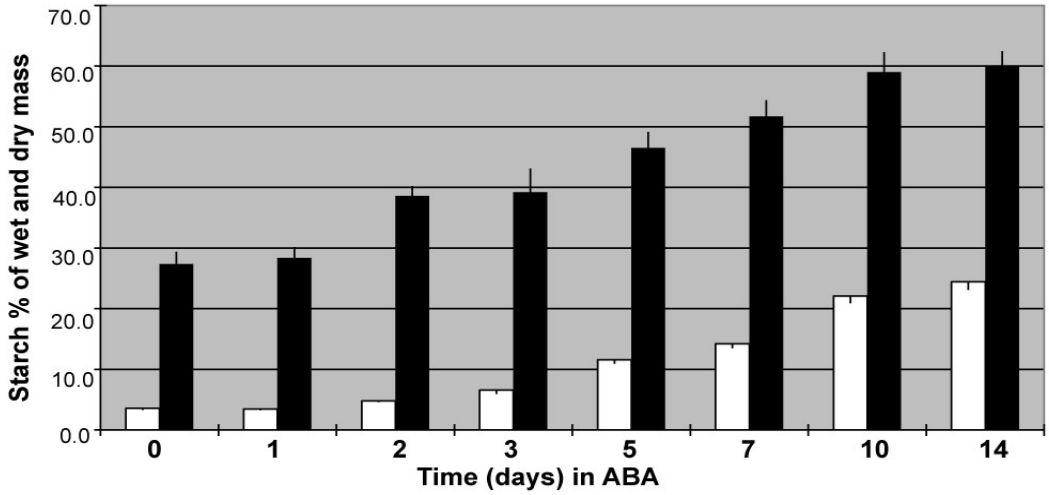


Figure 6.2 Starch accumulation during turion development.

White bars stands for wet tissue and black bars for dry tissue. Y-axis shows starch content (mg) for every 100 mg wet tissue or dry tissue.

Frond samples were then examined by electron microcopy. The frond cell had normal discal chloroplasts with a few small starch grains (Figure 6.3A and Figure 6.3C). Most frond cells contained a single larger vacuole and bigger intercellular air space, while turion cells have multiple smaller vacuoles and bigger air space between cells (Figure 6.3B). The turion cell accumulated many starch granules, which almost occupied 1/4 to 2/3 of cell volume (Figure 6.3B and Figure 6.3D). The kidney-shaped starch granule was surrounded with stacks of thylakoid membranes in chloroplasts (Figure 6.3E and Figure 6.3F). The increased starch granules at the expense of the vacuolar expansion also contributed to the distortion of chloroplasts (Figure 6.3E) and a shift in tissue density that caused turions to sink to the bottom of liquid medium (left panel of Figure 6.4A). Placed on filter paper, they looked like “green seeds” compared to fronds (Figure 6.4B).

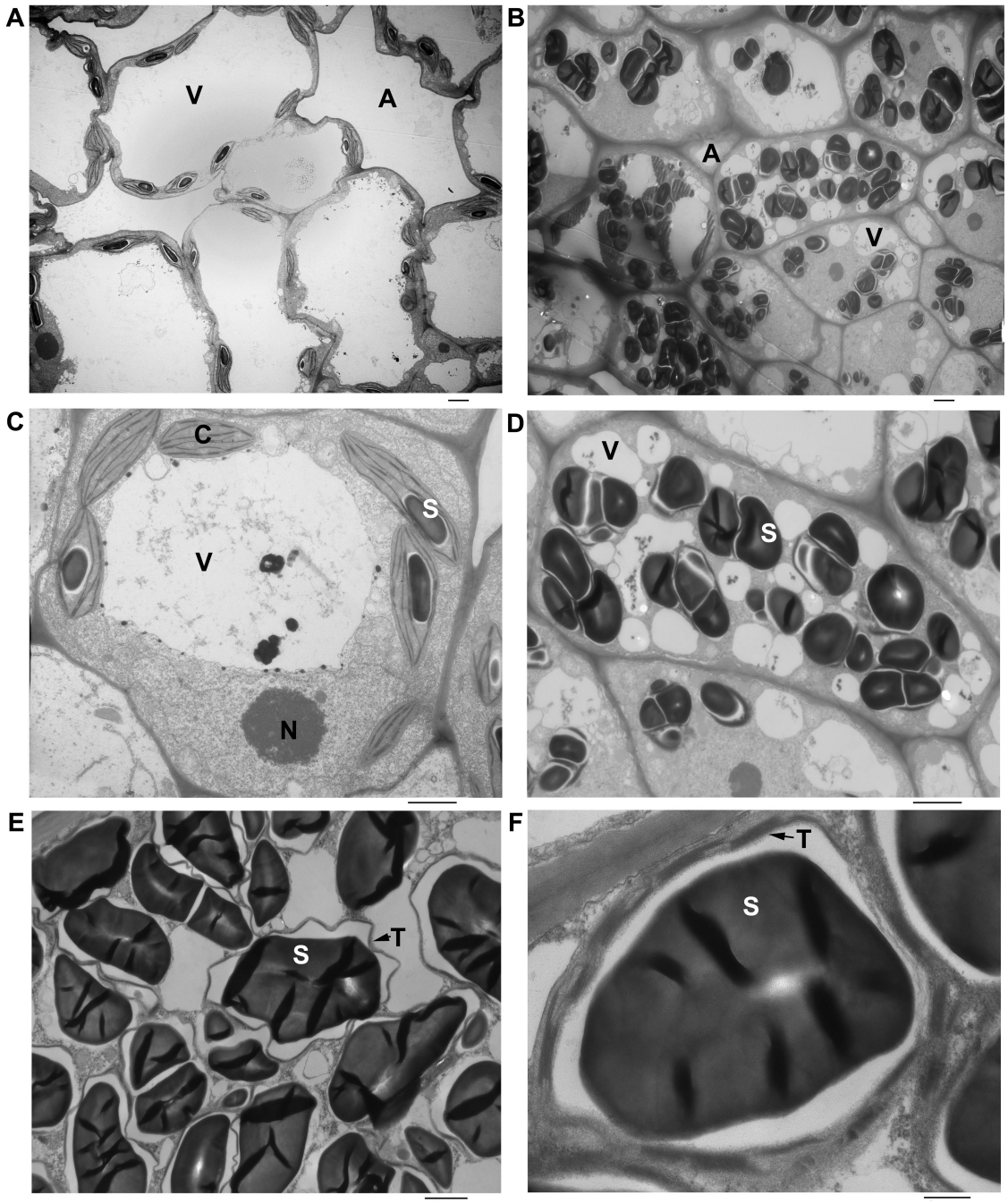


Figure 6.3 Microscopic study.

A) Transmission electron microscopic (TEM) picture of frond cells with lower magnification, Bars = 2 μm; B) TEM picture of frond cells with lower magnification, Bar

= 2 μm; C) TEM picture of a frond cell with higher magnification, Bar = 2 μm; D) TEM picture of a turion cell with higher magnification, Bar = 2 μm ; E) TEM picture of a section of a turion cell with higher magnification, Bar = 2 μm; F) TEM picture of a section of a turion cell with the highest magnification, Bar = 500 nm. Abbreviations are chloroplast (C), starch granule (S), vacuole (V), intercellular air space (A), thylakoid membrane (T), nucleus (N).

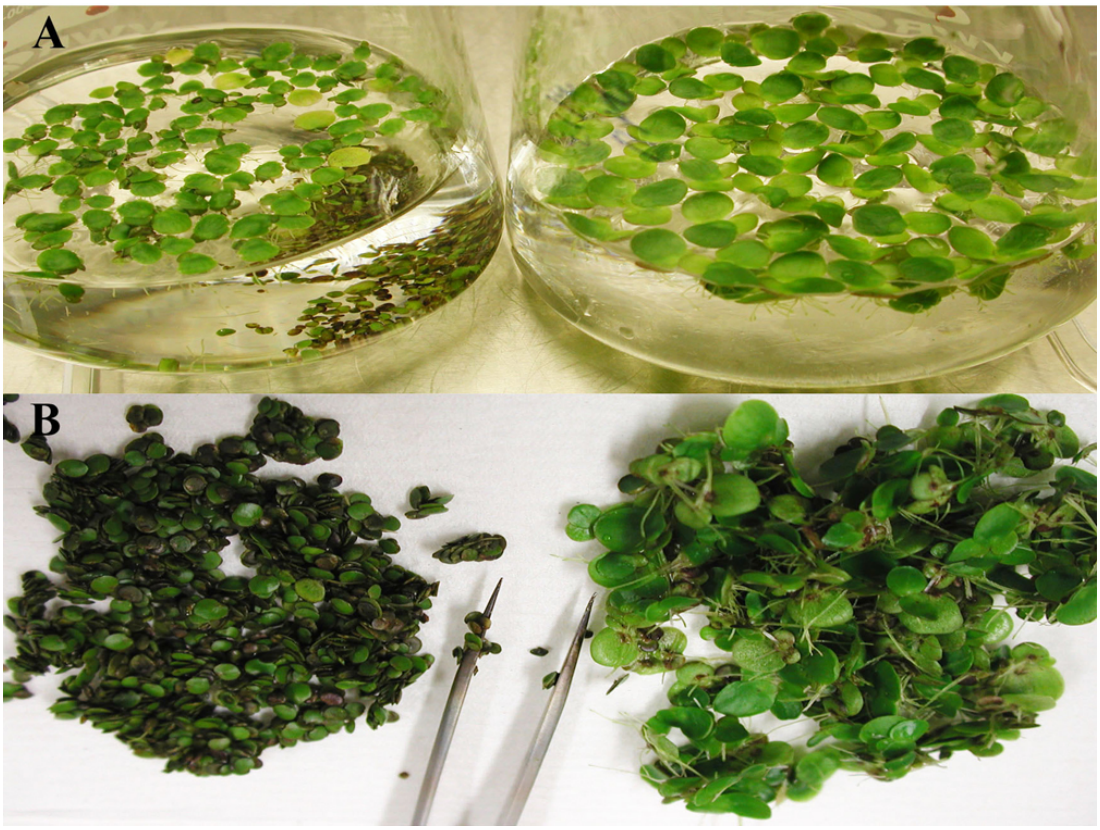


Figure 6.4 Fronds and Turions in flask.

A) Turions (left panel) on the bottom and fronds swimming with roots down (right panel) in flasks; B) turions (left) and fronds (right) placed on filter paper. Bars= 1 mm.

Cloning and sequencing of members of the *Spirodela APL* gene family

The level of starch accumulation in turions (Figure 6.2) and the convenience of collecting them from the flask bottom (Figure 6.4) are key features for biofuel applications as described above. To examine the metabolic regulation of these features, this study seeks to identify key enzymes, whose manipulation at the molecular level could optimize the timing and level of starch production. Common knowledge would then suggest investigating the differential expression of key enzymes in starch biosynthesis. Therefore, we decided to clone the large subunit of the ADP-glucose pyrophosphorylase gene family (*APLs*) from *Spirodela polyrhiza*. Because this gene is very conserved among angiosperms, we used the known *Arabidopsis* protein sequences to design degenerate primers to amplify *APL* coding sequences as described under Methods. Cloned DNA fragments were then sequenced and overlapping fragments were used to reconstruct the entire three cDNA-copies from *Spirodela*. We named them *SpAPL1*, *SpAPL2* and *SpAPL3* with Genbank accession numbers of JN180634, JN180635, JN180636. Based on the cDNA sequences primers were then designed to clone the corresponding gene sequences from total genomic DNA as described under Methods. The cloned genes of *SpAPL1*, *SpAPL2* and *SpAPL3* were then also sequenced and deposited into GenBank with accessions JN180631, JN180632, JN180633, respectively. After aligning cDNAs with their corresponding genomic sequences, all introns could be identified. Accordingly, all *SpAPLs* consisted of 15 exons and 14 introns

(Figure 6.5). Whereas the coding sequences of the *SpAPL1*, *SpAPL2* and *SpAPL3* genes were slightly different in length with 1,554, 1,611, 1,620 bp or 517, 536, and 539 amino acids, respectively, the corresponding genomic regions differed significantly with 8,449, 4,684 and 3,460 bp (Table 6.1), reflecting intron expansions.

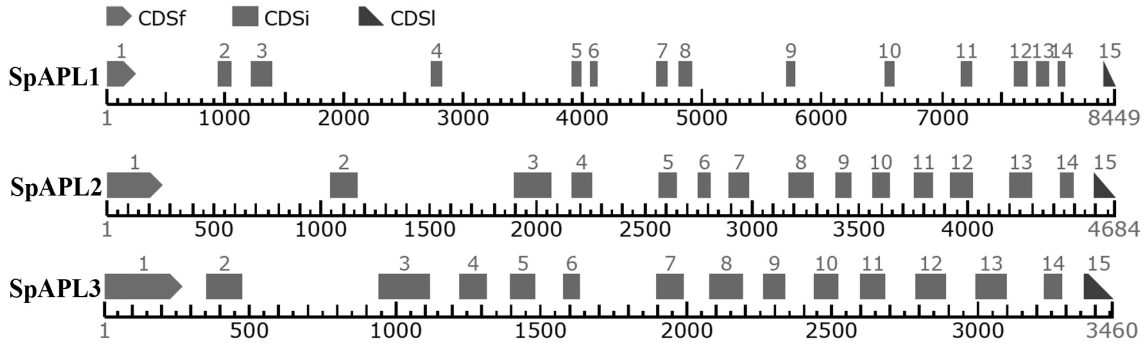


Figure 6.5 Structural organization of the *SpAPL* genes.

The coding exons were depicted as gray boxes. Introns were depicted as bar scaled by x-axis (bp). CDSf - First (Starting with Start codon), CDSi - internal (internal exon), CDSl - last coding segment (ending with stop codon).

Structure and phylogeny of members of the *Spirodela APL* gene family

The basis for the variation in protein sizes became clear when their primary structures were compared with other known APLs. Sequence alignments of the deduced amino acid sequences of SpAPL1, SpAPL2, and SpAPL3 proteins showed high homology except for their N-terminal regions. APLs are usually targeted to the plastid through a signal peptide at their amino-terminus. SpAPL2 and SpAPL3 had conserved

plastid-targeting signals with cleavage sites at positions 78 and 64 based on TargetP (http://www.cbs.dtu.dk/services/TargetP/). The shorter protein of SpAPL1 had a very weak targeting signal and internal deletions similar to the rice APLs. Although there were differences in the amino-terminal regions, the coding sequences from exon 3 to 15 were of the same size and very conserved.

The corresponding introns, however, have diverged significantly in length and composition. Interestingly, comparison to the TIGR Plant Repeat Database [15] indicated that expansion of introns could be largely due to miniature inverted-repeat transposable elements (MITEs). When the MUST system was applied that was used to predict MITEs rather than depending on sequence homology alone, the sequence data suggested then that MITEs had invaded the introns of *SpAPL1*, *SpAPL2*, and *SpAPL3*, comprising now 36%, 21%, and 7% of total intron sequences, respectively (Table 6.1).

Gene Name	**Gene Length (bp)**	**ORF Length (bp)**	**Putative Protein Length (aa)**	**Intron Length (bp)**	**MITE Length (bp)**	**Ratio=MITE/ Intron**
SpAPL1	8449	1554	517	6895	2507	0.36
SpAPL2	4684	1611	536	3073	659	0.21
SpAPL3	3460	1620	539	1840	126	0.07

Table 6.1 Gene features of *APL* family.

Using the amino acid sequence alignment of *APL*s from *S. polyrhiza*, rice, and

maize, we constructed a maximum likelihood phylogeny of the *APL* family. This phylogenetic tree separated the *APL*s into three main clades: *SpAPL1* clustered together with the plastidial forms of *OsAPL1* and *ZmAPL1* in branch APL-I. *SpAPL2* shares the branch APL-II with the plastidial forms of *OsAPL4* and *ZmAPL4*; *SpAPL3* shares a common ancestor with both plastidial (*OsAPL3* and *ZmAGP1*) and cytosolic forms (*OsAPL2* and *ZmSH2*) in rice and maize [16] (Figure 6.6).

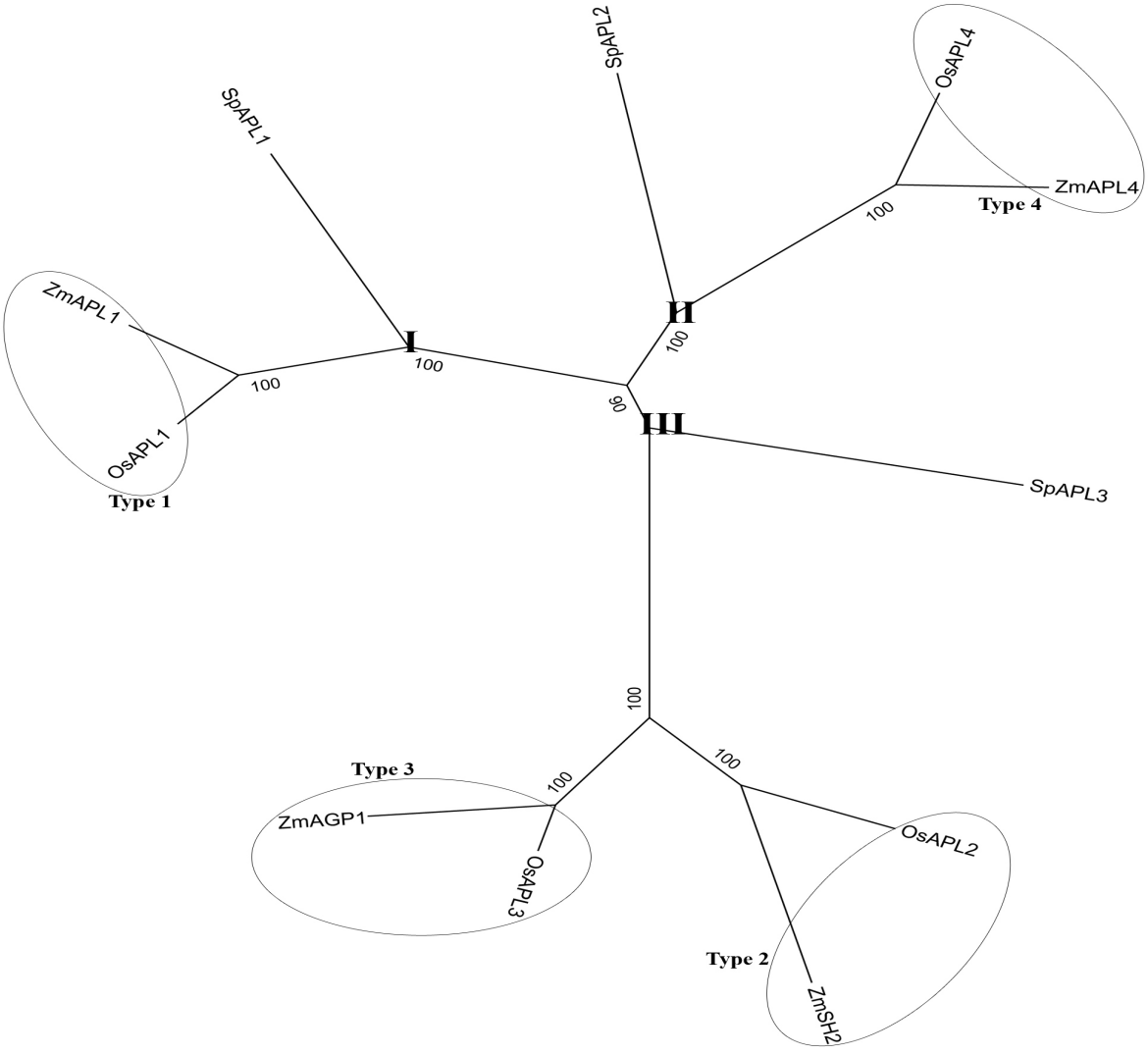


Figure 6.6 APL Phylogenetic tree.

The analysis is based on the amino acid sequence alignment of large subunits of AGPase (*APLs*) from *S. polyrhiza* (Sp), *Oryza sativa* (Os), and *Zea mays* (Zm). The protein names are those published previously or predicted from CDS: OsAPL1, NP_001051184; ZmAPL1, NP_001106017; SpAPL1, JN180634; OsAPL2, NP_001043654; ZmSH2, NP_001121104; OsAPL3, NP_001056424; ZmAGP1, NP_001105717; SpAPL3, JN180636; OsAPL4, NP_001059276; ZmAPL4, NP_001106058; SpAPL2, JN180635. Three clades were designated APL-I, APL-II, APL-III. The classification of APLs in grasses has previously been published [16].

A structural model of the APLs

To confirm the inference of their function, three-dimensional structures of SpAPLs were built by using the experimental protein structure (PDB 1yp3) from potato as a suitable template. Amino acid sequence alignment of the regulatory site of APLs from potato and *S. polyrhiza* showed five key conserved residues (P44, P52, P66, K414 and K452) (Figure. 7A) in all three SpAPLs. Molecular modeling analysis of APLs suggested a critical role of APLs for allosteric regulation in this region with binding sites for ATP and 3-PGA (Figure. 7B). P44 was important for accommodating ATP phosphate groups, as it was located between a conserved GGXGXRL loop region and the strongly conserved “PAV” region, which involved catalysis and allosteric regulation [17]. P52

was predicted to be located in flexible loops close to the lysine residues (K414 and K452), while P66 lied in a helix. Site-directed mutagenesis of the P52 and P66 in potato showed dramatic changes in affecting enzyme regulatory properties, while P44 mutants resulted in a nearly catalytically inactive enzyme [18]. K414 and K452 were shown to be involved in the increase of the affinity for the activator 3-PGA [17, 19]. Model structures of APL1, APL2 and APL3 were identical in these features. Therefore, only APL1 was shown in Figure. 7 as an example.

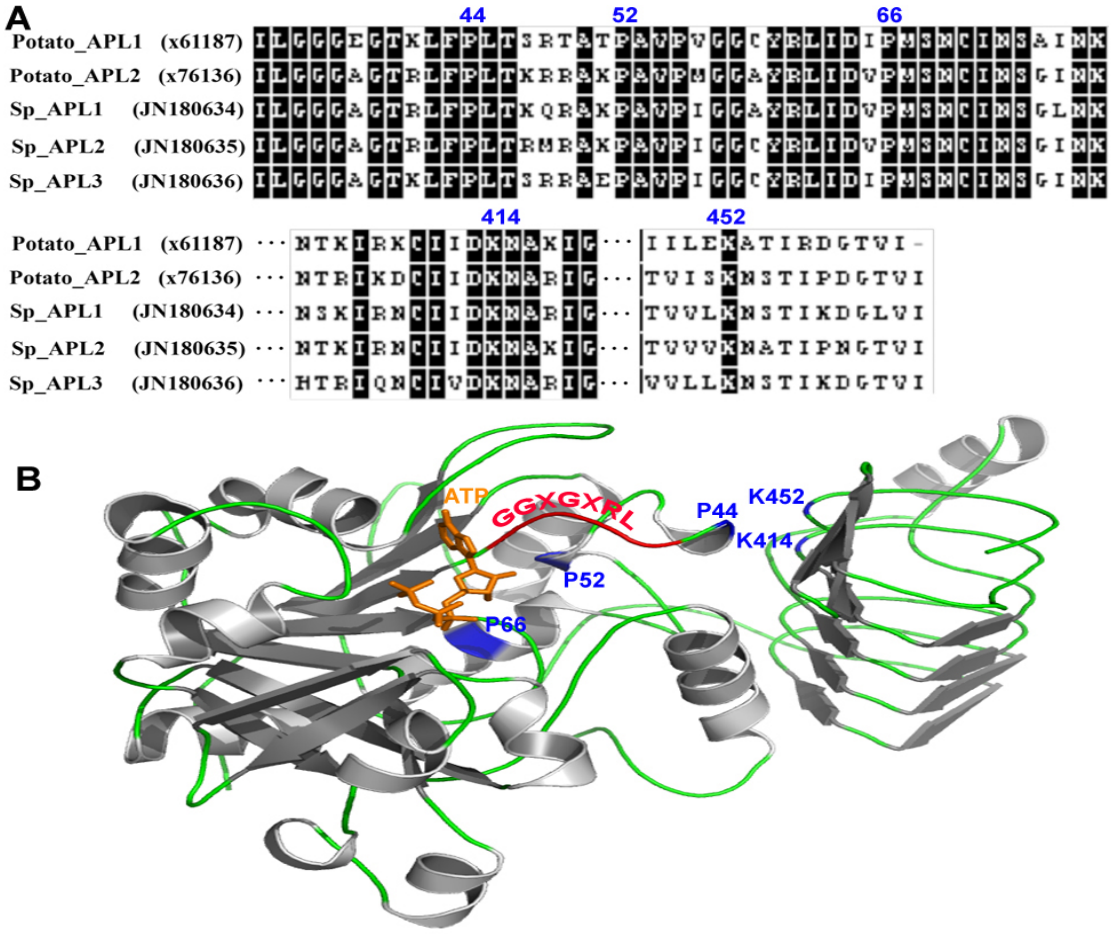


Figure 6.7 A structural model of the APL.

A) Amino acid sequence alignment for the regulatory sites of APLs between potato and *S. polyrhiza*. GenBank accession numbers were listed in parentheses. Important proline (P44, P52 and P66) and lysine (K414 and K452) residues critical for allosteric regulation were numbered corresponding to potato AGPase large subunit (x61187). Identical residues were shaded in black. B) Modeled structure of *S. polyrhiza* APL1. The N-terminal region containing the putative ATP binding site and the regions containing the putative 3-PGA binding sites of APL1 were modeled by comparison with the known structure of the potato AGPase small subunit (PDB 1yp3) with 54.93% identity. The modeled position of ATP in orange was shown. The α-helix and β-sheet were colored in gray and the loop was in green. Important proline (P) and lysine (K) residues in APL1 were indicated by blue color. The conserved GGXGXRL loop region was in red.

Expression patterns of *APL* genes in developing turions

With three different gene copies present in *Spirodela polyrhiza*, the question arises how each enzyme is expressed temporally during turion formation. We therefore isolated total mRNA from leaf-like tissue 0, 1, 2, 3, 5, and 7 days after the addition of ABA. To measure expression of each APL gene copy, we applied qPCR to mRNA samples using specific primer pairs to distinguish between transcripts from each gene. Expression of *SpAPL2* and *SpAPL3* dramatically increased two- and ten-fold,

respectively, as turion development was initiated (1 to 3 days). Furthermore, there seemed to be a difference in the expression of *SpAPL2* and *SpAPL3*. *SpAPL2* was critically in the first phase of induction, whereas *SpAPL3* seemed to replace *SpAPL2* in a second burst of activity. There was no obvious increased expression of *SpAPL1* after ABA induction. Indeed, *SpAPL1* appeared to be more active in initial fronds compared to *SpAPL3* (0 days of ABA application). When turions went into mature phase (after 5 days), the expression of all *SpAPLs* was leveling off (Figure 6.2B).

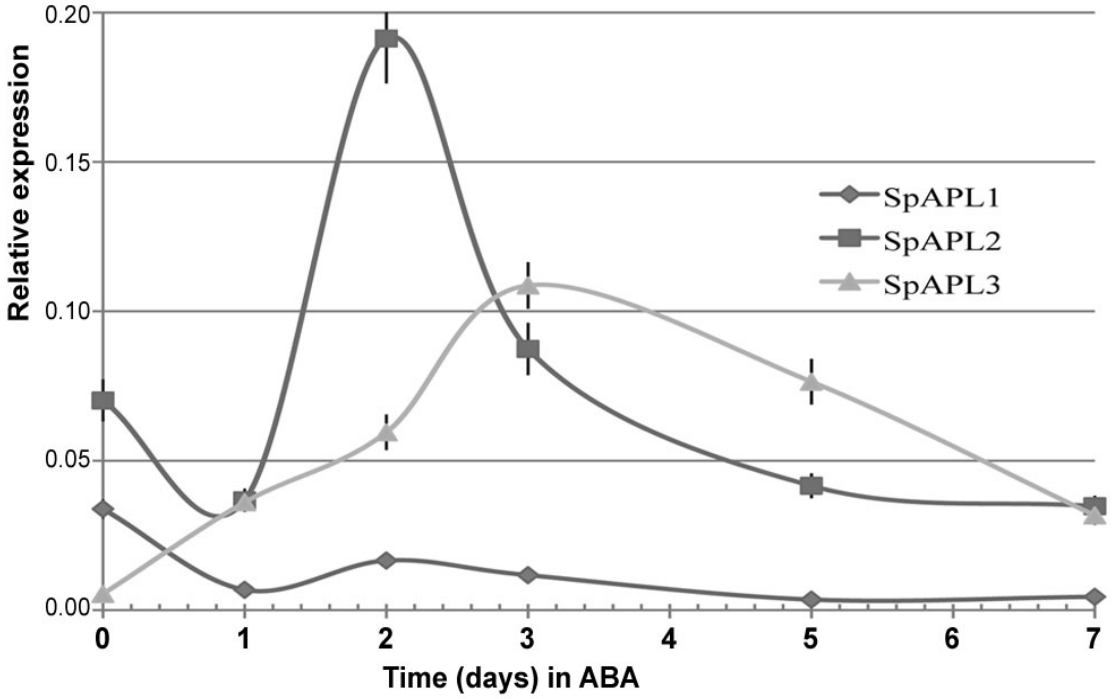


Figure 6.8 APL gene expression.

qPCR was used to quantify expression of APLs based on RNA from 0, 1, 2, 3, 5, 7 days of ABA treatment. Standard error was shown by vertical bar.

6.4 Discussion

We began dissecting the process of turion formation in duckweeds. Usually turion development occurs in late summer or early autumn because of starvation and lower temperatures [20]. *Spirodela* turions can also be induced under controlled laboratory conditions by increasing the concentration of ABA in the growth medium [6, 8], decreasing temperature [10], or depriving phosphorus in the medium [7]. Here, we have taken advantage of ABA as an inducer and could reproduce the morphological changes that occur during turion formation. Turions are germinated into new fronds in the presence of light and nitrogen in the following spring using starch storage as an energy source [21, 22]. Therefore, the drastic starch accumulation during turion formation marks a turning point in the switch process from low-starch fronds to high-starch turions.

The reported contents of starch varied from 14% to 43% depending on the species, developmental states (fronds, resting fronds, or turions) [23] and tested methods [24, 25]. Starch content could even go up to 75% of the dry weight in resting fronds of *Spirodela oligorrhiza* (renamed into *Landoltia punctata*) growing in phosphor-deficient cultures [7], a level that is comparable to cereal seeds of corn, sorghum and wheat [26]. Even though regular fronds have as low as 16% starch in dry mass, turions of *S. polyrhiza* can reach up to 62% starch [20]. Our use of exogenous ABA produces the same developmental switch, as the different morphological features are easily distinguishable. The switch is rapid, providing advantages for biochemical and physiological analysis [8].

We obtained 60.1% starch from dry mass after two weeks of ABA induction (Figure 6.2A), which is comparable to the Henssen's study. The size of mature starch grains from turions was around 4 μm in diameter as estimated by TEM (Figure 6.3E and 3F), whereas starch grains from wheat, corn and rice reach a size of 30 μm, 25 μm and 20 μm, respectively [27]. In a different study, the size of starch granules illuminated by red light for different times have also been measured using SEM scans arriving at similar values [28]. Interestingly, it has been suggested that smaller starch granules are more easily hydrolyzed into sugars than larger ones, regardless of botanical source [29]. After 72 h of continuous irradiation, the sizes of starch granules in turions are significantly reduced to about 1.5 μm [28]. Although duckweeds might have adapted to rapidly switch back to a growth phase faster than seed plants, this property also might provide a more efficient way for producing bio-ethanol than from maize.

Amyloplasts in non-photosynthetic tissue, such as seeds, roots, and stems, which lack chlorophyll and internal membranes, are the main organelles responsible for the synthesis and storage of starch granules in most plants. However, turions remain green or dark-green throughout their development (Figure 6.1B and 4B). The plastids in turions, where starch synthesis takes place, still retain abundant stacks of thylakoids (Figure 6.3E and Figure 6.3F). It therefore suggests that chloroplasts with a simple structure as in duckweeds can function both as source and sink. The starch-storing plastids of turions are directly derived from chloroplasts, and retain chloroplast-like characteristics throughout

their development. This adaptation greatly saves energy by directly depositing sucrose generated from photosynthesis into starch storage without the need for transport through a vascular system and the use of a glucose phosphate transporter [16]. A similar system exists also in a non-aquatic plants such as pea embryos, where starch-storing plastids also directly originate from chloroplast [30, 31]. Moreover, using TEM light-induced degradation of starch granules in turions of *Spirodela polyrhiza* also exhibited a transition from amyloplasts to chloroplasts [28]. Both studies would demonstrate that differentiation from chloroplast to amyloplast could be reversed based on physiological changes. Indeed, the cell structure of turions appears to be well organized for its function. Its lack of intercellular air space and presence of smaller vacuoles allow them to survive in deep water, where the temperature is more moderate than on the surface. The numerous starch grains provide a bank of energy when turions germinate in the following spring. This life cycle is also consistent with starch content in fronds and turions.

Because starch biosynthesis is an important feature for the developmental switch from fronds to turions, it also provides us with the first entry point to dissect the developmental regulation of turion formation. Therefore, we reasoned that the first step in this line of investigation consists of the identification and characterization of key regulatory genes known in starch biosynthesis, which are the ADP-glucose pyrophosphorylases. We successfully cloned three copies of *APLs* of *Spirodela polyrhiza*. *APLs* are expressed in different organs of grass species, type 1 in leaves, type 2 and type

3 in seeds, and type 4 in both seeds and leaves [16, 32]. Based on phylogeny and spatial expression of *SpAPLs* (Fig. 6 and Figure 6.2B), they have their homologs in grass species. *SpAPL2* and *SpAPL3* are active in turions, while *SpAPL1* is expressed at a higher level in fronds. The transcript level of *SpAPL2* and *SpAPL3* are active at an early phase of turion formation, while all transcript level of *SpAPLs* decline towards the end phase. It could account for the inhibition of total RNA synthesis after 3 days in ABA, which leads to the shutdown of all primary processes and onset of the dormant state [33]. Analysis of networks of gene expression during *Arabidopsis* seed filling has also shown that expression of carbohydrates occurred early in seed development [34]. Noticeably, the transcription of *SpAPL1* and *SpAPL2* is suppressed right after one day of ABA addition, which is quite consistent with previous findings that ABA could inhibit DNA, protein, and RNA synthesis during turion development [33]. But this inhibitory effect of ABA during turion development is selective for that the synthesis of certain turion specific proteins increases [33]. Indeed, the pattern of expression was consistent with a rate-limiting role for this protein in starch biosynthesis. Furthermore, the regulation of gene copies underwent divergence and probably sub-functionalization to permit metabolic differentiation.

In plants, ADP-glucose pyrophosphorylases consist of large and small subunits that share many amino acids due to the proposed origination from a common ancestral gene [35]. For example, APLs and APSs, which make up the heterotetrametic potato

enzyme, share significant sequence homology (53% identity and 73% similarity) [36]. Here we selected the large subunit for our analysis because we made the assumption that both are coordinately expressed and that the large subunit should suffice as a marker of the developmental switch between frond and turion stage of the life cycle. Furthermore, the current sequencing of the entire genome will provide an opportunity to locate the gene copies of the small subunit as well. The model structure of the large subunit confirms that N- and C-terminal regions of the SpAPLs are essential for the allosteric regulatory properties of the heterotetrameric enzyme AGPase (Figure. 7B) [18]. Even though APLs are considered as a catalytic-disabled subunit, the ability of binding effectors (3-PGA) and substrates (ATP) is likely to undergo a conformational transition similar to the APSs during its catalytic cycle [37].

Phylogenetic analysis showed that *SpAPL1* and *SpAPL2* descended from common ancestors of the plastidial form Type 1 and Type 4 of the grasses, respectively, while *SpAPL3* shares the same branch with the ancestor of cytosolic Type 2 and plastidial Type 3 of grasses (Fig. 6) [12]. Studies suggests that cytosolic Type 2 in grass evolved from a duplication of an ancestral gene encoding a Type 3 plastidial *APL* by loss-of-function of the transit peptide cleavage site [16]. A similar process might have taken place in *Spirodela*, where *SpAPL1* does not exhibit a clear transit peptide. Interestingly, the opposite seems to be true for *SpAPL3*, which clusters with cytosolic Type 2 *APL*s, but encodes a transit peptide. Based on this, we classify it as a plastidial Type 3 *APL* of the

grasses. The phylogenetic relationship will become clearer when we know whether these copies are clustered or dispersed in the *Spirodela* genome. Interestingly, there is differential invasion of MITES in the introns of these genes with the most pronounced invasion in the *SpAPL1* gene (Table 6.1). This is reminiscent of the grasses, where one of the smallest genomes, rice, had a relative high percentage of MITEs (13.3% of all repeat elements compared to 0.4% in maize), but low retrotransposon content (59.5% compare to 92.7% in maize). *Spirodela polyrhiza* was namely chosen for sequencing because of its small genome size. Given the genome size variation among *Lemnoideae*, perhaps a similar relationship of genome size and MITEs exists among *Lemnoideae* as has been found in grass species [38].

In summary, turions of *S. polyrhiza* contain high starch content, small size of starch granules, and low lignin proportion, which provides a solid foundation for developing them as an alternative biofuel source. For further investigation of the role of *SpAPL2* and *SpAPL3* genes in starch synthesis, studies using transgenic plants will be needed.

6.5 Materials and Methods

Plant material and growth conditions

For our studies we chose *S. polyrhiza* (Sp) 7498 because this will serve as a reference genome for the *Lemnoideae*. One cluster of 3-5 fronds was aseptically

transplanted into half-strength Schenk and Hildebrandt basal salt mixture (Sigma, S6765) with 1% sucrose liquid medium at pH 5.8. The cultures were kept in chamber maintained at 100 $\mu mol.m^{-2}.s^{-1}$ and 23 °C through a 16h-light, 8h-dark photoperiod. After a couple of days' growth, 1μM abscisic acid (ABA, Sigma, A1049) was added.

Microscopic analysis of frond and turion

Vegetative fronds without ABA treatment and turions with 14 days ABA treatment were fixed, embedded, and dehydrated as described [39]. Samples were fixed in 5% glutaraldehyde in 0.1 M sodium cacodylate buffer, pH 7.4, containing 2% Suc in a 2-ml tube at 4 °C overnight and another 3 h at room temperature. Rinsed by 0.1 M sodium cacodylate buffer, they were postfixed in buffered 1% osmium tetroxide at 4 °C overnight followed by dehydration in a graded series of acetone washings. The dehydrated samples were then embedded in epon resin. The 1-mm-thick sections were picked up on a glass slide, stained with methylene blue and scoped with a light microscope. For transmission electron microscopy (TEM), 90-nm-thin sections were cut on a Leica EM UC6 ultramicrotome, stained with a saturated solution of uranyl acetate and lead citrate and scoped at 80 kV with a Philips CM 12 transmission electron microscope.

Determination of starch content of developing turions

One hundred milligrams of fresh sample tissues were taken from a time course of

0 (no ABA), 1, 2, 3, 5, 7, 10, 14 days of ABA treatment and flash frozen in liquid nitrogen. Before 7 days, the whole plants including both mother and daughter fronds were collected. After 7 days, the developed turions were separated from mother fronds and collected, when they sunk to the bottom of flask. Three biological replicates were done for each time point. The quantification of starch content was determined colorimetrically following manufacturer's protocols of a "total starch assay" procedure (amyloglucosidase/α-amylase method) (Megazyme, K-TSTA). We used water as a blank control and D-glucose as a standard. Dry weight was counted by 500 mg fresh tissue after incubation in 65 °C chamber for 24 hours.

Genomic DNA and total RNA isolation

Total genomic DNA was extracted from whole plant tissue by the CTAB method [40]. Considering that only daughter fronds shorter than 0.7 mm in length respond to ABA treatment and undergo turion formation after ABA treatment [11], developing turions only with specific sizes were collected at their developmental stages after 0 (no ABA addition), 1, 2, 3, 5, 7 days of ABA treatment, respectively, for quantification of *APL* gene expression. For each time point we used again three biological replicates. High-quality total RNA was extracted with RNeasy Plant Mini Kit (Qiagen, 74904). The on-column DNase I was used to remove contaminating genomic DNA (Qiagen, 79254). The RNA quality and quantity were confirmed by analysis with Nanodrop 1000 (Nanodrop Technologies, Wilmington, DE). First-strand cDNA synthesis of all samples

was generated by kit of SuperScript™ III First-Strand Synthesis System for RT-PCR (Invitrogen, 18080) using oligo-dT as primer.

Retrieval of *APL* genes and CDS sequence

The conserved domains of APL proteins of *Arabidopsis* were used to set up degenerate primers. Degenerate PCR reactions were done with templates of cDNA extracted from samples of three days of ABA treatment. The program was: 35 cycles of 94°C 30s, 50°C 30s and 72°C 1 min. PCR products were cloned into the pGEM-T Easy Vector (Promega) and DNA fragment sequences were determined using the ABI 3730XL platform. Gene specific primers were designed based on the sequence of the cloned DNA to perform 5' and 3' RACE using the SMARTer^{TM} RACE cDNA Amplification Kit (Clontech, 634923). The RACE-ready cDNA was also generated from total RNA of samples treated three days with ABA. RACE reactions were performed under the following program: 5 cycles of 94°C 30s and 72°C 2 min; 5 cycles of 94°C 30s, 70°C 30s and 72°C 2 min; 25 cycles of 94°C 30s, 68°C 30s and 72°C 2 min. The RACE products were also cloned and sequenced. The full-length cDNA was confirmed with primers designed from 5' end of the 5' RACE sequence and the 3' end of the 3' RACE sequence. The same primers were used to amplify corresponding gene sequences using genomic DNA as template. Because of the size of the genes we used Expand Long Range dNTPack (Roche, #04829042001). The thermal cycling conditions were: 10 cycles of 94°C 15s, 55°C 30s and 68°C 9 min; 25 cycles of 94°C 15s, 55°C 30s and 68°C 9 min

with 10 more seconds for each cycle. Initially, primer sequences derived from APL cDNA were used to sequence genomic DNA. Subsequently, primers derived from genomic sequences were used in iterative rounds of sequencing until sufficient coverage was achieved. The sequences were assembled and analyzed with DNASTAR. MUST system, which tested the existence of a pair of terminal inverted repeats (TIRs) and a pair of direct repeats (DRs) [41] was used to predict miniature inverted-repeat transposable elements (MITEs) in *APL* introns..

Phylogenetic studies

An unrooted maximum likelihood phylogenetic tree was determined by using the MEGA 5 program [42] based on the amino acid sequence alignments under the WAG model with 1000 bootstrap replications. The corresponding subunit sequences from rice and maize were downloaded from GenBank.

Modeling of the three-dimensional structures

Sequences of the APL regulatory sites from potato and *S. polyrhiza* were aligned using MEGA 5. Homology modeling studies were performed using the Swiss Model server (http://swissmodel.expasy.org/) [43] and structures were visualized and prepared by an open source program PyMOL (The PyMOL Molecular Graphics System, Version 0_99rc6, Schrödinger, LLC.). The sequence used was that of SpAPL1, SpAPL2 and SpAPL3. The chosen suitable template was homodimeric AGPase of potato (PDB 1yp3)

[17] for which X-ray structure information was available, showing more than 52% sequence homology with SpAPLs. Key proline (P44, P52 and P66) and lysine (K414 and K452) residues were numbered based on AGPase large subunit of potato (x61187) [18]. Only APL1 modeled structure was shown in Figure. 7B as representative for the sake of simplicity.

Expression analysis of *APL* genes

Alignment of full length of cDNAs produced unique regions at the 5' UTR to design primers for qPCR. qPCR was performed for 0, 1, 2, 3, 5, 7 day of ABA treatment. All cDNAs were made with 2 μg of RNA using the SuperScript® III First-Strand Synthesis System kit (Invitrogen, 18080-051). cDNAs were diluted 20-fold and Real-time PCR was performed by using the iQTM SYBR Green Supermix (Biorad, 170-8880) following the manufacturer's standard instructions. All qPCRs were performed in triplicates. The relative quantification of each gene expressional level was calculated by calibrating CT values normalized to a standard dilution series over all samples [44].

Acknowledgement

This work has been published in BMC Plant Biology in 2012 by authors of Wang W, Messing J with the title of "Analysis of ADP-glucose pyrophosphorylase expression during turion formation induced by abscisic acid in *Spirodela polyrhiza* (greater duckweed)".

6.6 References

1. Landolt E: **The family of Lemnaceae - a monographic study, Vols. 1**. In.: Veroffentlichungen des Geobotanischen Institutes der Eidgenossischen Technischen Hochschule, Stiftung Rubel; 1986.
2. SMART CC, TREWAVAS AJ: **Abscisic-acid-induced turion formation in *Spirodela polyrrhiza* L. II. Ultrastructure of the turion; a stereological analysis**. *Plant, Cell and Environment* 1983, **6**(6):515-522.
3. Gray KA, Zhao L, Emptage M: **Bioethanol**. *Curr Opin Chem Biol* 2006, **10**(2):141-146.
4. Wyman CE: **Potential Synergies and Challenges in Refining Cellulosic Biomass to Fuels, Chemicals, and Power**. *Biotechnology Progress* 2003, **19**(2):254-262.
5. Cheng JJ, Stomp A-M: **Growing Duckweed to Recover Nutrients from Wastewaters and for Production of Fuel Ethanol and Animal Feed**. *CLEAN - Soil, Air, Water* 2009, **37**(1):17-26.
6. Perry TO, Byrne OR: **Turion induction in Spirodela polyrrhiza by abscisic acid**. *Plant Physiol* 1969, **44**(5):784-785.
7. Reid MS, Bieleski RL: **Response of Spirodela oligorrhiza to Phosphorus Deficiency**. *Plant Physiol* 1970, **46**(4):609-613.
8. Smart CC, Fleming AJ, Chaloupkova K, Hanke DE: **The Physiological Role of Abscisic Acid in Eliciting Turion Morphogenesis**. *Plant Physiol* 1995, **108**(2):623-632.
9. Appenroth K, Teller S, Horn M: **Photophysiology of turion formation and germination in Spirodela polyrhiza**. *Biol Plant* 1996, **38**(1):95-106.
10. Appenroth K-J, Nickel G: **Turion formation in Spirodela polyrhiza: The environmental signals that induce the developmental process in nature**. *Physiol Plant* 2010, **138**(3):312-320.
11. SMART CC, TREWAVAS AJ: **Abscisic-acid-induced turion formation in *Spirodela polyrrhiza* L. I. Production and development of the turion**. *Plant, Cell and Environment* 1983, **6**(6):507-514.
12. Georgelis N, Braun EL, Shaw JR, Hannah LC: **The two AGPase subunits evolve at different rates in angiosperms, yet they are equally sensitive to activity-altering amino acid changes when expressed in Bacteria**. *Plant Cell* 2007, **19**(5):1458-1472.
13. James MG, Denyer K, Myers AM: **Starch synthesis in the cereal endosperm**. *Curr Opin Plant Biol* 2003, **6**(3):215-222.
14. Emes MJ, Bowsher CG, Hedley C, Burrell MM, Scrase-Field ES, Tetlow IJ: **Starch synthesis and carbon partitioning in developing endosperm**. *J Exp Bot* 2003, **54**(382):569-575.

15. Ouyang S, Buell CR: **The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants**. *Nucleic Acids Res* 2004, **32**(Database issue):D360-363.
16. Comparot-Moss S, Denyer K: **The evolution of the starch biosynthetic pathway in cereals and other grasses**. *J Exp Bot* 2009, **60**(9):2481-2492.
17. Jin X, Ballicora MA, Preiss J, Geiger JH: **Crystal structure of potato tuber ADP-glucose pyrophosphorylase**. *EMBO J* 2005, **24**(4):694-704.
18. Hwang S-K, Hamada S, Okita TW: **Catalytic implications of the higher plant ADP-glucose pyrophosphorylase large subunit**. *Phytochemistry* 2007, **68**(4):464-477.
19. Greene TW, Woodbury RL, Okita TW: **Aspartic acid 413 is important for the normal allosteric functioning of ADP-glucose pyrophosphorylase**. *Plant Physiol* 1996, **112**(3):1315-1320.
20. Henssen A: **Die Dauerorgane von Spirodela polyrrhiza (L.) Schleid. in physiologischer Betrachtung.** *Flora* 1954, **141**:523-566.
21. Ley S, Dolger K, Appenroth KJ: **Carbohydrate metabolism as a possible physiological modulator of dormancy in turions of Spirodela polyrhiza (L.) Schleiden**. *Plant Science* 1997, **129**:1-7.
22. Appenroth KJ, Ziegler P: **Light-induced degradation of storage starch in turions of Spirodela polyrhiza depends on nitrate**. *Plant Cell Environ* 2008, **31**(10):1460-1469.
23. Pankey RD, Draudt HN, Desrosier NW: **Characterization of the Starch of Spirodela polyrrhiza**. *Journal of Food Science* 1965, **30**(4):627-631.
24. Landolt E, Riklef K: **The family of Lemnaceae - a monographic study, Vols. 2**. In.: Veroffentlichungen des Geobotanischen Institutes der Eidgenossischen Technischen Hochschule, Stiftung Rubel; 1987.
25. Fujita M, Mori K, Kodera T: **Nutrient removal and starch production through cultivation of Wolffia arrhiza**. *Journal of Bioscience and Bioengineering* 1999, **87**(2):194-198.
26. Lin Y, Tanaka S: **Ethanol fermentation from biomass resources: current state and prospects**. *Applied Microbiology and Biotechnology* 2006, **69**(6):627-642.
27. Gupta M, Bawa AS, Semwal AD: **Morphological, Thermal, Pasting, and Rheological Properties of Barley Starch and Their Blends**. *International Journal of Food Properties* 2009, **12**(3):587 - 604.
28. Appenroth K-J, Keresztes A, Krzysztofowicz E, Gabrys H: **Light-induced degradation of starch granules in turions of Spirodela polyrhiza studied by electron microscopy**. *Plant Cell Physiol* 2011, **52**(2):384-391.
29. Franco CML, Ciacco CF, Tavares DQ: **The Structure of Waxy Corn Starch: Effect of Granule Size**. *Starch* 1998, **50**(5):193-198.
30. Smith A, Quinton-Tulloch J, Denyer K: **Characteristics of plastids responsible for starch synthesis in developing pea embryos**. *Planta* 1990, **180**(4):517-523.

31. Burgess D, Penton A, Dunsmuir P, Dooner H: **Molecular cloning and characterization of ADP-glucose pyrophosphorylase cDNA clones isolated from pea cotyledons**. *Plant Mol Biol* 1997, **33**(3):431-444.
32. Ohdan T, Francisco PB, Jr., Sawada T, Hirose T, Terao T, Satoh H, Nakamura Y: **Expression profiling of genes involved in starch synthesis in sink and source organs of rice**. *J Exp Bot* 2005, **56**(422):3229-3244.
33. SMART CC, TREWAVAS AJ: **Abscisic-acid-induced turion formation in *Spirodela polyrrhiza* L III. Specific changes in protein synthesis and translatable RNA during turion development**. *Plant, Cell and Environment* 1984, **7**(2):121-132.
34. Ruuska SA, Girke T, Benning C, Ohlrogge JB: **Contrapuntal networks of gene expression during Arabidopsis seed filling**. *Plant Cell* 2002, **14**(6):1191-1206.
35. Bhave MR, Lawrence S, Barton C, Hannah LC: **Identification and molecular characterization of shrunken-2 cDNA clones of maize**. *Plant Cell* 1990, **2**(6):581-588.
36. Smith-White BJ, Preiss J: **Comparison of proteins of ADP-glucose pyrophosphorylase from diverse sources**. *J Mol Evol* 1992, **34**(5):449-464.
37. Figueroa CM, Esper MC, Bertolo A, Demonte AM, Aleanzi M, Iglesias AA, Ballicora MA: **Understanding the allosteric trigger for the fructose-1,6-bisphosphate regulation of the ADP-glucose pyrophosphorylase from Escherichia coli**. *Biochimie* 2011, **In Press, Corrected Proof**.
38. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A *et al*: **The Sorghum bicolor genome and the diversification of grasses**. *Nature* 2009, **457**(7229):551-556.
39. Wu Y, Messing J: **RNA Interference-Mediated Change in Protein Body Morphology and Seed Opacity through Loss of Different Zein Proteins**. *Plant Physiol* 2010, **153**(1):337-347.
40. Murray MG, Thompson WF: **Rapid isolation of high molecular weight plant DNA**. *Nucleic Acids Res* 1980, **8**(19):4321-4325.
41. Chen Y, Zhou F, Li G, Xu Y: **MUST: A system for identification of miniature inverted-repeat transposable elements and applications to Anabaena variabilis and Haloquadratum walsbyi**. *Gene* 2009, **436**(1-2):1-7.
42. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance and maximum parsimony methods**. *Mol Biol Evol* 2011.
43. Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T: **Protein structure homology modeling using SWISS-MODEL workspace**. *Nat Protocols* 2008, **4**(1):1-13.
44. Zdepski A, Wang W, Priest HD, Ali F, Alam M, Mockler TC, Michael TP: **Conserved daily transcriptional programs in Carica papaya**. *Trop Plant Biol* 2008, **1**(3-4):236-245.

CHAPTER 7 EXPRESSION PROFILING WITH ONSET OF DORMANCY

7.1 Abstract

Higher plants exhibit a remarkable phenotypic plasticity to adapt to adverse environmental changes. The Greater Duckweed Spirodela, as an aquatic plant, presents exceptional tolerance to cold winters through its dormant structure of turions in place of seeds. Abundant starch in turions permits them to sink and escape the freezing surface of waters. Due to their clonal propagation, they are the fastest growing biomass on earth, providing yet an untapped source for industrial applications.

We used next generation sequencing technology to examine the transcriptome of turion development triggered by exogenous ABA. A total of 208 genes showed more than a 4-fold increase compared with 154 down-regulated genes in developing turions. The analysis of up-regulated differential expressed genes in response to dormancy exposed an enriched interplay among various pathways: signal transduction, seed dehydration, carbohydrate and secondary metabolism, and senescence. On the other side, the genes responsible for rapid growth and biomass accumulation through DNA assembly, protein synthesis and carbon fixation are repressed. Noticeably, three members of late embryogenesis abundant protein family are exclusively expressed during turion formation. High expression level of key genes in starch synthesis are APS1, APL3 and GBSSI, which could artificially be reduced for re-directing carbon flow from photosynthesis to create a higher energy biomass.

The identification and functional annotation of differentially expressed genes open a major step towards understanding the molecular network underlying vegetative

frond dormancy. Moreover, genes have been identified that could be engineered in duckweeds for practical applications easing agricultural production of food crops.

7.2 Introduction

Plants, unlike animals, do not have a fur or can seek shelter to survive under food shortage and cold weather. Consequently, they adapt to dormancy to avoid adverse environments, such as poor nutrition, chilling temperature and drought. Dormancy is a complex state of plant development, in which the plant body exhibits little or no growth. They recover their growth once the conditions are favorable.

There are mainly two types of plant dormancy by forming seeds or buds. Seed dormancy has been observed for many plants species including our major crops [1-3]. Winter dormant buds are found for instance in woody plants, bulbs, rhizomes and tubers of herbaceous plants [5]. Studies on the molecular mechanisms of bud dormancy transitions in perennial woody plants have been conducted, including pear [4], oak [6], and poplar [7].

Spirodela polyrhiza, a floating aquatic monocot, develops a specific dormant organ called turion during its life cycle, which alternates between periods of clonal propagation and dormancy. Its leaf, stem and bud are extremely compact in form of a round-shaped frond, resembling a single leaf. Large numbers of Spirodela plants can be maintained like cell cultures under totally controlled medium and environmental conditions. They reproduce vegetatively through budding of fronds (growth phase) during spring and summer [8] and transition to turions (dormant phase), when there is shortage of nutrition in the fall or the temperature drops in the winter [9]. Noticeably, fronds perform photosynthesis and turions function as storage for starch and germinate in

the following spring [10-13]. Turion cells exhibit dense intercellular space, thick cell wall and are also rich in anthocyanins [14]. Therefore, turions provide a unique system to study both bud and seed dormancy because they reproduce like buds without sexual hybridization but functionally equivalent to seeds that could generate a progeny plant in the growing season. Previous studies have shown that addition of ABA into growth medium quickly leads to turion formation after 5 days of treatment in the laboratory [13, 16, 17]. Only 3 days after ABA treatment, the Spirodela primordium is irreversibly committed to turion development [16]. The ease of growth and its direct contact with water make Spirodela a model system to gain molecular insights into plant dormancy [18].

At the molecular level, some studies on turion development have already been performed. For example, the transcript level of D-myo-inositol-3-phosphate synthase is rapidly induced within 15 min of ABA application, an enzyme that plays a key role in the inositol metabolism of the cell wall [19, 20]. The expression of the key enzyme ADP-glucose pyrophosphorylase (APL) for starch production [13] is significantly changed during turion formation. Still, not much information is known about the global transcriptome profiling for turion formation in this model system. To further uncover the regulation of gene expression as the phase switches, we took advantage of RNA deep sequencing, and compared the transcriptome between fronds and developing turions. A more comprehensive understanding of the gene repertoire and its regulation during turion formation has also great potential for industrial applications including the redirection of carbon flow into higher energy products.

7.3 Results and discussion

Calibration and selection of tissue samples

A comprehensive study for turion formation has been done using abscisic acid (ABA) induction [14, 16, 18, 21, 22]. Three days after ABA induction, the Spirodela primordium is committed to turion development, which cannot be reversed. All primary biosynthesis of protein, mRNA and DNA are shut down resulting in the onset of the dormant state [21]. To calibrate our growing conditions with previous investigations, we used transmission electron microscopy (TEM) to investigate different developmental stages. We chose fronds and developing turions with 3 days after ABA treatment instead of 14 days because 14-day treatment is not a key transition state and RNA purification is greatly interfered by high content of starch, but mature turions with 14-days treatment provide a more complete structural image through TEM. Turion cells have thicker cell walls, multiple smaller vacuoles and distorted plastids filled with abundant starch granules, whereas frond cells differ with having well-shaped chloroplasts consistent with previous observations (Figure 7.1). Therefore, growing conditions and turion induction appear to be reproducible.

Mapping RNA-Seq reads

We used eight samples in total, with each condition having four biological replicates. To eliminate potentially technical variation from biological replicates, they were multiplexed, pooled, and sequenced with the SOLiD 5500 platform. A total of 15~41 million quality reads per sample were generated after filtering raw reads (Table 7.1).

The high quality reads were mapped to chloroplast [23], mitochondria [24], and nuclear genomes [25], respectively. We could clearly divide sequence reads into these three classes. Surprisingly, there was an abundance of organelle-derived transcripts with 28~39% of total reads. With this depth of data we could assemble sequences for complete plastid and mitochondrial transcriptomes. The high proportion of organelle reads stresses the important roles of their transcripts, provides us with their expression profiles and facilitates the phylogenetic analysis [26]. Based on the combined reads of nuclear and organelle RNAs, more than 89% of our RNA-Seq reads were mappable, attesting to the performance of the sequencing platform. It also suggests that part of previously unmapped reads in other studies remained undetected because of their organellar origin [4, 27-29]. We still found that 1~9% of total reads were derived from ribosomal RNA, which is an indication that the protocol for the depletion of ribosomal RNA from samples was reasonably successful. Such efficiency is critical for mainly uncovering the desired transcriptome with complete coverage and in a cost-effective manner [30].

Among the total reads, 53-61% originated from nuclear DNA, lower than in other cases with about 80% of mappable sequences [27, 29]. The reason could be the method we used through ribosomal RNA removal rather than polyA selection. In case of polyA selection, organelle transcripts are automatically removed due to the lack of the polyA tail in organelle transcripts, whereas most of them were captured by our method of ribosomal RNA removal. Excluding the abundant organelle and rDNA reads, nuclear reads corresponded to 29~72X coverage for all annotated genes (Table 7.1), exhibiting the depth used in our study was sufficient to cover the Spirodela nuclear transcriptome.

Sample	Qualified total reads	Reads # map nuclear genome	Map nuclear genome	Nuclear coverage	Map chloroplast	Map mitochondria	Map rDNA
fronds 1	24,356,014	12,795,916	53%	42	35%	1%	4%
fronds 2	41,310,111	22,039,845	53%	72	37%	3%	4%
fronds 3	28,333,911	16,444,539	58%	54	29%	2%	6%
fronds 4	28,188,669	16,282,775	58%	53	30%	2%	9%
turions 1	26,484,522	15,431,023	58%	50	28%	2%	1%
turions 2	28,466,211	16,123,639	57%	53	34%	2%	2%
turions 3	25,754,050	15,697,393	61%	51	26%	2%	3%
turions 4	14,996,833	8,824,987	59%	29	29%	2%	1%

Table 7.1 Summary of sequence read alignments to three genome references.

Identification and validation of differentially expressed genes

Comparison of frond and developing turion samples provided us with 362 differentially expressed (DE) genes. A total of 117 had greater than 10-fold difference in mRNA levels and 208 genes were up-regulated and expressed at higher levels in developing turions than in fronds, whereas 154 genes were down-regulated, indicating lower expression in turions than in fronds (Table 7.2).

Fold change	4.0-5.0	5.1-10	10.1-15	15.1-20	>20	Sum
Genes expressed lower in turions than fronds	37	73	12	10	22	154
Genes expressed higher in turions than fronds	38	97	25	15	33	208

Table 7.2 Fold change in differentially expressed genes.

Previous studies had indicated that a small number of biological replicates might not be robust enough because it is impossible to know whether expression patterns are specific to individuals or are characteristic for the total population. Even for RNA deep

sequencing, a sufficient number of biological replicates are still required to have confidence in measurements [31-33]. Because two biological replicates usually are not sufficient to account for sample variability, we increased this number to four independent biological replicates. The coefficient of variation to the power of two (CV^2), a normalized measure of cross-replicate variability that can be useful for evaluating the quality of RNA-Seq data, was calculated to exhibit the biological variation (Figure 7.2). As expected, the data showed that the abundance of the genes varied between replicate RNA samples, especially for ones with lower FPKMs. However, with four biological replicates, which take the target population variation into account and also counteract random technical variation [34, 35], we were very confident to assess gene expression levels with accuracy.

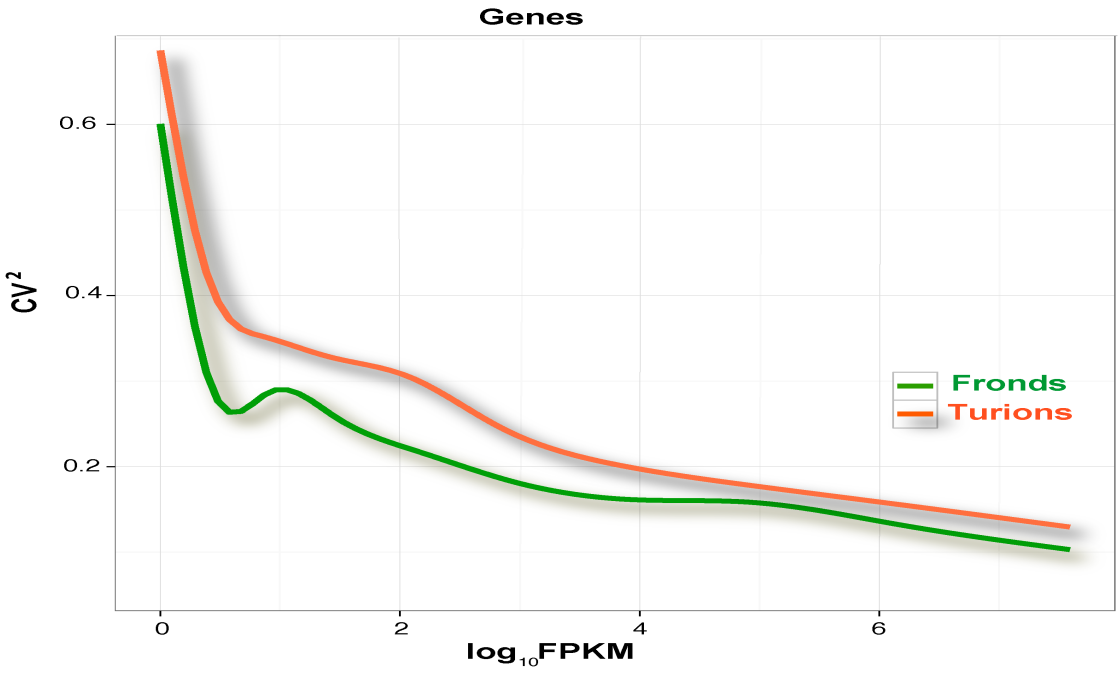


Figure 7.1 Biological variation for biological replicates.

Biological variation was represented by the square coefficient of variation of FPKM values for each gene (CV^2).

As another quality control, we could rely on our measurements of the 3 transcripts of ADP-glucose pyrophosphorylases (APLs) for starch synthesis [13], which were done with qRT-PCR, and compared with the RNA-Seq data. Indeed, the correlation co-efficient of 0.992 indicated that the two independent measurements were consistent and showed similar patterns: APL1 (GenBank Acession #JN180634) was highly expressed in fronds and APL3 (GenBank Acession #JN180636) showed the most abundance in developing turions. However, APL2 (GenBank Acession #JN180635) was not identified as DE gene due to only 1.5 times of difference at the time point of 0 and 3rd day by the threshold value of 4 (Figure 7.3). A fourth gene, tur4, provided us with a negative control from an independent study [36]. The tur4 gene has the Gene ID Spipo7G0013500 in the sequenced genome of Spirodela. Expression of this gene during turion formation was studied with Northern blot analysis. Although the tur4 gene responded to ABA treatment within hours, it appeared to return to nearly normal levels of expression thereafter. Northern blot analysis showed no induction at day 3 after ABA treatment, whereas we could still detect a 2-fold increase in tur4 expression with RNA-Seq, indicating that our method is more sensitive than Northern blot analysis. However, given both the APLs and tur4 results, we selected a cut-off for DE genes at 4-fold expressional change.

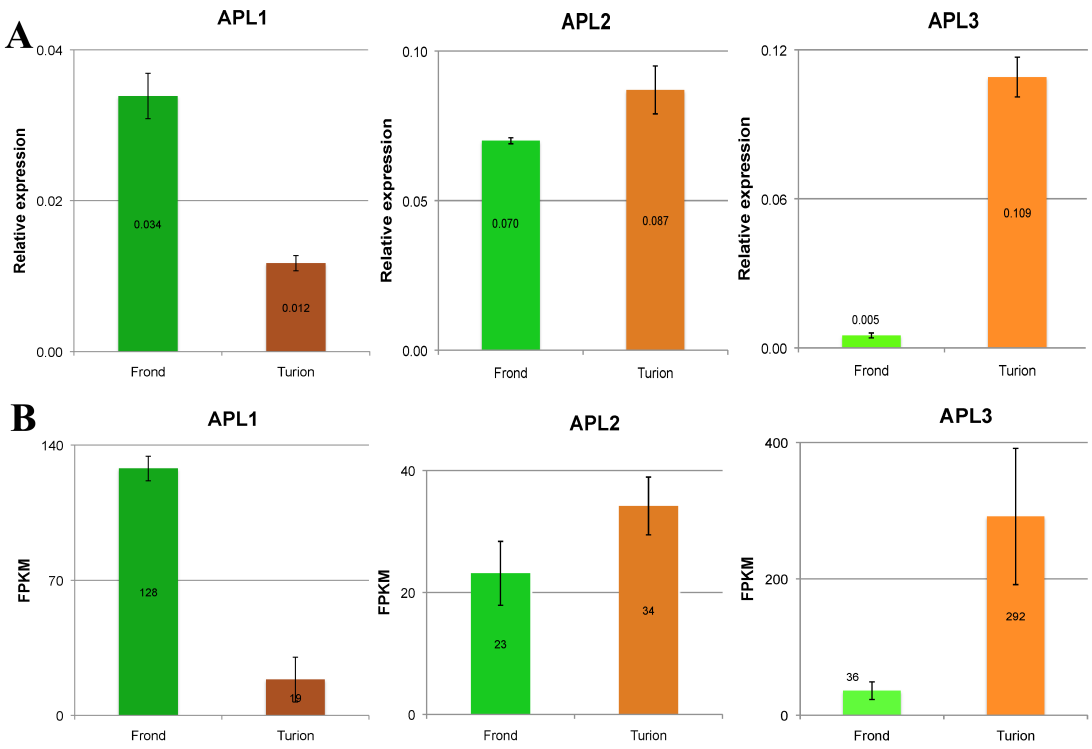


Figure 7.2 Comparison of RNA-Seq vs. qRT-PCR.

A. APL gene expression from qRT-PCR; B. APL gene expression from RNA-Seq data.

Response to ABA stimulus

The plant hormone abscisic acid (ABA) plays a major role as a signal in seed development and plant dormancy [37, 38] and regulates many important aspects, such as the synthesis of seed storage proteins, starch and lipids [39, 40]. In Spirodela, the exogenous ABA could easily trigger the dormant state (turions) from growth phase (fronds) [16]. We found 25 up-regulated DE genes in response to ABA stimulus or regulation based on their GO annotation (Table 7.3 and S1). The pathway of ABA signal transduction and response seemed to be interwoven with enzyme metabolism (kinase, synthase, and phosphatase) and other signaling pathways (transporter, ethylene).

Northern blot analysis shows that ABA rapidly up-regulates tur4 transcriptional level that encodes a peroxidase, which could stimulate turion formation and growth inhibition [36].

Gene ID	Fold change	Frond FPKM	Turion FPKM	Annotation
Spipo6G0001100	146	0.3	45.3	Peripheral-type benzodiazepine receptor
Spipo5G0029200	57	0.8	48.1	Major facilitator superfamily protein
Spipo19G0014500	43	0.5	22.4	Galactinol synthase
Spipo26G0007700	17	8.4	140.0	Late embryogenesis abundant protein LEA
Spipo8G0058900	16	1.2	19.4	Flowering locus T/Terminal flower 1-like protein
Spipo4G0016300	15	15.2	235.2	Annexin
Spipo18G0029800	15	28.7	420.3	O-acetyltransferase-like
Spipo3G0078900	14	0.4	6.2	Stachyose synthase, putative
Spipo0G0155100	13	1.7	21.5	Ethylene-responsive transcription factor 1
Spipo0G0130700	9	1.6	15.5	C4-dicarboxylate transporter
Spipo7G0041900	7	1.4	9.8	ABC transporter G family member
Spipo3G0031800	7	10.3	74.2	Ethylene-responsive transcription factor 2
Spipo8G0062500	6	7.2	44.3	Receptor-like protein kinase
Spipo14G0026800	5	4.0	18.3	Eukaryotic aspartyl protease family protein
Spipo12G0003900	4	37.9	162.9	myb domain protein 73
Spipo5G0040500	8	23.1	189.7	Alpha-dioxygenase
Spipo0G0156500	6	15.5	97.9	Alpha-dioxygenase
Spipo0G0180000	6	98.0	561.3	Alpha-dioxygenase
Spipo0G0156600	5	19.5	104.4	Prostaglandin G/H synthase
Spipo8G0046200	67	0.2	15.6	Protein phosphatase 2c, putative
Spipo3G0013100	38	0.5	20.2	NAC domain-containing protein 67
Spipo23G0012800	32	1.1	34.2	Protein phosphatase 2C
Spipo21G0022300	10	6.1	59.1	Protein phosphatase 2c, putative
Spipo1G0021700	6	3.2	18.9	Protein phosphatase 2c, putative
Spipo6G0056800	4	11.5	46.5	NAC domain-containing protein 67

Table 7.3 FPKM for Up-regulated DE genes in response to ABA stimulus.

Growth inhibition

Dormancy is generally defined by the lack of visible growth. The shoot apices cease active growth in perennial plants when a state of dormancy is reached. The seed

dormancy is observed in seeds with a quiescent phase preventing germination. The same phenomenon was investigated for Spirodela in the presence and absence of growth. When we looked at DE genes associated with Spirodela growth by RNA-Seq data, we found genes of histone H3 (Spipo9G0039400, Spipo0G0046100 and Spipo13G0007500) and H4 (Spipo28G0019000), ribosomal protein (Spipo1G0126300), expansins (Spipo22G0026300), aquaporins (Spipo11G0033800, Spipo17G0045100), ribulose-1,5-bisphosphate carboxylase oxygenases (RuBisCO) (Spipo19G0027700, Spipo23G0013400) for carbon fixation were down-regulated in turions (Table 7.4). In eukaryotic cells, DNA replication requires the synthesis of histone proteins to package newly replicated DNA into nucleosomes. Expansins are a key endogenous regulator of plant cell enlargement [41]. Aquaporins support cell growth and especially contributes to cell expansion and cell division. The gene that is highly expressed in fronds (69 times higher than in turions) is aquaporin (Spipo11G0033800) (Table 7.4). Over-expression of aquaporin stimulates cell growth in tobacco [42] or in Arabidopsis [43]. These results further confirm our knowledge that fronds are mainly responsible for rapid growth through actively DNA assembly, protein synthesis and carbon fixation, leading to a quick biomass increase, in comparison to the turions, where these processes are greatly decreased. Previous studies also suggested this mechanism of the turion formation by measuring DNA, RNA and protein content, which showed that DNA, protein and RNA biosynthesis were largely inhibited, resulting in the decrease of cell division, expansion and differentiation [21].

Gene ID	Fold change	Frond FPKM	Turion FPKM	Annotation
Spipo11G0033800	69	33.6	0.5	Aquaporin
Spipo17G0045100	5	86.3	17.8	Aquaporin
Spipo22G0026300	5	186.4	40.4	Expansin
Spipo9G0039400	7	68.2	9.4	Histone H3
Spipo0G0046100	7	112.4	16.4	Histone H3
Spipo13G0007500	6	159.5	27.5	Histone H3
Spipo28G0019000	5	77.9	14.8	Histone H4
Spipo3G0024800	14	1018.4	71.2	Pre-rRNA-processing protein PNO1
Spipo1G0126300	5	1371.4	293.1	60S ribosomal protein L10-like protein
Spipo19G0027700	29	6951.7	241.3	Ribulose bisphosphate carboxylase small chain
Spipo23G0013400	5	476.1	93.1	Ribulose-1 5-bisphosphate carboxylase/oxygenase activase

Table 7.4 FPKM for Down-regulated DE genes associated with growth.

Late embryogenesis abundant protein (LEA) genes are a valuable marker for dormancy

On the other hand, we found some specific mRNAs were increased in developing turions, for example LEAs. Although there were five members of LEA genes (Spipo14G0001200, Spipo5G0015500, Spipo0G0166800, Spipo1G0033500, Spipo26G0007700) with increased expression inturions, the LEA gene (Spipo0G0166800) was the most up-regulated DE gene, two other LEA genes (Spipo5G0015500 and Spipo14G0001200) were exclusively expressed in developing turions (Table 7.5). Indeed, the promoter of these LEA genes would be ideal to ensure expression of other coding regions exclusively in turions through transgenic approaches. Additionally, LEA was found to protect other proteins against desiccation, cold, and high salinity [44] and especially accumulates when plant seeds desiccate [45]. Given their high induction, they provide valuable markers for dormancy in general. In response to

dehydration, endogenous ABA levels increased dramatically followed by induction of LEA [46]. As expected, when Spirodela fronds are destined to dormant turions triggered by ABA, desiccation is an indispensable step, in which LEA proteins play pivotal roles to preserve the cellular structures and nutrients in turions.

Gene ID	Fold change	Frond FPKM	Turion FPKM	Directionality	Annotation
Spipo14G0001200	NA	0.0	31.0	up-regulated	Late embryogenesis abundant protein LEA
Spipo5G0015500	NA	0.0	45.8	up-regulated	Late embryogenesis abundant protein LEA
Spipo0G0166800	170	1.4	235.2	up-regulated	Late embryogenesis abundant protein LEA
Spipo1G0033500	34	3.4	114.8	up-regulated	Late embryogenesis abundant protein LEA
Spipo26G0007700	17	8.4	140.0	up-regulated	Late embryogenesis abundant protein LEA
Spipo4G0008600	5	6.1	33.1	up-regulated	bZIP transcription factor A
Spipo8G0037600	11	1.1	11.6	up-regulated	Heat shock transcription factor A2
Spipo9G0002000	5	14.2	67.8	up-regulated	Heat shock transcription factor A2
Spipo0G0155100	13	1.7	21.5	up-regulated	Ethylene-responsive transcription factor 1
Spipo3G0031800	7	10.3	74.2	up-regulated	Ethylene-responsive transcription factor 2
Spipo20G0027700	5	10.6	53.1	up-regulated	Ethylene-responsive transcription factor 3
Spipo11G0028200	7	32.7	4.4	down-regulated	Ethylene-responsive transcription factor 4
Spipo8G0045500	7	11.3	1.7	down-regulated	WRKY transcription factor, putative
Spipo2G0055800	4	17.1	4.0	down-regulated	bZIP transcription factor I

Table 7.5 FPKM for Turion-specific genes and DE transcriptional factors.

Genes involved in carbon partitioning

Starch is the major carbon reserve in plant storage organs, and ABA has a signaling role by inducing starch biosynthetic gene expression and co-ordinate carbohydrate partitioning [47]. In our study, four genes (Spipo12G0062400, Spipo18G0038500, Spipo16G0027000 and Spipo27G0011300) (Table 7.6) participating in starch biosynthesis were significantly enhanced in developing turions. The qRT-PCR experiment confirmed the key enzyme of large-subunit ADP-glucose pyrophosphorylase 3 (APL3) for starch biosynthesis was highly expressed in turion development [13]. The

RNA-Seq study for *Landoltia punctata* also revealed gene expression involved in starch biosynthesis was up-regulated under nutrient starvation [48]. Another way to accumulate starch content is to redirect carbon flow to starch biosynthesis. We found seven genes participate in the degradation of lipids by alpha- (Spipo0G0156600, Spipo0G0180000, Spipo0G0156500, Spipo5G0040500) or beta-oxidation (Spipo0G0179100, Spipo3G0031300, Spipo1G0110400), which probably allocate carbon to starch rather than fatty acids to achieve denser turions that sink to the bottom of streams during seasons (Table 7.6). Previously, it has been shown that the carbon flow into seeds can be rebalanced between different macromolecules with different energy content [49]. Reallocation of carbon is critical for the improvement of oil production in novel crops in the future. In oilseed species, numerous biotechnological approaches have been carried out that were aimed to maximize the flow of carbon into oil by over-expression of enzymes of the TAG assembling network [50]. Although one might argue that turions would no longer be able to sink in water when filled with lipids, in those applications biomass would be accumulated under constant temperature.

Another way to investigate the balance of carbon partitioning can be derived from the average FPKM value (Fragments Per Kilobase of transcript per Million mapped reads) of all the key genes encoding both pathways. The genes encoding for lipid production were expressed relatively low with FPKM of 28 and 22 in fronds and turions, respectively. Therefore, the level of lipids remains low throughout development (Table 7.6). Given the high level of starch in turions, genes in lipid production are not induced, whereas the ones for starch biosynthesis are during turion formation, providing us with a correlation between metabolic products and the regulation of the corresponding

pathways. Given this correlation, we hypothesize that we could redirect carbon flow into lipids by blocking key genes of such as AGPS1, AGPL3, GBSSI and ACCase4, GPAT1, DGAT2, and over-express transcripts of the lipid pathway (Table 7.6) together with turion-specific promoters, like LEAs (Spipo14G0001200, Spipo5G0015500, Spipo0G0166800) (Table 7.5).

Pathway	Gene ID	Enzyme	FPKM in frond	FPKM in turion	Average FPKM in frond	Average FPKM in turion
Lignin	Spipo0G0185100	CCR1	32.6	23.5	23	41
	Spipo11G0026200	CCR2	3.4	10.2		
	Spipo6G0037000	CCR3	45.3	35.5		
	Spipo28G0002300	CCR4	1.0	0.9		
	Spipo23G0040600	CCR5	6.7	16.7		
	Spipo11G0026400	CCR6	5.3	10.8		
	Spipo10G0016700	CCR7	0.3	0.6		
	Spipo10G0000200	CCR8	26.2	167.8		
	Spipo8G0071400	CCR9	20.0	24.2		
	Spipo5G0064600	CCR10	16.7	18.6		
	Spipo7G0010700	CCR11	10.1	64.8		
	Spipo0G0172200	CCR12	215.3	306.7		
	Spipo7G0010800	CCR13	0.4	1.4		
	Spipo14G0054900	CCR14	2.0	9.9		
	Spipo12G0004300	CAD1	18.1	32.5		
	Spipo17G0012300	CAD2	10.8	7.8		
	Spipo1G0069500	CAD3	2.6	0.7		
	Spipo2G0124600	CAD4	3.9	2.3		
Starch	Spipo28G0001400	APS1	264.1	242.5	70	86
	Spipo3G0049000	APL1	127.9	18.6		
	Spipo6G0024200	APL2	23.1	34.2		
	Spipo18G0038500	APL3	36.0	291.5		
	Spipo26G0026900	SSI	21.7	28.6		
	Spipo0G0050800	SSII	3.4	1.6		
	Spipo14G0048800	SSIII	33.4	24.0		
	Spipo14G0042000	SSIV	45.7	15.6		

	Spipo1G0057900	GBSSI	327.8	333.2		
	Spipo1G0057400	BEI	19.6	10.3		
	Spipo0G0008100	BEII	40.8	72.0		
	Spipo12G0062400	ISA1	2.8	14.6		
	Spipo3G0051400	ISA2	8.1	12.9		
	Spipo20G0022100	ISA3	25.1	98.0		
Lipid	Spipo0G0127900	ACCase1	24.0	17.4	28	22
	Spipo10G0023400	ACCase2	6.3	4.6		
	Spipo12G0034900	ACCase3	4.5	2.5		
	Spipo12G0063600	ACCase4	85.6	44.3		
	Spipo15G0009000	ACCase5	22.4	21.4		
	Spipo4G0043600	ACCase6	20.9	11.0		
	Spipo4G0047600	ACCase7	73.7	56.7		
	Spipo30G0006700	GPAT1	127.7	52.2		
	Spipo7G0013300	GPAT2	21.0	20.4		
	Spipo3G0111400	AGPAT1	1.4	0.3		
	Spipo4G0068200	AGPAT2	17.6	18.2		
	Spipo6G0030100	AGPAT3	15.8	13.5		
	Spipo7G0018900	AGPAT4	4.3	3.3		
	Spipo7G0051900	AGPAT5	1.6	0.9		
	Spipo21G0027500	DGAT1	6.0	5.7		
	Spipo28G0006400	DGAT2	70.9	117.6		
	Spipo1G0066600	DGAT3	0.9	1.5		
	Spipo20G0011900	DGAT4	11.2	8.2		
	Spipo3G0079500	DGAT5	14.6	23.1		

Table 7.6 Expression patterns for lignin, starch and lipid biosynthesis.

Turion-specific pathays

We found that the transcriptome also closely links the turion phenotypic variation with thick cell wall and abundant secondary metabolites like pigment. The expressions of eight members of the UDP-glycosyltransferase superfamily (Spipo2G0010600, Spipo2G0043800, Spipo16G0044000, Spipo2G0039000, Spipo14G0034300,

Spipo2G0124000, Spipo5G0014300, Spipo2G0077900) and two of the cellulose synthases (Spipo28G0017100, Spipo7G0044000) involved in cell call biosynthesis were increased. Three dihydroflavonol reductases (Spipo7G0010700, Spipo10G0000200, Spipo14G0054900) and one flavonoid 3', 5'-hydroxylase (Spipo0G0155000) involved in the anthocyanin pathway were up-regulated. In addition, we found the average FPKM value for all key enzymes of lignin biosynthesis were 23 in fronds but 41 in turions, which may explain the rigidity of cell wall in turion cells to defend water pressure at the bottom of waters (Table 7.6).

To gain a broad overview into the biological functions for DE genes, we next performed an analysis of gene ontology (GO) enrichment (Methods). We found a total of 24 enriched pathways ($p<0.01$) in developing turions, whereas no enriched GO was found in fronds under the null hypothesis of the entire gene set of Spirodela (Young et al., 2010). The clustered DE genes were mainly related to response to ABA, fatty acid oxidation, and ion transportation. The GO functions of leaf senescence and cell wall modification were also highlighted (Table 7.7).

Enriched GO ID	description
GO:0001561	fatty acid alpha-oxidation
GO:0033539	fatty acid beta-oxidation using acyl-CoA dehydrogenase
GO:0010167	response to nitrate
GO:0015706	nitrate transport
GO:0055114	oxidation-reduction process
GO:0009830	cell wall modification involved in abscission
GO:0009651	response to salt stress
GO:0010106	cellular response to iron ion starvation
GO:0010150	leaf senescence
GO:0009737	response to abscisic acid stimulus
GO:0006826	iron ion transport
GO:0001676	long-chain fatty acid metabolic process
GO:0001666	response to hypoxia
GO:0046487	glyoxylate metabolic process
GO:0071732	cellular response to nitric oxide
GO:0010286	heat acclimation
GO:0071446	cellular response to salicylic acid stimulus
GO:0072329	monocarboxylic acid catabolic process
GO:0019579	aldaric acid catabolic process
GO:0009751	response to salicylic acid stimulus
GO:0042542	response to hydrogen peroxide
GO:0046686	response to cadmium ion
GO:0009788	negative regulation of abscisic acid mediated signaling pathway
GO:0009414	response to water deprivation

Table 7.7 Functional GO enrichment in developing turions.

Transcriptional regulation of differentially expressed genes

Transcription factors (TFs) are crucial components of regulatory systems, which initiate vital changes in gene expression. Thus, we examined TF gene models and found nine TFs were significantly changed including two ABA-responsive element binding factors (bZIP, Spipo4G0008600 and Spipo2G0055800), four Ethylene-responsive

element binding factors (ERFs, Spipo0G0155100, Spipo3G0031800, Spipo20G0027700 and Spipo11G0028200), two heat shock TFs (HSFs, Spipo8G0037600 and Spipo9G0002000), and one WRKY TF (Spipo8G0045500) (Table 7.5 and S1).

ABA-responsive element binding factor

The bZIP trancription factors regulate plant development through a basic region and a leucine zipper dimerization motif that binds to DNA [51, 52]. In the complete sequence of Spirodela genome [25], an exhaustive search of the bZIP superfamily was performed and 41 members identified. Among them, seven genes belong to the ABA-responsive element binding factors (ABFs), i.e., the bZIP superfamily group A due to their structural features with conserved regions C1-C2, basic regions, and leucine zippers (Figure 7.4) [51, 53]. This group is thought to play a central role in controlling ABA-responsive gene expression in seeds and vegetative tissues via binding to ABA-responsive-elements (ABREs). For example, ABI5, one member of ABFs, induces LEA expression by binding to its promoters during seed maturation [53]. Here, all seven genes showed differentially increased expression levels, whereas only SpABF1 (Spipo4G0008600) was defined as a DE gene due to a significant change (Table 7.5). Noticeably, SpbZIP (Spipo2G0055800), another bZIP transcription factor, was significantly decreased in developing turions (Table 7.5). It shared leucine residues in the basic domain but missing other 2 conserved regions, corresponding to bZIP group I in Arabidopsis. Studies of group I genes from several species indicate that they might play a role in vascular development [51]. SpbZIP might positively regulate xylem and phloem development, too. Because both structure and function of turions are equivalent to seeds,

less vascular tissue is needed in turions compared to fronds and the expression of SpbZIP is decreased accordingly. Thus, we conclude that a specific subset of bZIP transcription factors are involved in turion formation.

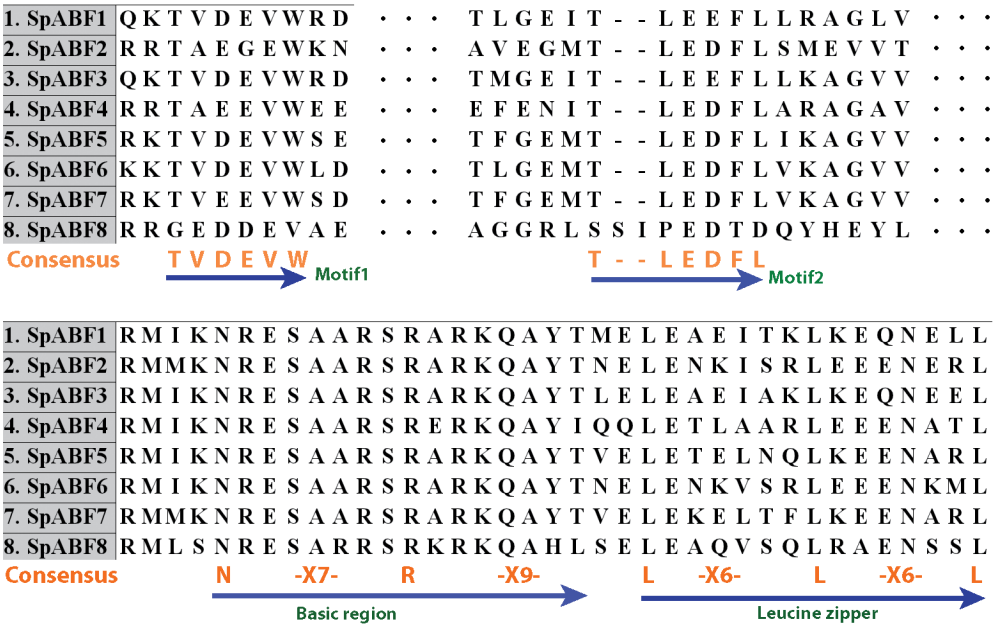


Figure 7.3 Alignment of ABF domain from Spirodela.

The amino acid sequences of bZIP protein sequence from Spirodela were aligned and the conserved regions were demonstrated here. The consensus amino acids were labeled from conserved region indicate motif 1, motif 2, primary structure of bZIP domains (basic region and leucine zipper). All members contain these four domains except SpbZIP, which only has basic region and leucine zipper. SpABF1-Spipo4G0008600; SpABF2-Spipo6G0055300; SpABF3-Spipo15G0021000; SpABF4-Spipo4G0111500; SpABF5-Spipo7G0034500; SpABF6-Spipo3G0017700; SpABF7-Spipo13G0002500; SpbZIP-Spipo2G0055800.

Other TFs involved in ABA-mediated gene expression

In addition to ABF TFs, other TFs were also identified to be involved in turion development. Ethylene-responsive element binding factors (ERFs) are transcription factors that are specific to plants. A highly conserved DNA binding domain, known as the ERF domain interacting directly with the GCC box in the ethylene-responsive-element (ERE), is the unique feature of this protein family [54] (Figure 7.5). ERFs also play a role in a variety of developmental processes such as flower, seed development [55], and fruit ripening [56]. We identified 57 ERF genes in the Spirodela genome, where SpERF1 (Spipo0G0155100), SpERF2 (Spipo3G0031800), and SpERF3 (Spipo20G0027700) were significantly up-regulated and SpERF4 (Spipo11G0028200) down-regulated in response to turion development (Table 7.5). It had been reported that AtERF1, AtERF2, ATERF5 functioned as activators of GCC box-dependent transcription in Arabidopsis leaves, but AtERF3 and AtERF4 acted as repressors [52, 54]. It also was shown that ERF2 and ERF4 enhanced the transcription of a reporter gene in tobacco protoplasts [57]. The three highly up-regulated ERFs in Spirodela turions should therefore play an important role in turion development.

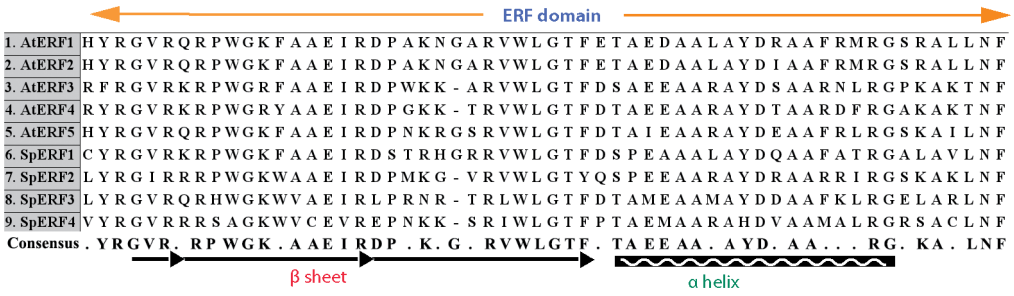


Figure 7.4 Alignment of the ERF domain from Arabidopsis and Spirodela.

The bar and black arrows indicate β sheet which interact with the GCC box in target DNA. The cross-hatched box indicates the α helix. The consensus amino acids are underlined in ERF domain. The accession numbers are: AtERF1-BAA32418; AtERF2-BAA32419; AtERF3-BAA32420; AtERF4-BAA32421; AtERF5-BAA32422; SpERF1-Spipo0G0155100; SpERF2-Spipo3G0031800; SpERF3-Spipo20G0027700; SpERF4-Spipo11G0028200.

Heat shock transcription factors (HSFs) are transcriptional activators of heat shock genes. An increasing number of studies indicated that some HSFs appeared during the maturation stage of the seed, when cell division ceased and seeds adapted to desiccation and long-term survival [58]. Here, the increased expression of two HSFs (Spipo9G0002000 and Spipo8G0037600) (Table 7.5) might also indicate an important function for turion desiccation and survival during long periods of winter.

WRKY transcription factors (TFs) are key regulators of many plant processes, including the responses to biotic and abiotic stresses, senescence, seed dormancy, and seed germination [59]. *In vivo* and *in vitro* promoter-binding studies showed that WRKY TFs could either activate or repress the expression of downstream ABFs through W-box sequences present in their promoters [60]. However, whether the Spirodela WRKY TF (Spipo8G0045500) (Table 7.5) is a repressor or activator needs to be further investigated.

Together, the significant changes in the expressions of ABFs, ERFs, HSF and WRKY TF reflected their obligatory regulation during turion development. Their involvement in the transition from fronds to turions and their control of spatial and

temporal expression of target genes provides us also with new tools to create specialized traits through tailoring of chimeric genes.

cis-element

Control of gene expression is achieved through the binding of transcription factors to specific *cis*-elements in promoter regions of target genes [61]. To predict potential pairs of TFs and *cis*-elements, we scanned a 1-kb region upstream of DE genes with the PLACE database [62]. We found 30 of up-regulated DE genes containing the *cis*-element of ABA-responsive element (ABRE: YACGTGGC) and 119 with ethylene-responsive element (ERE: GCCGCC). These target genes of ABFs and ERFs are associated with seed dehydration (like late embryogenesis abundant proteins), regulatory transcription factor, protein kinases and phosphatases (like CPK, MAPK), carbohydrate and secondary metabolism (like cellulose synthase and stachyose synthase), and senescence-associated proteins (like Glutathione-S-transferase).

Model of Spirodela dormancy

ABA is essential for seed maturation and also enforces a period of seed dormancy so that the seeds do not germinate prematurely during unseasonably conditions. The same behavior is seen in dormant Spirodela turions that are induced by low temperature, limited nutrition, or exogenous ABA [63]. The external stimuli rapidly induce both Ca^{2+} influx and endogenous ABA synthesis [64]. In maturing seed, ABA-regulated genes include those required for the synthesis of storage reserves and the acquisition of desiccation tolerance. Ca^{2+} can act as secondary messenger to activate the expression of

cascade components of calcium-dependent protein kinase (CPK) and mitogen-activated protein kinase (MAPK). The structure of CPK shows there are four Ca^{2+}-binding EF hand domains allowing the protein to function as a Ca^{2+} sensor. In addition to Ca^{2+}, reversible phosphorylation also regulates kinase activity [65]. A number of studies have demonstrated that MAPKs in Arabidopsis are associated with hormone biosynthesis and signaling including ethylene and ABA [38]. Both of CPK and MAPK could phosphorylate a wide range of target proteins, including other kinases and/or transcription factors [39, 52], in particular SpERF of Spipo0G0155100, Spipo3G0031800 and Spipo20G0027700, SpABF of Spipo4G0008600 and Spipo2G0055800, SpHSF of Spipo8G0037600 and Spipo9G0002000, and SpWRKY of Spipo8G0045500 (Table 7.5). The activation of TFs ultimately regulates their target genes to cease cell division but begin to accumulate secondary metabolites. As shown in flowering seeds, aspects of reserve accumulation and late embryogenesis abundant (LEA) gene expression are controlled largely by the coordinated action of transcription factors [39]. Taken together, we generated a model summarizing the signal transduction leading to Spirodela dormancy based on integration of our result and previous knowledge (Figure 7.5).

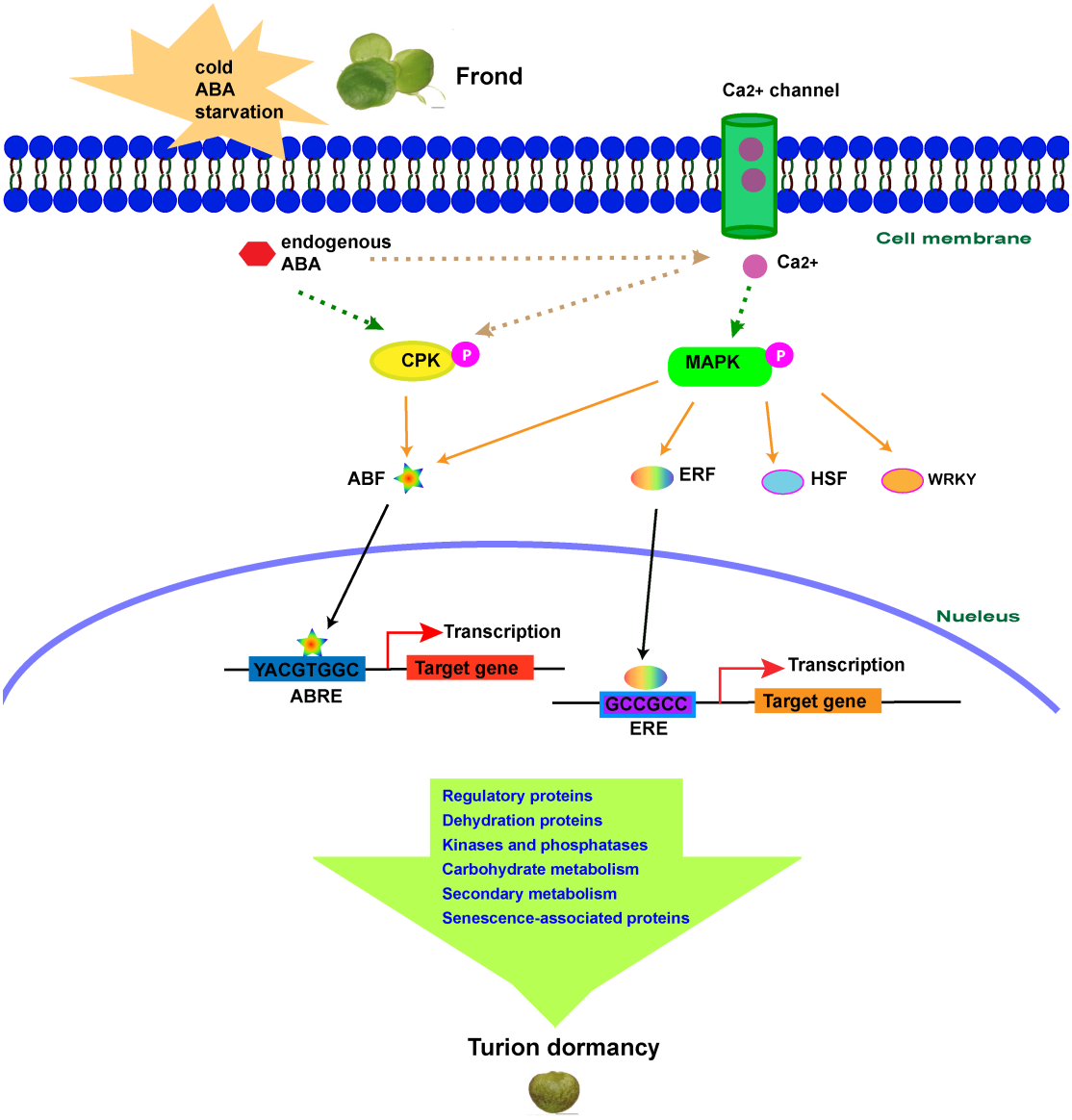


Figure 7.5 A model of development of Spirodela dormancy.

Phosphorylated proteins are labeled as pink circles with a P inside. Solid lines represent direct connections. The dotted line indicates indirect connection. Not all linkages and details of pathway are shown in this diagram in order to simplify the model. Abbreviations: ABA (abscisic acid), CPK (calcium-dependent protein kinase), MAPK (mitogen-activated protein kinase), ABF (ABA-responsive element binding factor), ERF (ethylene-responsive element binding factor), HSF (heat shock transcription factor), WRKY (WRKY transcription factors), AREB (ABA-responsive element), ERE (ethylene-responsive element).

7.4 Conclusions

Many studies have been concerned with seed development in plants. Seeds are the product of sexual reproduction and the segregation of Mendelian traits. They also represent a dormant state in the life cycle of the plant and they compartmentalize nutrients for growth in the absence of photosynthesis. Agriculture could not exist without these properties of plants. Here, we studied a plant that propagates by clonal division and can undergo dormancy without forming seeds. The aquatic plant Spirodela could not survive on water surface without human intervention, when the water freezes. It simply switches to dormancy and accumulates starch that allows it to sink to the bottom of the water to escape the ice. Besides low temperature, however, the same switch can be achieved with the hormone ABA that has been shown to perform the same change for seed maturation. Using such an induction with Spirodela, we can study genes that regulate dormancy. Here, we isolated total RNA exclude ribosomal RNA before and at the onset of dormancy, sequenced them with next-generation technology, and identified the transcripts by mapping them back to the genome sequence. The detailed analysis of the transcriptional landscape of differentially expressed genes provides the first comprehensive view at the dormancy of the aquatic plants. On the other hand, research studies have been initiated with the goal of developing duckweed species as an alternative to algae for oil production with the fact of fast growth and quick biomass accumulation [66]. The expression data for lipid and starch biosynthesis together with the turion-specific transcriptional genes from our RNA-Seq data would be the ideal targets to develop duckweeds into oil crops.

7.5 Materials and methods

Sample preparation

Spirodela polyrhiza 7498 was grown into half-strength Schenk and Hildebrandt basal salt mixture (Sigma, S6765) with 1% sucrose liquid medium under 6-hrs light, 8-hrs dark photoperiod. Plant tissues from four biological replicates for fronds without ABA treatment and developing turions with 3-day 10 μM ABA were collected and frozen in liquid nitrogen. 10 μg of total RNA was extracted for each sample by RNA-easy Qiagen kit with RLC buffer due to second metabolites. Ribosomal RNA was depleted with a kit from Epicenter (MRZPL116) in order to increase the coverage of other RNA classes. Vegetative fronds and turions with 14 days ABA treatment were fixed, embedded, and examined under transmission electron microscope as described [13, 67].

Library construction and sequence quality control

We started with ~300 ng rRNA-depleted total RNA, fragmented the RNA, performed reverse transcription and size-selected the cDNA, used Emulsion PCR to amplify the complex gene libraries and prevent formation of chimeric cDNA products. All steps followed the manufacturer's guide (SOLiD^{TM} total RNA-Seq kit). To minimize potential experimental batch effect, eight samples were barcoded, pooled, and evenly distributed into three lanes. The single-end reads with the size of 75bp were generated with our in-house SOLiD 5500 platform. The Exact Call Chemistry (ECC) module was utilized in the sequencing run, which is an optional kit that is used to further enhance sequencing accuracy by generating reference-free bases directly. After quality trimming with score of 20, reads with a minimum length of 40 bp were saved.

Read mapping and quantifying gene expression

The remaining reads were mapped to the reference genome *Spirodela polyrhiza* 7498 (GenBank Accession #ATDW01000000), which was recently sequenced, assembled, and annotated, by using TopHat 2 [68] with Bowtie [69]. TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to reference genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons. Gene expression levels were normalized using fragments per kilobase of exon per million mapped reads (FPKM). Transcript abundance and deferential gene expression were calculated with Cufflinks [35]. DE genes were defined, as when their absolute value of log2 fold change was higher than 2 and their P value was less than 0.01.

As a positive control for our measurements, we used independent data obtained in a separate study under the same induction conditions as in this study from the expression of ADP-glucose pyrophosphorylase genes with qRT-PCR [13]. As a negative control, we used northern blot data of the expression of the tur4 gene obtained in yet another study [36].

Functional annotation and *cis*-element predictions

For each DE gene, GO annotation was obtained with the program of blast2go, which uses a blast algorithm to assign GO terms to sequences based on similarity [70]. GO enrichment was performed in two groups of gene sets, respectively, one of highly expressed transcripts in turions, the other one of highly expressed transcripts in fronds

based on the whole gene set of the Spirodela genome using GOseq, which adjusts the bias from gene lengths [71]. The *cis*-acting regulatory DNA elements were predicted by signal scan search from PLACE database [62]. PLACE is a database of motifs found in plant *cis*-acting regulatory DNA elements, all from previously published reports. We dissected 1-kb regions upstream of DE genes and scanned them for potential pairs of TFs and cis-elements.

7.6 References

1. Jones SI, Vodkin LO: **Using RNA-Seq to profile soybean seed development from fertilization to maturity**. *PLoS One* 2013, **8**(3):e59270.
2. Bentsink L, Hanson J, Hanhart CJ, Blankestijn-de Vries H, Coltrane C, Keizer P, El-Lithy M, Alonso-Blanco C, de Andres MT, Reymond M *et al*: **Natural variation for seed dormancy in Arabidopsis is regulated by additive genetic and molecular pathways**. *Proc Natl Acad Sci U S A* 2010, **107**(9):4264-4269.
3. Liu A, Gao F, Kanno Y, Jordan MC, Kamiya Y, Seo M, Ayele BT: **Regulation of wheat seed dormancy by after-ripening is mediated by specific transcriptional switches that induce changes in seed hormone metabolism and signaling**. *PLoS One* 2013, **8**(2):e56570.
4. Liu G, Li W, Zheng P, Xu T, Chen L, Liu D, Hussain S, Teng Y: **Transcriptomic analysis of 'Suli' pear (Pyrus pyrifolia white pear group) buds during the dormancy by RNA-Seq**. *BMC Genomics* 2012, **13**:700.
5. Vegis A: **Dormancy in Higher Plants**. *Annual Review of Plant Physiology* 1964, **15**(1):185-224.
6. Ueno S, Klopp C, Leple JC, Derory J, Noirot C, Leger V, Prince E, Kremer A, Plomion C, Le Provost G: **Transcriptional profiling of bud dormancy induction and release in oak by next-generation sequencing**. *BMC Genomics* 2013, **14**:236.
7. Ruttink T, Arend M, Morreel K, Storme V, Rombauts S, Fromm J, Bhalerao RP, Boerjan W, Rohde A: **A molecular timetable for apical bud formation and dormancy induction in poplar**. *Plant Cell* 2007, **19**(8):2370-2390.
8. Landolt E: **The family of Lemnaceae - a monographic study, Vols. 1**: Veroffentlichungen des Geobotanischen Institutes der Eidgenossischen Technischen Hochschule, Stiftung Rubel; 1986.
9. Appenroth K-J, Nickel G: **Turion formation in Spirodela polyrhiza: The environmental signals that induce the developmental process in nature**. *Physiologia Plantarum* 2009, **138**(3):312-320.

10. Appenroth KJ, Teller S, Horn M: **Photophysiology of turion formation and germination in Spirodela polyrhiza**. *Biologia Plantarum* 1996, **38**(1):95-106.
11. Appenroth KJ, Ziegler P: **Light-induced degradation of storage starch in turions of Spirodela polyrhiza depends on nitrate**. *Plant Cell Environ* 2008, **31**(10):1460-1469.
12. Appenroth KJ, Keresztes A, Krzysztofowicz E, Gabrys H: **Light-induced degradation of starch granules in turions of Spirodela polyrhiza studied by electron microscopy**. *Plant Cell Physiol* 2011, **52**(2):384-391.
13. Wang W, Messing J: **Analysis of ADP-glucose pyrophosphorylase expression during turion formation induced by abscisic acid in Spirodela polyrhiza (greater duckweed)**. *BMC Plant Biol* 2012, **12**:5.
14. Smart CC, Trewavas AJ: **Abscisic-acid-induced turion formation in Spirodela polyrrhiza L. II. Ultrastructure of the turion; a stereological analysis**. *Plant, Cell and Environment* 1983, **6**(6):515-522.
15. Cernac A, Andre C, Hoffmann-Benning S, Benning C: **WRI1 is required for seed germination and seedling establishment**. *Plant Physiol* 2006, **141**(2):745-757.
16. Smart CC, Trewavas AJ: **Abscisic-acid-induced turion formation in Spirodela polyrrhiza L. I. Production and development of the turion**. *Plant, Cell and Environment* 1983, **6**(6):507-514.
17. Smart CC, Fleming AJ, Chaloupkova K, Hanke DE: **The physiological role of abscisic acid in eliciting turion morphogenesis**. *Plant Physiol* 1995, **108**(2):623-632.
18. **Plant dormancy: physiology, biochemistry and molecular biology**. In*: 1996; Wallingford*. CAB INTERNATIONAL: xx + 386 pp.
19. Smart CC, Fleming AJ: **A plant gene with homology to D-myo-inositol-3-phosphate synthase is rapidly and spatially up-regulated during an abscisic-acid-induced morphogenic response in Spirodela polyrrhiza**. *Plant J* 1993, **4**(2):279-293.
20. Flores S, Smart CC: **Abscisic acid-induced changes in inositol metabolism in Spirodela polyrrhiza**. *Planta* 2000, **211**(6):823-832.
21. Smart CC, Trewavas AJ: **Abscisic-acid-induced turion formation in Spirodela polyrrhiza L III. Specific changes in protein synthesis and translatable RNA during turion development**. *Plant, Cell and Environment* 1984, **7**(2):121-132.
22. Smart CC, Trewavas AJ: **Abscisic-acid-induced turion formation in Spirodela polyrrhiza L. IV. Comparative ion flux characteristics of the turion and the vegetative frond and the effect of ABA during early turion development**. *Plant, Cell and Environment* 1984, **7**(7):521-530.
23. Wang W, Messing J: **High-Throughput Sequencing of Three Lemnoideae (Duckweeds) Chloroplast Genomes from Total DNA**. *PLoS ONE* 2011, **6**(9):e24670.
24. Wang W, Wu Y, Messing J: **The mitochondrial genome of an aquatic plant, Spirodela polyrhiza**. *PLoS One* 2012, **7**(10):e46747.
25. Wang W, Haberer, G., Gundlach, H., Gläßer, C., Nussbaumer, T., Luo, M.-C., Lomsadze, A., Borodovsky, M., Kerstetter, R.A., Shanklin, J., Byrant D., Mockler, T., Appenroth, K.J., Grimwood, J., Jenkins, J., Chow, J., Choi, C.,

Adam, C. , Cao,XH., Fuchs, J., Schubert, I., Rokhsar, D., Schmutz, J., Michael, T.P., Mayer, K.F.X., and Messing, J.: **Reduced gene content in the genome of Greater Duckweed reflects essentials in plant morphogenesis**. *Submitted* 2013.

26. Smith DR: **RNA-Seq data: a goldmine for organelle research**. *Brief Funct Genomics* 2013.
27. Xu H, Gao Y, Wang J: **Transcriptomic analysis of rice (Oryza sativa) developing embryos using the RNA-Seq technique**. *PLoS One* 2012, **7**(2):e30646.
28. Kakumanu A, Ambavaram MM, Klumas C, Krishnan A, Batlang U, Myers E, Grene R, Pereira A: **Effects of drought on gene expression in maize reproductive and leaf meristem tissue revealed by RNA-Seq**. *Plant Physiol* 2012, **160**(2):846-867.
29. Socquet-Juglard D, Kamber T, Pothier JF, Christen D, Gessler C, Duffy B, Patocchi A: **Comparative RNA-Seq analysis of early-infected peach leaves by the invasive phytopathogen Xanthomonas arboricola pv. pruni**. *PLoS One* 2013, **8**(1):e54196.
30. Raz T, Kapranov P, Lipson D, Letovsky S, Milos PM, Thompson JF: **Protocol dependence of sequencing-based gene expression measurements**. *PLoS One* 2011, **6**(5):e19287.
31. Hansen KD, Wu Z, Irizarry RA, Leek JT: **Sequencing technology does not eliminate biological variability**. *Nat Biotechnol* 2011, **29**(7):572-573.
32. Fang Z, Cui X: **Design and validation issues in RNA-Seq experiments**. *Brief Bioinform* 2011, **12**(3):280-287.
33. Kvam VM, Liu P, Si Y: **A comparison of statistical methods for detecting differentially expressed genes from RNA-Seq data**. *Am J Bot* 2012, **99**(2):248-256.
34. Robles JA, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, Taylor JM: **Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing**. *BMC Genomics* 2012, **13**:484.
35. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L: **Differential gene and transcript expression analysis of RNA-Seq experiments with TopHat and Cufflinks**. *Nat Protoc* 2012, **7**(3):562-578.
36. Chaloupkova K, Smart CC: **The abscisic acid induction of a novel peroxidase is antagonized by cytokinin in Spirodela polyrrhiza L**. *Plant Physiol* 1994, **105**(2):497-507.
37. Christmann A, Moes D, Himmelbach A, Yang Y, Tang Y, Grill E: **Integration of abscisic acid signalling into plant responses**. *Plant Biol (Stuttg)* 2006, **8**(3):314-325.
38. Rodriguez-Gacio Mdel C, Matilla-Vazquez MA, Matilla AJ: **Seed dormancy and ABA signaling: the breakthrough goes on**. *Plant Signal Behav* 2009, **4**(11):1035 - 1049.
39. Finkelstein RR, Gampala SS, Rock CD: **Abscisic acid signaling in seeds and seedlings**. *Plant Cell* 2002, **14 Suppl**:S15-45.

40. Chen J, Huang B, Li Y, Du H, Gu Y, Liu H, Zhang J, Huang Y: **Synergistic influence of sucrose and abscisic acid on the genes involved in starch synthesis in maize endosperm**. *Carbohydr Res* 2011, **346**(13):1684-1691.
41. Cosgrove DJ: **Loosening of plant cell walls by expansins**. *Nature* 2000, **407**(6802):321-326.
42. Reisen D, Leborgne-Castel N, Ozalp C, Chaumont F, Marty F: **Expression of a cauliflower tonoplast aquaporin tagged with GFP in tobacco suspension cells correlates with an increase in cell size**. *Plant Mol Biol* 2003, **52**(2):387-400.
43. Lin W, Peng Y, Li G, Arora R, Tang Z, Su W, Cai W: **Isolation and functional characterization of PgTIP1, a hormone-autotrophic cells-specific tonoplast aquaporin in ginseng**. *J Exp Bot* 2007, **58**(5):947-956.
44. Hundertmark M, Hincha DK: **LEA (late embryogenesis abundant) proteins and their encoding genes in Arabidopsis thaliana**. *BMC Genomics* 2008, **9**:118.
45. Espelund M, Saeboe-Larssen S, Hughes DW, Galau GA, Larsen F, Jakobsen KS: **Late embryogenesis-abundant genes encoding proteins with different numbers of hydrophilic repeats are regulated differentially by abscisic acid and osmotic stress**. *Plant J* 1992, **2**(2):241-252.
46. Ried JL, Walker-Simmons MK: **Group 3 Late Embryogenesis Abundant Proteins in Desiccation-Tolerant Seedlings of Wheat (Triticum aestivum L.)**. *Plant Physiol* 1993, **102**(1):125-131.
47. Rook F, Corke F, Card R, Munz G, Smith C, Bevan MW: **Impaired sucrose-induction mutants reveal the modulation of sugar-induced starch biosynthetic gene expression by abscisic acid signalling**. *Plant J* 2001, **26**(4):421-433.
48. Tao X, Fang Y, Xiao Y, Jin YL, Ma XR, Zhao Y, He KZ, Zhao H, Wang HY: **Comparative transcriptome analysis to investigate the high starch accumulation of duckweed (Landoltia punctata) under nutrient starvation**. *Biotechnol Biofuels* 2013, **6**(1):72.
49. Ekman A, Hayden DM, Dehesh K, Bulow L, Stymne S: **Carbon partitioning between oil and carbohydrates in developing oat (Avena sativa L.) seeds**. *J Exp Bot* 2008, **59**(15):4247-4257.
50. Weselake RJ, Taylor DC, Rahman MH, Shah S, Laroche A, McVetty PB, Harwood JL: **Increasing the flow of carbon into seed oil**. *Biotechnol Adv* 2009, **27**(6):866-878.
51. Jakoby M, Weisshaar B, Droge-Laser W, Vicente-Carbajosa J, Tiedemann J, Kroj T, Parcy F: **bZIP transcription factors in Arabidopsis**. *Trends Plant Sci* 2002, **7**(3):106-111.
52. Fujita Y, Fujita M, Shinozaki K, Yamaguchi-Shinozaki K: **ABA-mediated transcriptional regulation in response to osmotic stress in plants**. *J Plant Res* 2011, **124**(4):509-525.
53. Bensmihen S, Rippa S, Lambert G, Jublot D, Pautot V, Granier F, Giraudat J, Parcy F: **The homologous ABI5 and EEL transcription factors function antagonistically to fine-tune gene expression during late embryogenesis**. *Plant Cell* 2002, **14**(6):1391-1403.
54. Fujimoto SY, Ohta M, Usui A, Shinshi H, Ohme-Takagi M: **Arabidopsis ethylene-responsive element binding factors act as transcriptional activators**

or repressors of GCC box-mediated gene expression. *Plant Cell* 2000, **12**(3):393-404.
55. Riechmann JL, Meyerowitz EM: **The AP2/EREBP family of plant transcription factors**. *Biol Chem* 1998, **379**(6):633-646.
56. El-Sharkawy I, Sherif S, Mila I, Bouzayen M, Jayasankar S: **Molecular characterization of seven genes encoding ethylene-responsive transcriptional factors during plum fruit development and ripening**. *J Exp Bot* 2009, **60**(3):907-922.
57. Ohta M, Ohme-Takagi M, Shinshi H: **Three ethylene-responsive transcription factors in tobacco with distinct transactivation functions**. *Plant J* 2000, **22**(1):29-38.
58. Schoffl F, Prandl R, Reindl A: **Regulation of the heat-shock response**. *Plant Physiol* 1998, **117**(4):1135-1141.
59. Rushton DL, Tripathi P, Rabara RC, Lin J, Ringler P, Boken AK, Langum TJ, Smidt L, Boomsma DD, Emme NJ *et al*: **WRKY transcription factors: key components in abscisic acid signalling**. *Plant Biotechnol J* 2012, **10**(1):2-11.
60. Antoni R, Rodriguez L, Gonzalez-Guzman M, Pizzio GA, Rodriguez PL: **News on ABA transport, protein degradation, and ABFs/WRKYs in ABA signaling**. *Curr Opin Plant Biol* 2011, **14**(5):547-553.
61. Himmelbach A, Yang Y, Grill E: **Relay and control of abscisic acid signaling**. *Curr Opin Plant Biol* 2003, **6**(5):470-479.
62. Higo K, Ugawa Y, Iwamoto M, Korenaga T: **Plant cis-acting regulatory DNA elements (PLACE) database: 1999**. *Nucleic Acids Res* 1999, **27**(1):297-300.
63. Lang GA: **Plant dormancy: physiology, biochemistry and molecular biology**: CAB International; 1996.
64. Webb AA, Larman MG, Montgomery LT, Taylor JE, Hetherington AM: **The role of calcium in ABA-induced gene expression and stomatal movements**. *Plant J* 2001, **26**(3):351-362.
65. Cheng SH, Willmann MR, Chen HC, Sheen J: **Calcium signaling through protein kinases. The Arabidopsis calcium-dependent protein kinase gene family**. *Plant Physiol* 2002, **129**(2):469-485.
66. Stomp AM: **The duckweeds: a valuable plant for biomanufacturing**. *Biotechnol Annu Rev* 2005, **11**:69-99.
67. Wu Y, Messing J: **RNA interference-mediated change in protein body morphology and seed opacity through loss of different zein proteins**. *Plant Physiol* 2010, **153**(1):337-347.
68. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq**. *Bioinformatics* 2009, **25**(9):1105-1111.
69. Langmead B, Salzberg SL: **Fast gapped-read alignment with Bowtie 2**. *Nat Methods* 2012, **9**(4):357-359.
70. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research**. *Bioinformatics* 2005, **21**(18):3674-3676.
71. Young MD, Wakefield MJ, Smyth GK, Oshlack A: **Gene ontology analysis for RNA-Seq : accounting for selection bias**. *Genome Biol* 2010, **11**(2):R14.

Supplementary Information

1. Systematics, Morphology, and Development of *Spirodela polyrhiza*

The subfamily of the *Lemnoideae* (duckweeds) are aquatic plants which are distributed worldwide from tropical to moderate climates and grow floating on the surface of still waters like lakes and ponds. Typical features of the *Lemnoideae* are their rapid and predominantly vegetative growth and their small sizes ranging from a few centimeters down to less than a millimeter frond size in *Wolffia*, a genus containing the smallest flowering plants in the world. The morphology of duckweeds, however, is not a simple miniaturization of the body plan commonly found for angiosperms but rather a reduction and simplification of it (Figure S1A). Roots are either completely absent in *Wolffiella* and *Wolffia* but are present in Spirodela, *Landoltia* and *Lemna*. Up to 21 short and slender adventitious roots do not have lateral roots and root hairs. The leaf-like body is generally called a 'frond,' containing a shoot reduced to a single vegetative point. In contrast to leaves, daughter fronds originate from pouches of the basal section of mother fronds and remain connected to the mother via a short tissue band, the stipule. The daughter frond itself produces more new fronds thereby repeating the pattern of its own generation. The connected fronds may form small groups, called cluster, unless they are separated by external conditions. This process of rapid vegetative reproduction enables duckweeds to double their biomass and cover surfaces of lakes in a very short time. To survive in moderate climates, duckweeds undergo a seasonable developmental change (Figure S1C). To escape freezing water surface during cold periods, fronds at lower temperature become turions, which are denser and sink in water. Turions represent a

dormant stage and achieve their sinking property with the accumulation of starch, also an energy storage that can be utilized in spring, when turions cycle back to fronds. Such a change can also be induced with abscisic acid in the laboratory and involves the onset of specific gene expression for starch biosynthesis[1].

Within the *Lemnoideae* clade, a clear trend of reduction to increasingly simpler structures is evident from the more ancient Spirodela to the more derived species[2-4]. The reduction affects all parts of the plant, decreasing number and size of roots, flowers and flowering frequencies, size and branching of transport tissues like the xylem as well as the anatomy of the frond. The frond of duckweeds is not a 'frond' by strict botanical definition but has rather been described as a thalloid, leaf- and stem-like structure with juvenile or embryonic characteristics[5]. One common interpretation proposes that the frond remains at the cotyledon or early seedling stage. In this interpretation, the peculiar structure of fronds of Spirodela and other duckweed species results from heterochrony – or more specifically from neoteny, i.e. a decreased change in the rate of developmental processes compared to their ancestors. Support for this interpretation comes from the reduction within the *Lemnoideae* as well as fossil records of an duckweed ancestor of the cretaceous period, Limnobiophyllum (Figure S1B), which shows an elaborated vein system and a complex root organization with lateral roots and root hairs[6].

Due to the overall highly miniature size and simple anatomy of duckweeds, systematics of this family has been controversial for a long time. Molecular and morphological studies placed the *Lemnoideae* into the family of the *Araceae* and the order of the *Alismatales*, an early offshoot of the monocots[4, 7-9]. A simplified phylogenetic tree of flowering plants was presented in Figure 1.

2. Genome and Transcriptome Sequencing and Assembly

2.1 Genomic DNA isolation

One cluster of 3-5 fronds of *Spirodela polyrhiza* ecotype 7498, which were clonally grown from one plant to reduce potential sequence polymorphism, was aseptically transplanted into half-strength Schenk and Hildebrandt basal salt mixture (Sigma, S6765) with 1% sucrose liquid medium at pH 5.8. The cultures were kept in a growth-chamber, maintained at 100 μmol.m^{-2}.s^{-1} and 23°C through a 16h-light, 8h-dark photoperiod[1]. High molecular weight of nuclear DNA was extracted by adaption of a nuclei isolation procedure[10] and the CTAB method[11]. Simply, after grinding 10 g of frozen tissue in liquid nitrogen, nuclei were isolated with a sucrose-based buffer, and then suspended in 50 ml CTAB extraction buffer. The isolated DNA was digested with 50 μg/ml RNase for one hour at 37°C. The quality and quantity were checked with a 1% gel and measured with Nanodrop 1000.

2.2 Haploid Genome Size

For flow cytometric genome size estimations 10 mg of fresh duckweed tissue were chopped together with similar amounts of an internal reference standard *Raphanus sativus* 'Voran' (IPK gene bank accession number RA 34, 543 Mb) with new razor blades in propidium iodide-containing nuclei isolation buffer[12]. Measurements were performed on a FACSStar^{PLUS} cell sorter (BD Biosciences). The calculated average value is based on at least 10 independent measurements per reference standard performed on

separate days. The genome size of *Spirodela polyrhiza* ecotype 7498 was estimated to be ~158 Mb (Figure S2).

2.3 Genome Sequencing

A high-quality genome sequence was produced with the Roche/454 and Sanger ABI-3730 platforms, using the whole-genome shotgun (WGS) sequencing method[13, 14]. Sequencing reads for the nuclear genome were collected with the Roche 454 XLR next-generation sequencing platform at the Department of Energy Joint Genome Institute in Walnut Creek, California (http://www.jgi.doe.gov/sequencing/protocols/prots_production.html). Two linear Roche 454 libraries (8 runs, 2.95 Gb) and one 5.7 kb insert size paired library (343.4 Mb) were sequenced with standard XLR protocols. Paired ends were also generated from BAC and fosmid libaries and 24 entire fosmids were obtained using standard protocols on ABI3730XL machines at the HudsonAlpha Institute in Huntsville, Alabama. A total of 113.6K fosmid ends from a 39.5kb insert library and 30,720 BAC ends from a 101.8kb insert library were collected. We generated a total of 10,519,519 reads, or 22.53-fold genome coverage, of which the Sanger reads provided 0.75-fold coverage and the 454 reads provided 21.78-fold coverage. A summary of the input sequence data, together with their mean insert size, number of reads from each library and estimated sequence depth used for assembly is presented in Table S1.

2.4 Assembly, Scaffolding, and Contig Quality Assessment

The assembly was generated using Newbler version 2.6 with default parameters after trimming poor bases from ends and masking vector sequences. In total, 1071

scaffolds were formed with an N50 of 3.7 Mb. Over 97% (141.8 Mb) of the total sequence is represented in 252 scaffolds (Table S2). The average, N50, minimum and maximum contig sizes have been added to the supplement (Table S2). The N50 (14.5 kb) is comparable with other plant genomes published to date and only 9% of the scaffolded genome is in gaps of Ns.

A Spirodela BAC library was constructed from high molecular weight nuclear DNA[10]. The DNA was partially digested with Hind III, double size-selected and ligated into pIndigoBAC-5 (Hind III-cloning Ready, Epicentre, BACH095H). A total of 15,360 Spirodela BAC clones with an average insert size of 110 kb representing 10X genome equivalents were fingerprinted with the SNaPshot HICF fingerprinting method[15]. Of the 15,360 fingerprinted clones, 11,770 clones (76.6%) were suitable for contig assembly. There was on average one restriction fragment every 1.19 kb. The final FPC map spanned 200 Mb and contained 269 singletons and 11,501 BAC clones, which were integrated into 320 contigs. In the physical map, 23 contigs had more than 100 clones each, 62 contigs had 50-99 clones each, 160 contigs had 10-49 clones, and the residual 75 had less than 10 clones. Based on the physical map integration, the Newbler scaffolds were ordered by the FPC draft sequence function and pseudomolecules were constructed from joined scaffolds. Scaffolds within one pseudomolecule were interlaced by a stretch of 500 undefined bases (‘N’s). In total, we obtained with the aid of the physical map 32 pseudomolecules named as “pseudo1-32”. All remaining scaffolds were deposited into “pseudo0”. Sizes for each pseudomolecule is shown in Table S3.

To assess the accuracy of the assembly, 24 fosmid clones with the size of 40 kb were randomly picked, sequenced, and reconstructed into contigs. The comparison of the

chosen fosmid clones and the finished genome assembly confirmed coverage and sequence-level accuracy. Among 24 fosmids, 20 matched exactly the Newbler assembly. Fosmid 13497 was split between pseudo0 and pseudo10. Fosmid 13502 was one from the chloroplast genome (JN160603) (Table S6). In total, 89.7% of fosmid sequences were represented in the Newbler assembly; almost 10% of fosmid sequences could not be matched due to the failure of Newbler to assemble repetitive elements. The actual bp error rate was lower than 8 in 10,000 bp (>98.22%).

To align the 32 BAC-based pseudomolecules with linkage groups represented by individual chromosome pairs, we applied fluorescence in situ hybridization (FISH). BACs with a low repeat content according to RepeatMasker (http://www.repeatmasker.org/) analysis were assembled into contigs spanning the physical map of *Spirodela polyrhiza*. These BAC contigs were subdivided into appropriate probes for FISH. Metaphase chromosome spreads of Spirodela were prepared from young root tips treated with 0.002 M 8-hydroxyquinoline for 1 h on ice before fixation in absolute ethanol: acetic acid (3:1, v/v) for 48 h and digestion for 90 min at 37°C in 1% cellulase and 1% pectinase in 0.01 M sodium citrate at pH 4.8. Digested root tips were squashed in 75% acetic acid on slides and frozen on dry ice. BAC probe labeling, FISH, microscopic evaluation and image processing were performed as described[16]. We found 40 chromosomes of the 2n genome in *S. polyrhiza*. FISH data supported the coherence of 30 BAC-based pseudomolecules in distinct chromosome pairs (Figure 4). In two cases (pseudo #7 and #22) individual BACs were located on other chromosomes than the remaining contig. For instance 002B12 and 035P14 labeled another chromosome than the other 3 BACs of pseudo #7. Thus, these two

pseudomolecules were chimeric and each of the two arms had to be separated and joined to another arm to from one of the 20 chromosomes (Figure S4). Chimerism of pseudomolecules could be due to the short reads of the 454 sequencing–platform and repeat-dense regions of the chromosomes. Whereas future work will have to convert pseudomolecules into chromosomes, the current analysis of the gene content and order remained unaffected.

To confirm that no other chimerism underlies the overall assembly quality and completeness, we localized telomeric repeats in the pseudomolecules. In land plants, telomeres are characterized by tandem repeats of the conserved hexa- or heptamer sequences TTAGGG or TTTAGGG. Clusters of telomeric repeats were exclusively identified at the ends of the pseudomolecules indicating that there were no hybrids of chromosomal arms (Figure S3). Confirmation of the accuracy of sequence assembly was also possible with the distribution of repeat elements in the pseudomolecules described after the analysis of repeat elements below (see section 3.1.6).

2.5 Transcriptome

To collect a diverse set of expressed genes, *Spirodela polyrhiza* was grown under different light-dark cycles (24h/0h, 16h/8h, 12h/12h, 8h/16h, 0h/24h) and collected at time points that represent various states of the circadian clock (2AM, 6AM, 10AM, 14PM, 18PM, 22PM). In addition, *Spirodela polyrhiza* cultures were treated by various stress conditions (heat treatment at 37°C, cold treatment at 0°C, desiccation on agar plate, high pH value of 9, UV exposure, 20 mg/l $CuCl_2$, 300 mg/l KNO_3, 250 nM ABA, 10 µM kinetin, 300 mM mannitol) and samples were collected at different exposure times (0.5h,

1h, 3h, 6h, 12h, 24h) (Table S9). Fresh tissue (0.2 g) was collected from each conditions and flash-frozen in liquid nitrogen. High-quality total RNA was extracted with the RNeasy Plant Mini Kit (Qiagen, 74904). The on-column DNase I was used to remove contaminating genomic DNA (Qiagen, 79254). We used gel electrophoresis and Nanodrop 1000 to assess RNA quality and quantity. Finally, equal amounts of RNA were pooled from each sample.

The library was constructed for RNA samples and sequenced by 454. ESTs were assembled using the JGI EST sequence-processing pipeline. Briefly, raw 454 EST sequences were trimmed for vector and adaptor/linker sequences and poor reads. Contaminants were also screened and filtered by BLAST alignment. ESTs were clustered using malign and assembled using CAP3 to build tentative consensus sequences.

To assess the coverage of *Spirodela polyrhiza* genome assembly, we aligned 379,502 ESTs against the assembled scaffolds using BLAT. In total, 363,065 (95.7%) could be mapped to the genomic sequence with more than 85% of sequence coverage indicating a largely complete genomic coverage of the gene space by the assembled pseudomolecule-sequences (Table S10).

3. Genomic Elements

3.1 Repeats

To identify common and special features of the Spirodela genome the repeat data was put into a comparative context with the similar sized *Arabidopsis thaliana* At (tigr 8 version)[17] and three to four monocot genomes of different sizes, *Brachpodium*

distachyon Bd[18], *Oryza sativa* Os (rice)[19], *Sorghum bicolor* Sb[20] and *Zea mays* Zm (maize)[21] (Figure S10).

3.1.1 Kmer Frequencies

Kmer frequencies are a repeat library independent and thus unbiased method to access the repetitive portion of a genome. The program tallmyer[22] from the program suite genome tools (http://genometools.org/) was used to calculate the frequency of each 16-mer in the respective genome assemblies and other sequence sets.

The major sources of repeat elements in the genome are transposons and variable number tandem repeats (VNTRs). Further sources, like high copy number genes (e.g. rRNA genes) and different degrees of duplications (polyploidy, segmental duplications, tandem genes) usually only contribute to a far smaller extent to the repeat content. Comparing the 16mer frequency of Spirodela with other plant genomes (Figure S10) showed that the kmer curve of Spirodela followed a similar trend as the equally sized Arabidopsis genome. In both genomes kmers occurring >= 10 times are only found in ~3-4 % of the sequence amount. In the larger monocot genomes, there is a continuous rise towards increase of genome size with kmer counts starting from 12% in Brachypodium up to 63% in sorghum, which are repeated >= 10-times.

3.1.2 Detection and Annotation of Transposon and Simple Sequence Repeats

Complete LTR retrotransposons where identified in a de-novo approach with the program LTR-STRUC[23] yielding 217 candidate sequences. After quality filtering, which required < 30% tandem repeat content, at least one typical inner protein domain, manual dot plot inspection, and removal of overlapping sequences, we obtained 170

Spirodela specific full-length LTR-retrotransposons that were added to mips-REdat, a comprehensive plant repeat data base[24]. An additional 94 complete LTR-retrotransposons were detected by homology search against the 170 full-length sequences, leading to a total of 264 elements having both LTRs. The insertion age of those full-length LTR-retrotransposons was derived from the divergence (emboss distmat with Kimura 2 parameter distance) between the left and right LTR sequences, which were identical after transposition as described elsewhere[25].

Transposons and rRNA genes were annotated by the wublast version of RepeatMasker-open-3-3-0 (http://www.repeatmasker.org) against the mipsREdat (REdat_v9.3, 387 Mb, 56,169 entries). The RepeatMasker output was subjected to two post-processing filter steps, removal of low confidence hits (length <50 bp or score <250 or identity<60%) and purification of overlapping annotations in a priority-based approach, where higher score hits were assigned first and overlapping lower score hits either shortened or, if the overlap exceeded 90% of their length, removed.

Tandem repeat sequences were detected with the program Tandem Repeats Finder [26] under default parameters. Classification of the tandem repeats were based on their monomer length and divided into microsatellites (2-9 bp), minisatellites (10-99), and satellites (>= 100bp). Overlapping annotations were joined and classified as hybrid type, if they contained more than one of the three classes.

3.1.3 Insertion Age Distribution of Full-Length LTR-Retrotransposons

Spirodela harbors almost the same amount of full-length LTR-retrotransposons as the similar sized Arabidopsis genome, but the insertions are distinctly older (average 4.6

vs 2.0 my) and very young elements are completely missing. The atypical age distribution suggests an "ancient" genome state without much recent transposon activity in combination with small removal rates. The common picture in plant genomes of younger copia and older gypsy LTR-retrotransposons is weakly visible in Spirodela (Figure S18).

3.1.4 Repeat Content and Composition

A homology search with de novo full-length LTR-retrotransposons traces 13% of the Spirodela genome as LTR-retrotransposon derived, which was perfectly in line with the small genome size (Figure S11). An annotation attempt with RepeatMasker against mips-REdat_v9.3[24] gave an additional 2.5 % of retroelements and 0.23% of DNA-transposons. A closer inspection revealed, that the DNA transposon hits were only based on small stretches of simple sequence repeats, which occurred within the template transposon sequence. The same and related problems were true for the additional retroelement hits from other species. Due to their largely unspecific nature all transposon hits from non-Spirodela template sequences were removed from the final annotation. The observed lack of transposon similarity confirmed the large evolutionary distance between Spirodela and sequenced monocot genomes.

Table S13 placed the Spirodela repeat annotation into the context of Arabidopsis and 3 other monocotyledonous genomes. The single values for Spirodela followed more or less the ones from Arabidopsis and related to its small genome size. One exception was the ratio gypsy/copia LTR-retrotransposons of 3.5 in Spirodela, next to sorghum the second highest among the 5 genomes. The percentage of tandem repeats was relatively

independent of genome size and ranged usually between ~ 2 to 3%. The higher satellite repeat content of sorghum could be explained by its fully sequenced centromeres for three chromosomes. Spirodela had an exceptional high proportion of microsatellite tandem repeats, 50% versus 3 to 6% in four reference genomes (Figure S12). This amplification even influenced the absolute amounts of microsatellite repeats, as Spirodela topped the list with 1 Mb followed over the much larger sorghum genome (0.9 Mb) (Figure S13). A detailed breakdown of microsatellite repeats into the different monomer sizes (Figure S14) showed that one of the four possible dinucleotide repeats, namely "GAGA", is responsible for the noticeable increased numbers of microsatellite repeats. Due to their high repetitively, they severely impeded elongation during sequence assembly and were prevalently found at one or both ends of a pseudomolecule together with very high 16mer frequencies, especially often in the unplaced contigs of pseudo 0 (Figure S17).

3.1.5 Chromosomal Architecture

Heat maps and stacked bar charts are used to visualize and compare specific chromosomal content from a bird's eye perspective. The higher-level heat map data was created by sliding along the chromosome with a 0.1-Mb window size and 0.02-Mb shift length and determining for each window the number and percentage of bp coverage of the respective element type, like genes or LTR-retrotransposons. For kmer-frequencies the mean and median values per window was used. The density values were corrected

for the number of Ns per window, if the N content exceeded 60% the value was set to null and drawn in gray color. The number value was extrapolated to number per Mb to facilitate comparisons. The heat maps where created from the obtained density values using the python pylab module in combination with the jet color map (low to high values from blue to red).

A more detailed insight into the annotation structure was achieved with the integrative genome viewer (http://www.broadinstitute.org/igv/)[27] by defining customized tracks for the different element types and kmer values in combination with special color codes and display options. The heat map of the 32 Spirodela pseudomolecules followed the known pattern of anti-correlation between gene and LTR-retrotransposon densities (Figure S15). LTRs were preferentially eliminated from gene-rich regions but accumulate in gene-poor regions. There is a perfect correlation between prominent retrotransposon/kmer peaks and known centromeric locations in many other plant genomes (e.g. Sorghum, Brachypodium, Rice, Cotton, Tomato, Arabidopsis). Here, based on high LTR-retrotransposon (> 80-90%) content together with high kmer values, the constrictions were to be seen as the most likely positions of the centromeres. The whole length of pseudo 2 viewed in IGV is detailed (Figure S16).

3.1.6 Assembly Check

The so called "assembly checker" is based on sequence content assessment by masking one sequence set with another via the program vmatch [http://www.vmatch.de].

The approach introduces a versatile alternative method for the quantification of assembly completeness. Whole genome sequence sets, like transcripts, reads and BES are used as test sets to determine their percent base-pair coverage with the genome assembly. After the evaluation of different matching stringency, the parameters setting "-l 50 -e 1" (= minimum hit length 50 bp, maximal 1 mismatch or indel per 50 bp) was found to be suitable for the Spirodela sequence sets. Here the matching of two different random sets of 1x genome coverage against each other gave 63% coverage, which was exactly the Lander Waterman expectation.

Figure S5 compared the Spirodela pseudomolecules assembly with three different sized sets of random sampled 454 reads. These sets represented 1, 2, and 5X genome coverage corresponding to Lander Waterman statistics of 63%, 87%, and 99%. The Spirodela assembly contained 80% of the 1X read test set and 90% of the EST and BES test sets. The content values for ESTs and BES were almost identical to the 5X read set, which should represent the whole sequence amount. Overall the assembly completeness could be verified with the described new masking method to be at least 90% for genic sequences and >=80% for the rest, which was in the same range as the values given in Table S10 and Table S8 (95.7% for ESTs, 83% for AraCyc genes).

3.2 Genes

3.2.1 Gene prediction and characteristics of the gene complement

Gene models were derived from consensus gene predictions based on *de novo* gene finders, transcript data and protein homologies. EST assemblies of Spirodela and of two sea grasses, *Posidonia oceanica* and *Zostera marina*[28], were used as transcript

evidences. Heterologous protein evidence was based on protein sequences of four monocotyledonous species - *Brachypodium*, Maize, Sorghum and Rice - and three dicotyledonous species, Arabidopsis, Poplar and Wine. For evidence by homology, spliced alignments were generated by GenomeThreader[29] using an initial seed size of 7 aa for protein and 16 bp for nucleotide alignments. For *de novo* gene finders, a training set was derived from mapping the Spirodela EST assemblies and high quality protein families to scaffold sequences. AS high quality proteins, we selected orthologous gene families from the PLAZA database[30], for which at least five distinct plant species had members differing by a maximum of 2% in the protein sizes from the mean family sequence size. Next, we performed multiple sequence alignments of the selected families applying MUSCLE[31] to confirm sequence similarity and size consistency in the alignments. Spliced alignments to Spirodela genomic scaffolds were computed using GenomeThreader and filtered for full-length alignments including start and stop codons, high similarity (blosum62 score ≥ twice size of alignment) and size consistency with the respective gene family. Remaining gene models for training were selected to be non-redundant both in terms of individual, overlapping mappings of members of one family as well as mappings of one family to multiple genomic locations. To derive full-length Spirodela transcripts from the EST assemblies, candidate ORFs were predicted applying ORFpredictor[32] with pre-computed tblastx comparisons against a database compilation of Arabidopsis, Sorghum, and *Brachypodium* proteins. Only ORFs with similarity to known proteins and aligning them by their entire length to a genomic position of Spirodela scaffolds including a start and stop codon were retained. Lastly, we compiled a non-redundant training set as described above for the PLAZA proteins.

Using the non-redundant data sets described above, we trained four *de novo* gene prediction tools, Augustus, Snap, GlimmerHMM and GeneID[33-36] and determined genome-wide predictions using the Spirodela-specific parameter sets of each tool. An additional gene finder, Fgenesh+, was run using a monocotyledonous-specific parameter matrix[37]. Next, the statistical combiner Jigsaw was trained using our training set, mapped homologies and gene models predicted by our set of *de novo* gene finders[38]. The resulting gene models constituted version 1.0.

For historical reasons, an independent set of predictions was made with a new self-training *ab initio* gene finder, GeneMark-ES-GC, developed for compositionally heterogeneous genomes (Lomsadze and Borodovsky, manuscript in preparation). Local variability in GC composition in the Spirodela genome exceeds variability observed in other plant genomes (Figure S6). Also, the differences between exon and intron GC content vary significantly between genes (Figure S7). Still, the new *ab initio* algorithm was able to predict in a single run 19,327 genes, the number close to the final number of genes in annotation. The accuracy of the new algorithm was shown to be sufficiently high by assessment on a test set generated from mapping the transcripts and high quality proteins (Sn/Sp of exact prediction of internal exons: 87.2%/74.5%).

The GeneMark-ES-GC gene predictions were later merged with the models produced by computational homologies and models the other *de novo* predictions including the Jigsaw models. Upon merging *de novo* models with no significant homology (blastp e-value > 10^{-10} versus UniprotKB/Swissprot) were included, if at least two independent predictions supported an identical gene structure. Finally, models from

the training set were integrated into this set of gene models to obtain the final set of consensus gene predictions, to which we refer as version 2.0.

In total, we predicted 19,623 gene models for the 32 Spirodela pseudomolecules and pseudo 0. Annotations for the Spirodela gene set were derived as described in Supplement 5.1. A summary of gene characteristics and a comparison to genes to other higher plant genomes was shown in Table S7. Mean exon and coding sequence sizes were similar in all 5 genomes. However, Spirodela shared with banana significantly larger gene sizes, which apparently resulted from larger introns. Several lines of evidences supported the considerably lower gene number in Spirodela compared to other higher plant species. First, three independent gene prediction pipelines - the one described above, a self-training version of genemark and an approach using homology and transcriptome data – consistently predicted a gene number below 20,000 genes. Second, we observed a similar coverage of the AraCyc pathway genes[39] between monocotyledonous species: 87%, 86% and 83% of the AraCyc genes were represented by a homolog (≥40% sequence identity, ≥70% alignment coverage) in the annotations of rice, banana and Spirodela, respectively (Table S8).

In comparison to dicots, increased GC contents were observed in monocotyledonous coding sequences, mainly due to a mutational bias of G and C in the third codon position. Spirodela protein-coding sequences exhibited a pronounced GC3 bias (Figure S8), which was the highest amongst currently sequenced monocot genomes. Elevated GC3 contents were found in Spirodela specific genes as well as genes shared with monocots and dicots and thus seemed to be a general feature of Spirodela coding sequences (Figure S9). In contrast to the reported distinct bimodal distributions in rice

and maize, Arabidopsis genes showed a sharp unimodal distribution. Broader distributions resulting from a composite of genes with low and high GC3 content were observed both for banana and Spirodela genes (Figure S8) suggesting that the distinct multimodal patterns evolved specifically, whereas high GC3 biases might have evolved independently and several times in the monocotyledonous lineage.

3.2.2 Tandem Genes

An undirected graph was constructed from a self-comparison of each proteome using blastp with protein identifiers as nodes and edges, which specified similarity matched between two proteins and were weighted by expectation values. A first filter removed all matches above a threshold e-value $E > 10^{-10}$. Next, only edges connecting two proteins with a genomic distance of less than 10 dissimilar intervening genes were retained. Tandem clusters were determined as connected components from this trimmed graph. Number of clusters and tandem repeated genes for 7 species are shown in Table S11. Arabidopsis, Tomato, Rice, Sorghum, and Brachypodium contained on average ~20% of their genes arranged in tandem clusters whereas only ~15.2% of Spirodela genes and 7.5% of banana genes were tandemly repeated in their genomes.

Interestingly, genome sequences of the latter two genomes are largely based on assemblies of next generation sequencing technologies and some very closely related tandem genes may have been collapsed in the assembly of short reads. It has to be determined by future studies, whether the two species indeed have a lower number of tandem genes or assembly artifacts cause the lower tandem gene count. Independent of this uncertainty, the observed lower number of tandem genes can only partially compensate for the lower gene count in Spirodela.

3.3 Non-coding RNAs

3.3.1 tRNAs

We applied tRNAscan-SE to the entire unmasked sequence of the 33 pseudomolecules of Spirodela to predict candidate tRNAs[40]. Excluding pseudo-tRNAs, we detected in total 191 tRNAs, of which 7 contained an intron. The tRNAs covered the whole set of 21 amino acids including one tRNA for seleno-cysteine.

3.3.2 miRNAs

Mature miRNA sequences of all plant species present in miRBase version 19[41] were mapped to the Spirodela whole-genome assembly using vmatch[42] allowing up to two mismatches. Subsequently, 150 bp flanking sequences adjacent to the 5'- and 3'-boundaries of putative miRNAs were retrieved and their secondary structure was predicted using RNAfold[43] with standard settings. The structure was evaluated using MIRcheck with default settings[44]. Putative miRNAs passing MIRcheck were retained and overlapping loci of matched miRNAs were concatenated and annotated as one miRNA. A summary of annotated candidate miRNAs is shown here (Table S12).

One miRNA family, miR156, was highly overrepresented in *Spirodela polyrhiza* with a total of 32 members while only 23, 19, 10, and 4 members were identified in maize, rice, Arabidopsis, and Brachypodium applying an identical approach, respectively. In Arabidopsis, miR156 has been shown to be required and sufficient for the promotion of the juvenile phase and repression of the phase transition from vegetative to reproductive growth[45]. The high abundance of this miRNA family may be causative

for a suppressed shoot in the *Lemnoideae* and their predominant vegetative growth by forming daughter fronds.

Interestingly, the opposite is true for miR169 and miR172. These miRNAs are highly expressed in sweet sorghum, which exhibited late flowering and drought tolerance[46], properties that were not required for the growth of Spirodela. Therefore, the expansion of miRNA gene families was consistent with the finding that the size of gene families correlates with the phenotypic traits of the organism. Still, we also would like to emphasize that the evidence for the reported candidate miRNAs were based solely on sequence similarities and phenotypes or target genes were not functionally evaluated.

3.4 Organellar Insertions

The assembled nuclear genome of Spirodela was compared by blastn against the Spirodela plastid genome (JN160603)[47] and the mitochondrial genome (JQ804980)[48] respectively, to identify the insertion of organellar sequences into the nuclear genome. We retained all hits longer than 50 bp and hits were categorized by their size (Table S4) (Table S5).

A total of 1,385 chloroplast DNA fragment insertions covering 240,242 bp (0.15%) of the nuclear genome were identified. 1,320 insertions detected were shorter than 500 bp, with 34 between 0.5 and 1 kb, 21 between 1 and 2 kb, and only 10 exceeding 2 kb with the largest being 5,197 bp (Table S4). A total of 1,589 mtDNA insertions into the nuclear genome covering 207,711 bp (0.13%) had been detected. Similar to the findings for the chloroplast insertions, 1,554 were less than 500 bp in

length, with 31 between 0.5 and 1 kb, 3 between 1 and 2 kb, and 1 exceeding 2 kb (2,185 bp) (Table S5).

4. Synteny Analysis

4.1 Methods

Analysis of genome duplications were based on all-against all blastp comparisons between non-redundant gene sets of the respective species (E-value cutoff $E< 10^{-10}$). Intra- and intergenomic duplicated segments were identified by a combination of quota-alignments[49] (exploring various quota settings for the expected evolutionary history of genome duplications) and manual inspection and curation of dot plots. Genes between candidate duplications were aligned by a global alignment similar to the methods described elsewhere[50]. Statistical significance of candidate duplications was evaluated by a Monte Carlo test. Briefly, the gene order in one genome was randomly shuffled 1,000 times by exchanging gene identifiers and gene alignments were recomputed according to the initial genomic borders of candidate segments. Alignments of the random genomes were ranked by the number of aligned homologs and all candidate duplications with a p-value < 0.001 were retained.

Synonymous K_s and non-synonymous K_a substitution rates for duplicated genes were determined with Smith-Waterman alignments of protein sequences and subsequently derived codon-based alignments[51]. Rates were computed by the Nei-Gojobori method as implemented in the KaKsCalculator tool[52]. Previous studies had shown a strong dependency of K_s values on the GC3 composition of gene pairs[49, 53]. We therefore analyzed Ks values of gene pairs separately for GC3-high (GC3 > 75% for

both genes), -medium (exactly one gene with GC3 > 75%) and -low (GC3 ≤ 75% for both genes) pairs. Divergence time estimates were based on histogram peak K_s values of the low pairs (Figure S21) (Figure S24) and a molecular clock of $\lambda=6.5x10^{-9}$ synonymous substitutions per site and year[54]. Divergence times T were computed as $T = 2\lambda K_s$. We emphasize, however, that all estimates might be biased by the unusual GC3 content of Spirodela genes and possible rate differences that had been reported in monocots[54].

4.2 Intragenomic Duplications in Spirodela

Quota-alignments and dot plots suggested the occurrence of two independent large-scale or whole genome (WGD) duplications in Spirodela (Figure S19). Copy numbers of duplicated chromosomal segments provided further support for two rounds of WGDs in Spirodela (Figure S20). For about one third of the genome, no duplicated counterpart was observed. However, segments with copy numbers of four comprised approximately a quarter of the available genome sequence and were the second largest copy number class, followed by segments with three and two copies in the genome.

Syntenic conservation between segmental blocks was significantly lower compared to those reported for grasses. Whereas syntenic regions between sorghum and rice contained on average 58% of the genes in collinear blocks[20], duplications in Spirodela showed a sparse conservation with a mean of 11.3% of syntenic paralogous pairs in collinear order (Table S14). This number might be an underestimate because global gene-based alignments between two blocks might miss small inversions or local translocations. Nevertheless, the reduced number was consistent with the older age of the presumed WGDs and a continuous loss of duplicated genes[55-57]. Synonymous

substitution rates showed a unimodal distribution indicating that both WGDs occurred within a short period of time (Figure S21) and that they could not be separated by their divergence times. We therefore refer to the WGDs in Spirodela as α^{SP}/β^{SP}. There was a distinct shift in the mean peak Ks values for GC3-high (mean ~0.85) and GC3-low (mean ~1.23) gene pairs. In this study and in agreement with other reports[49], we used the GC3-low paralogous pairs to estimate the occurrence of both WGDs at approximately 95 mya, which was older than the grass family.

4.3 Comparison of Genome Duplications Between Spirodela and Rice

We determined the syntenic relations between rice and Spirodela as described in the supplementary paragraphs above. The dot plot suggested a quota for the syntenic relation of 2:4 for rice and Spirodela duplicated segments, respectively (Figure S22) indicating that the well-known ρ-WGD in grasses and the α^{SP}/β^{SP}-WGDs in Spirodela occurred independently of each other. The reported σ-duplication in grasses predating ρ[49] had recently been placed after the split of the Zingiberales and Poales[58]. This study also reported an additional γ-WGD that was specific to the Zingiberales and hence had occurred after the split of the Alismatales and the core monocots. The Alismatales - together with the Acorales - represented the most ancient monocotyledonous clade that diverged from the core monocots, which included for example the Commelinids, Asparagales, and Liliales, approximately 130 mya ago[59]. This placed the α^{SP}/β^{SP}-WGDs in the Alismatales branch (Figure S23).

Syntenic conservation of collinear gene pairs was surprisingly slightly higher than those of the α^{SP}/β^{SP}-WGDs, with a mean of 15% of co-orthologous pairs in the

chromosomal segments showing conserved order. In total, the segments spanned 20,451 loci in rice and 11,479 in Spirodela with 4,275 and 3,710 non-redundant collinear genes, respectively. Syntenic gene pairs between rice and Spirodela showed a pronounced bimodal distribution that was clearly caused by the superimposition of two unimodal distributions of GC3-high and GC3-low gene pairs (Figure S24). Following the rationale in 4.2, we determined a mean peak Ks of ~1.7 for the GC3-low distribution translating into a divergence time of ~130 mya. This estimate closely agrees with the divergence of the Alismatales and the core monocots that had been estimated to occur between 128-131 mya[59].

5. Comparative Genomics

For comparative and functional analysis, we employed genome sequences and annotations of *Arabidopsis thaliana* TAIR10[17], *Vitis vinifera* v1.4[60], *Populus trichocarpa* v2.0[61], *Solanum lycopersicum*[62], *Brachypodium distachyon* v1.2[18], *Zea mays* AGPv2[21], *Oryza sativa* MSU7[19, 63, 64], *Sorghum bicolor* high confidence gene set v1.4[20] and *Musa acuminata*[58]. Alternative splice variants of protein and coding sequences were not considered in our comparative studies. To obtain a non-redundant set of loci for each genome, we either used representative sequence sets precompiled by the respective genome project or selected the model with the longest ORF.

5.1 Unified Datasets for Comparative Analysis

Functional classifications and annotations like protein domains or gene ontologies are predominantly inferred electronically for all plant genomes. These annotations were

based on different evidences and reflected current knowledge at the time of the respective genome project. A notable exception were rice and particularly *Arabidopsis thaliana*, for which efforts of large research communities and up to more than 10 years of ongoing manual curation created a rich and up-to-date view on the biological roles of thousands of genes[17, 65]. Outdated annotations as well as annotations highly biased in their detail might severely impair comparability of data sets for genome-wide inter-species comparisons. To achieve consistency between all genomes used in this study, we annotated all genomes in this study including *Spirodela polyrhiza* applying the AHRD pipeline[62]. Briefly, for each proteome, we computed blastp comparisons to the *Arabidopsis* TAIR10, UniprotKB/Swissprot and UniprotKB/Trembl databases[17, 66]. Next, protein domains and gene ontologies were determined by the standalone version of InterProScan[67]. Using this computational evidence, description lines as well as homogenous functional classifications were derived by AHRD for all genomes.

5.2 Orthologous Gene Sets

We applied orthoMCL to cluster the proteomes of five genomes into gene family groups, which comprised (co-)orthologs and very closely related (in-)paralogs. Parameter settings for the all-against-all blastp comparison as well as for clustering parameters were as recommended[68]. Results were illustrated for the five-way clustering (Figure 3). We classified all orthoMCL family groups into divisions, with one division containing all groups of one particular species combination. In addition, we defined a cluster class as the set of divisions containing equal species numbers (Figure 3). Divisions were named by abbreviations of the species constituting the respective division, cluster classes by the number of constituting species.

In the five-way analysis, we identified 18,765 orthoMCL clusters in total, comprising 110,787 genes. The fraction of genes that were partitioned into family clusters ranged from 61.8% for rice, 68.2% for banana, 71% for tomato, 75.2% for Spirodela and up to 81.4% for Arabidopsis. As expected, class 5 with the largest division, 'AtMuOsSpTo' - was most abundant containing 43.8% of all clusters and 58.7% of all genes. Class-1 comprising the species-specific divisions was the second largest class that included 19.2% and 29.4% of all clusters and genes, respectively. In general, abundance of divisions in one class reflected expected phylogenetic relationships between the five species. For example, in class-2 the dicot-specific division 'AtTo' and in class-3 the monocot-specific division 'OsSpMu' showed the highest cluster counts. One division - the class-4 division 'AtMuOsTo', which was missing Spirodela genes, was notably the largest division. We assumed that this division contained genes, which might have been either lost due to neoteny or might have undergone divergent evolution in Spirodela. Generally, however, the basic gene content of an angiosperm plant seemed to be well represented in Spirodela. Excluding cluster class-1 families, the total number of orthoMCL families showed only small differences between the five genomes (Figure S25), indicating a high overlapping coverage of the qualitative gene content. Although the number of families containing Spirodela genes was the lowest (10,596), we detected for the rice genome with twice the gene count as Spirodela also only 11,007 families of cluster class-2 or higher. Thus, the orthoMCL family count did not or only in part explain the lower gene number in Spirodela. However, mean copy number of Spirodela genes per family was significantly lower for Spirodela (Table S15). This suggested either loss of paralogs or reduced gene family expansions, i.e generation of in-paralogs in this species.

5.3 Gene Ontologies

Based on comparable annotations of plant genomes as described above, we investigated enriched gene ontology terms for selected divisions. Overrepresentation was computed using the R package GOstats with conditional probabilities[69]. As expected, division 'AtMuOsSpTo' showed enrichment in numerous basic biological processes like sugar, lipid, and amino acid metabolic processes, RNA processing, translation, and replication (data not shown). Genes in the Spirodela-specific division were enriched for various defense related processes and included antimicrobial peptides and NBS-LRRs (Table S17). The likewise enriched proteinases and peroxidases had also been implicated – besides many other processes - to adapted immune responses. Strong overrepresentations of immune and defense genes were also commonly seen in the other species-specific divisions (data not shown) and underpin the importance to adapt to an environment with host-specific pathogens.

As previously described, we assumed the division comprising genes of each of Arabidopsis, Tomato, Banana, and rice, but missing Spirodela members as candidates for functions that had been highly modified – either lost or divergently evolved – in Spirodela. A survey of the division 'AtMuOsTo' revealed several biological processes that were sufficiently specific to analyze consistency with the biology of Spirodela. Corresponding GO identifiers were shown in Table (Table S16). The small body size and aquatic environment of Spirodela made water transporters (GO:0006833) like aquaporins less likely relevant for a small, aquatic plant. The high buoyance of water and a floating lifestyle of Spirodela with continuous contact to water allowed for a less rigid plant body and likely required distinct cell wall architecture as compared to erect land plants.

Laccases are extracellular multicopper oxidases that act on a range of substrates including (mono) phenols and aromatic amines[70]. They have been associated – besides other processes - to cell wall cross-linking and both lignification and lignin degradation thus maintaining cell wall structure and xylem fibers. Two other terms (GO:0009664, GO:0005975) supported cell wall specific processes, which were present in the four other species but were missing or modified in Spirodela. Genes found in GO:0009664 - 'plant-type cell wall organization', were all α-expansins, which have important roles in biological processes that require cell wall loosening[71]. Although the term GO:000597 – 'carbohydrate metabolic process', could specify a broad range of metabolic processes, closer inspection of the genes underlying the overrepresentation in the four species revealed strong enrichment for hydrolases acting on cell wall and storage polysaccharides[72]. Out of 222 genes, 62% (137) encoded glycosidases, including α- and β-endoglucanases, xyloglucan hydrolases, xylanases, and polygalacturonases. The latter class of glycosidases had been associated with pectin degradation and fruit ripening, a process that was triggered by the plant hormone ethylene[21, 73]. Several terms indicated a partial loss of isoforms of two key enzymes in ethylene biosynthesis, 1-aminocyclopropane-1-carboxylate (ACC) synthase and oxidase[74]. The term GO:0009692 ('ethylene biosynthesis') contained 20 ACC-synthases of the four species in the division 'AtMuOsTo', for which no ortholog or close paralog was detected by orthoMCL. Interestingly, in the term GO:0055114 ('oxidoreduction'), 14 genes encoding isoforms of ACC oxidases were represented.

5.4 Analysis of Selected Gene Families

Gene families were selected based on prior knowledge about Spirodela biology and on biased representations of gene families, domains, and biological processes, identified in our analysis of orthoMCL clusters as well as global inter-species comparisons. Figure S26 provided an outline of our applied pipeline for genome-wide surveys of targeted gene families. Briefly, we compiled a list of gene identifiers either from publicly available, curated gene families or by selection of genes with specific PFAM/InterPro domains from the AHRD annotations[62]. For this starting list of gene identifiers, we collected all orthologous genes for this list from the orthoMCL clusters analysis of five plant species. We referred to this data set as strict gene family list. To extend the strict list and include closely related candidate in-paralogs from distinct clusters, we determined for each cluster the minimal intra-cluster similarity/threshold T of its members using an all-against all blastp comparison between all genomes. The minimal intra-cluster threshold T_i was defined as the minimal expectation value E of all pairwise similarity comparisons between members of cluster i. In addition, it was restricted to a maximal value of $E \leq 10^{-30}$. Next, we expanded our strict list by including all matches that exceeded threshold T_i to any member of the i^{th} cluster and derived by this procedure the extended gene list. For both gene lists, multiple protein sequence alignments were computed using MUSCLE[75]. Alignments were checked by Gblocks[76] and manual curation. Phylogenetic trees were constructed using FastTree[77]. Visualization and analysis of phylogenetic trees was performed with custom-made python scripts, the python module ETE2 and iHOP[78, 79]. Trees were manually inspected for reduced and amplified copy numbers of orthologous genes

between Spirodela and other plant species and results were analyzed by searches of the known literature.

5.4.1 Cellulose Biosynthesis Genes

Cell wall polysaccharides are the most abundant organic molecules on our planet, but only a few conserved genes involved in primary cell-wall biogenesis have been identified, including Cellulose synthases (Ces), Cellulose synthase-like (Csl) genes and Glycosyl transferases (GT).

We conducted the blastP analysis (e^{-20}) by querying Spirodela protein database with cell wall protein sequences in rice downloaded from the Purdue cell wall genomics website: http://cellwall.genomics.purdue.edu/families/index.html. The predicted candidates were further checked for their functional domain by InterProScan search. We summarized the gene numbers for each family and each species analyzed (Table S18). After multiple alignments for sequences of Arabidopsis, rice, and Spirodela by ClustalW, the dendrograms were constructed using the Neighbor Joining method with 1,000 of bootstrap replications in MEGA5[80]. Dendrograms were redrawn in program of TreeDyn (Figure S29) (Figure S30). The summary of results was described in the main text.

5.4.2 Expansins

The expansin superfamily comprised four families, α-, α-like, β- and β-like expansins[81]. Expansins were identified as cell-wall loosening proteins[71] involved in many plant processes including cell growth and expansion, root, and root hair expansion, fruit softening, ripening and abscission[81]. In our study, we analyzed α- (Figure S27)

and β-expansins (Figure S28), which were strongly reduced in numbers in Spirodela. Several clades of α-expansins were missing orthologous Spirodela genes including AtEXP 2, 8, 17, 11, 7, and 18. The latter two expansins had been implicated in root hair initiation with AtEXP7 restoring a short root hair phenotype in rice[82, 83]. Lack of these expansins was consistent with the reduced size and aqueous environment of Spirodela roots.

Monocots had experienced a great expansion of β-expansins, with 10 detected in banana and on average 20 members in *Poaceae* species in this study (Figure S28). However, we detected only three β-expansins in Spirodela, indicating that the expansion continually progressed along the monocot diversification or a selective decrease of this gene family in Spirodela.

5.4.3 Lignin-monomer Biosynthesis

Lignin, as a major component of secondary cell wall, played an important role for the support structure, water transport, and stress responses in vascular plants. In support of the cellulose and hemicellulose fabric a solid mesh-like structure formed by cross-linking, which gave them mechanical strength necessary for upright stature. The emergence of lignin was believed to be a critical factor for plants to adapt on land. Furthermore, lignin content and composition was relevant to the digestibility of plant cell walls. The elucidation of the pathway involved in lignin biosynthesis will greatly improve bioenergy feedstock by genetic modification.

In higher plants, monolignol biosynthesis appears to be performed by 10 gene families: Phe ammonia lyase (PAL), Trans-cinnamate 4-hydroxylase (C4H), 4-

Coumarate: CoA ligase (4CL), Hydroxycinnamoyl-CoA:shikimate/quinate hydroxycinnamoyltransferase (HCT), p-Coumarate 3-hydroxylase (C3H), Caffeoyl-CoA 3-O-methyltransferase (CCoAMT), Cinnamoyl-CoA reductase (CCR), Ferulate 5-hydroxylase (F5H), Caffeic acid O-methyltransferase (COMT), Cinnamyl alcohol dehydrogenase (CAD)[84]. A total of 63 protein sequences of the 10 gene families for monolignol biosynthesis were retrieved from the Arabidopsis database http://www.arabidopsis.org based on previous genome-wide lignin study[85]. We followed the same protocol as comparative genome analysis of lignin biosynthesis gene families across the plant kingdom[86]. Briefly, in order to find the candidate lignin biosynthesis genes, 63 protein sequences served as the base for the lignin gene identification and similarity, searched by blastP against the Spirodela database at e^{-30} except CCR at e^{-24} and COMT at e^{-08}. Secondly, we continued to select the gene family members mainly on the basis of the conserved functional domain from the candidate genes by InterproScan analysis. The protein sequences without the well-defined functional domain were removed. The final result was summarized in Table S19 and the main text.

5.4.4 Laccases

Laccases are multicopper oxidases found in many microbes, fungi, and plants and are acting on a broad spectrum of phenols and amines *in vitro*[70]. Based on their ability to oxidatively couple monolignols, it has been proposed that they play important roles in cross-linking cell wall components and in lignification. In addition, laccase activity is associated with xylem fibers and many laccases show expression patterns most prominent in the proximity of vasculature tissue[70]. Several mutants and transgenic lines in

Arabidopsis laccases have recently provided direct evidence for a role in lignification. TRANSPARENTA TESTA 10 (TT10), a loss-of-function mutation in laccase 15 (LAC15), is impaired in the oxidative polymerization of flavonoids and contains reduced lignin content[87, 88]. Significantly reduced lignin levels are observed in double mutants of LAC4 and LAC17, two laccases that contribute to the constitutive lignification of Arabidopsis stems[89].

Consistent with the low requirement of rigid supporting tissue, Spirodela lacked or had strongly reduced representations in three of the Laccase clades (Figure S31). In particular, there was no ortholog in the clade with LAC17 and at most a highly divergent Spirodela gene in the group with LAC4, two genes implicated in lignification in Arabidopsis stems. For another group completely lacking Spirodela orthologs but harboring monocotyledonous members (LAC 2, 5, 12, 13) no functional information was available.

5.4.5 Starch Biosynthesis Genes

Aquatic plants like Spirodela present exceptional tolerance to cold winter through its dormant turions in place of seeds. In fact, abundant starch in turions could be an important carbon source, not competing with seed crops like corn. The members of starch-metabolizing enzyme gene families include ADP-glucose pyrophosphorylase (AGPase), starch synthase (SS), granule-bound starch synthase (GBSS), branching (BE) and debranching enzyme (DBE)[90].

The phylogenetic analysis of AGPase genes across different plant species clearly revealed two main groups: small and large subunits except AtAPS2 (Figure S32) (Table

S20). There was one member of small subunit and three large ones in Spirodela. The dendrograms indicated that the starch synthase was phylogenetically separated into five subgroups corresponding to SSI, SSII, SSIII, SSIV and GBSS. Spirodela contained all homologues for each subclass. There were two main groups of BE. Whereas Arabidopsis, rice, and maize had two copies for group II, Spirodela had only one. We found the third independent branch from SpBEIII, which might split from the common ancestor of BE and develop Spirodela specific function. DBE were split into four subclasses: ISA1, ISA2, ISA3 and PUL. All four species had counterparts of DBE.

5.4.6 Nitrogen Assimilation: GOGAT

Nitrogen is often the major limiting nutrient for plant growth. Glutamate synthase (glutamine-oxoglutarate aminotransferase or GOGAT) is a key enzyme involved in the assimilation of inorganic nitrogen in higher plants, which contain two isoforms, one with NADH and one Ferredoxin (FD) as cofactor. Besides its role in primary nitrogen assimilation, Fd-GOGAT also plays a crucial role in the re-assimilation of ammonia released during photorespiration due to the oxidization of RuBisCo[91]. We found 10 Fd-GOGAT isoforms in the Spirodela genome compared to one in other monocots (Figure S33). The highly amplified copy number of Fd-GOGAT could explain the detoxification capabilities of duckweeds with fertilizer run-offs. Because the water surface also reflected light, these isoforms could be in addition an adaption of Spirodela to an environment of high light exposure and excessive energy, which dissipated in the photorespiratory pathway.

5.4.7 Juvenile-to-adult transition

Seed plants undergo in their life cycle a series of distinct developmental steps. After embryogenesis, and seed maturation, a plant passes through a period of juvenile vegetative growth before the adult plant acquires the competency to flower and enters the reproductive phase, producing flowers and setting seeds after fertilization. Each developmental phase is characterized and accompanied by unique and specific changes of the apical meristems and transitions are performed as 'all-or-none' switches, which are regulated by an intricate network of endogenous regulators and various environmental factors like temperature and photoperiod. Within the last ~20 years, biochemical, forward, and reverse genetic studies have revealed identity and interactions of many of the molecular players involved in the transition from the juvenile to the adult flowering phase[92-94].

We observed a recurrent pattern of copy number changes in our study of Spirodela genes that were orthologs or close paralogs to key players of this phase transition in Arabidopsis. Frequently, genes promoting adult phases showed reduced copy numbers in Spirodela, whereas repressors were retained or even amplified in the genome compared to rice and Arabidopsis. We were aware that several limitations might exist for our analysis. Missing or false positive gene annotations and absent genomic sequence counterparts in assemblies could result in incorrectly estimated copy numbers. However, such a bias should be balanced between promoters and repressors. In addition, we assumed that co-orthologous genes have similar functions between species. Although such knowledge transfer had succeeded in numerous studies, the derived functions for the Spirodela genes should be regarded as computational evidence.

The microRNA miR156, which is highly abundant in Spirodela (Supplement 3.3.2), has been shown to act as a potent promoter of the juvenile phase in Arabidopsis by repressing the SPB-(Squamosa promoter binding protein-)-like proteins SPL3, SPL4, SPL5, SPL9 and SPL10, which in turn trigger progression into the adult phase[45, 92]. Whereas the clade containing SPL3 showed conserved copy numbers between Arabidopsis, rice, and Spirodela (3 copies versus 2 copies for the latter two), we detected only one Spirodela copy of the SPL10/11 clade versus 3 and 4 copies in Arabidopsis and rice, respectively (Figure S34). The two paralogs, SPL9 and SPL15, had been reported to synergistically promote the progression into the adult phase[95, 96]. This pair was again present with two copies in Arabidopsis and at least two copies in the grasses while the Spirodela genome contains only one copy (Figure S34).

The promoting effect of the SPB genes in Arabidopsis was exerted by inhibition of (at least) six repressors of flowering. These six genes – *APETALA2 (AP2), TOE (TARGET OF EAT) 1, 2* and *3*, and *SCHLAFMÜTZE (SMZ)* and *SCHNARCHZAPFEN (SNZ)* - belonged to the AP2-EREB transcription factor gene family. We detected in the orthoMCL clusters and by blastp searches (E-value $\leq 10^{-70}$) all six members in Spirodela. This was consistent with the general pattern that repressors of the adult phase were preferentially retained despite the lower gene count in Spirodela.

In higher plants, many members of the MADS-box gene family participate in a multilayer regulatory network of flowering activators and inhibitors to integrate environmental and endogenous inputs of flowering pathways and to control the transition from an adult vegetative shoot to an inflorescence meristem characterizing the beginning of the adult reproductive phase. One subfamily, SOC1 and SOC1-like proteins act as

integrators of four different flowering pathways, the autonomous, the vernalization-, the gibberellin- and the photoperiod-dependent pathway, to activate the inflorescence and floral meristem identity genes LEAFY and the synergistically promoting MADS-box genes AP1, CAULIFLOWER (CAL), and FRUITFUL (FUL)[92, 97, 98]. In Spirodela, we observed only one copy of AP1/CAL/FUL and no copy for the SOC1-like gene family, whereas in rice and Arabidopsis, at least 3 copies for each of these MADS-box clades were present (Figure S35). Additionally, the floral meristem and organ identity genes of the SEPALLATA (SEP)-subfamily were present in four and at least five copies in Arabidopsis and the grasses, respectively, but only one member could be detected in Spirodela[99]. In contrast to the decreased numbers of MADS-box subfamilies in Spirodela activating floral meristems, the number of Spirodela genes (6 copies) in a clade containing potent repressors of flowering and flower meristems including FLOWERING LOCUS C and M (FLC, FLM), SHORT VEGETATIVE PHASE (SVP) and MADS AFFECTING FLOWERING (MAF) 2 and 3[97, 100-102] were comparable to those in Arabidopsis (8) and higher to the number of four rice homologs (Figure S35). At least one of these genes, FLC, repressing the phosphatidylethanolamine binding proteins (PEBP) FLOWERING LOCUS T (FT) and TWIN SISTER OF FT (TSF), were required for AP1 activation and thereby promoting the establishment of a floral meristem[97]. Two other genes in Arabidopsis that also encoded PEBPs, TERMINAL FLOWER 1 (TFL1) and At-CENTRORADIALIS (At-CEN) were antagonists to TFL1/TSF by maintaining an indeterminate meristem[103]. Consistent with the pattern of less floral promoters and retained floral repressors, we observed only one orthologous copy in Spirodela for FT and TSF, which were present as pairs both in Arabidopsis and rice,

whereas both copies of TFL1/TSF were retained in Spirodela (Figure S36). However, we also note a clade comprising PEBP genes that is specifically amplified in the monocotyledonous species and for which we could not deduce candidate functions (Figure S36). Interestingly, a third clade of the PEBP-family defined by the Arabidopsis gene MOTHER OF FT AND TFL1 (MFT1), had with four copies the highest number of co-orthologs in Spirodela compared to only one in Arabidopsis and two in rice. MFT1 had been implicated in the promotion of embryonic growth[104] and the high amplification in Spirodela might be consistent with a prolonged juvenile growth phase.

In summary, the Spirodela genome had retained gene functions promoting the juvenile phase and preferentially lost or reduced functions acting in the progression to adult and floral transition phases. Table S21 shows a summary of this paragraph.

6. References

1. Wang W, Messing J: **Analysis of ADP-glucose pyrophosphorylase expression during turion formation induced by abscisic acid in Spirodela polyrhiza (greater duckweed)**. *BMC Plant Biology* 2012, **12**(1):5.
2. Landolt E: **The family of Lemnaceae - a monographic study, Vols. 1**: Veroffentlichungen des Geobotanischen Institutes der Eidgenossischen Technischen Hochschule, Stiftung Rubel; 1986.
3. Landolt E: **The family of Lemnaceae - a monographic study, Vols. 2**: Veroffentlichungen des Geobotanischen Institutes der Eidgenossischen Technischen Hochschule, Stiftung Rubel.; 1987.
4. Les DH, Landolt E, Crawford DJ: **Systematics of theLemnaceae (duckweeds): Inferences from micromolecular and morphological data**. *Plant Systematics and Evolution* 1997, **204**(3-4):161-177.
5. Hillman W: **The Lemnaceae, or duckweeds**. *The Botanical Review* 1961, **27**(2):221-287.
6. Stockey R, Hoffman G, Rothwell G: **The fossil monocot Limnobiophyllum scutatum: resolving the phylogeny of Lemnaceae**. *Am J Bot* 1997, **84**(3):355.
7. Cabrera LI, Salazar GA, Chase MW, Mayo SJ, Bogner J, Davila P: **Phylogenetic relationships of aroids and duckweeds (Araceae) inferred from coding and noncoding plastid DNA**. *Am J Bot* 2008, **95**(9):1153-1165.

8. Les DH, Crawford DJ, Landolt E, Gabel JD, Kimball RT: **Phylogeny and Systematics of Lemnaceae, the Duckweed Family**. *Systematic Botany* 2002, **27**(2):221-240.
9. Rothwell GW, Van Atta MR, Ballard HE, Jr., Stockey RA: **Molecular phylogenetic relationships among Lemnaceae and Araceae using the chloroplast trnL-trnF intergenic spacer**. *Mol Phylogenet Evol* 2004, **30**(2):378-385.
10. Peterson D, Tomkins J, Frisch D, Wing R, Paterson A: **Construction of plant bacterial artificial chromosome (BAC) libraries: An illustrated guide.** *Journal of Agricultural genomics* 2000, **5**.
11. Murray MG, Thompson WF: **Rapid isolation of high molecular weight plant DNA**. *Nucl Acids Res* 1980, **8**(19):4321-4326.
12. Dolezel J, Greilhuber J, Suda J: **Estimation of nuclear DNA content in plants using flow cytometry**. *Nat Protoc* 2007, **2**(9):2233-2244.
13. Messing J, Crea R, Seeburg PH: **A system for shotgun DNA sequencing**. *Nucleic acids research* 1981, **9**(2):309-321.
14. Gardner RC, Howarth AJ, Hahn P, Brown-Luedi M, Shepherd RJ, Messing J: **The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing**. *Nucleic acids research* 1981, **9**(12):2871-2888.
15. Luo MC, Thomas C, You FM, Hsiao J, Ouyang S, Buell CR, Malandro M, McGuire PE, Anderson OD, Dvorak J: **High-throughput fingerprinting of bacterial artificial chromosomes using the snapshot labeling kit and sizing of restriction fragments by capillary electrophoresis**. *Genomics* 2003, **82**(3):378-389.
16. Schubert V, Klatte M, Pecinka A, Meister A, Jasencakova Z, Schubert I: **Sister chromatids are often incompletely aligned in meristematic and endopolyploid interphase nuclei of Arabidopsis thaliana**. *Genetics* 2006, **172**(1):467-475.
17. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M *et al*: **The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools**. *Nucleic Acids Res* 2012, **40**(Database issue):D1202-1210.
18. The_International_Brachypodium_Initiative: **Genome sequencing and analysis of the model grass Brachypodium distachyon**. *Nature* 2010, **463**(7282):763-768.
19. International_Rice_Genome_Sequencing_Project: **The map-based sequence of the rice genome**. *Nature* 2005, **436**(7052):793-800.
20. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A *et al*: **The Sorghum bicolor genome and the diversification of grasses**. *Nature* 2009, **457**(7229):551-556.
21. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA *et al*: **The B73 maize genome: complexity, diversity, and dynamics**. *Science* 2009, **326**(5956):1112-1115.
22. Kurtz S, Narechania A, Stein JC, Ware D: **A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes**. *BMC Genomics* 2008, **9**:517.

23. McCarthy EM, McDonald JF: **LTR_STRUC: a novel search and identification program for LTR retrotransposons**. *Bioinformatics* 2003, **19**(3):362-367.
24. Nussbaumer T, Martis MM, Roessner SK, Pfeifer M, Bader KC, Sharma S, Gundlach H, Spannagl M: **MIPS PlantsDB: a database framework for comparative plant genome research**. *Nucleic Acids Res* 2013, **41**(Database issue):D1144-1151.
25. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL: **The paleontology of intergene retrotransposons of maize**. *Nat Genet* 1998, **20**(1):43-45.
26. Benson G: **Tandem repeats finder: a program to analyze DNA sequences**. *Nucleic Acids Res* 1999, **27**(2):573-580.
27. Thorvaldsdottir H, Robinson JT, Mesirov JP: **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration**. *Brief Bioinform* 2013, **14**(2):178-192.
28. Wissler L, Dattolo E, Moore AD, Reusch TB, Olsen JL, Migliaccio M, Bornberg-Bauer E, Procaccini G: **Dr. Zompo: an online data repository for Zostera marina and Posidonia oceanica ESTs**. *Database (Oxford)* 2009, **2009**:bap009.
29. Gremme G, Brendel V, Sparks ME, Kurtz S: **Engineering a software tool for gene structure prediction in higher organisms**. *Information and Software Technology* 2005, **47**(15):965-978.
30. Proost S, Van Bel M, Sterck L, Billiau K, Van Parys T, Van de Peer Y, Vandepoele K: **PLAZA: a comparative genomics resource to study gene and genome evolution in plants**. *Plant Cell* 2009, **21**(12):3718-3731.
31. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity**. *BMC Bioinformatics* 2004, **5**:113.
32. Min XJ, Butler G, Storms R, Tsang A: **OrfPredictor: predicting protein-coding regions in EST-derived sequences**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W677-680.
33. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B: **AUGUSTUS: ab initio prediction of alternative transcripts**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W435-439.
34. Korf I: **Gene finding in novel genomes**. *BMC Bioinformatics* 2004, **5**:59.
35. Majoros WH, Pertea M, Salzberg SL: **TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders**. *Bioinformatics* 2004, **20**(16):2878-2879.
36. Guigó R, Agarwal P, Abril JF, Burset M, Fickett JW: **An Assessment of Gene Prediction Accuracy in Large DNA Sequences**. *Genome Research* 2000, **10**(10):1631-1642.
37. Salamov AA, Solovyev VV: **Ab initio Gene Finding in Drosophila Genomic DNA**. *Genome Research* 2000, **10**(4):516-522.
38. Allen JE, Salzberg SL: **JIGSAW: integration of multiple sources of evidence for gene prediction**. *Bioinformatics* 2005, **21**(18):3596-3603.
39. Mueller LA, Zhang P, Rhee SY: **AraCyc: a biochemical pathway database for Arabidopsis**. *Plant Physiol* 2003, **132**(2):453-460.

40. Lowe TM, Eddy SR: **tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence**. *Nucleic acids research* 1997, **25**(5):955-964.
41. Kozomara A, Griffiths-Jones S: **miRBase: integrating microRNA annotation and deep-sequencing data**. *Nucleic Acids Res* 2011, **39**(Database issue):D152-157.
42. Abouelhoda MI, Kurtz S, Ohlebusch E: **The enhanced suffix array and its applications to genome analysis**. *Algorithms in Bioinformatics, Proceedings* 2002, **2452**:449-463.
43. Denman RB: **Using RNAFOLD to predict the activity of small catalytic RNAs**. *Biotechniques* 1993, **15**(6):1090-1095.
44. Jones-Rhoades MW, Bartel DP: **Computational identification of plant microRNAs and their targets, including a stress-induced miRNA**. *Mol Cell* 2004, **14**(6):787-799.
45. Wu G, Park MY, Conway SR, Wang JW, Weigel D, Poethig RS: **The sequential action of miR156 and miR172 regulates developmental timing in Arabidopsis**. *Cell* 2009, **138**(4):750-759.
46. Calviño M, Bruggmann R, Messing J: **Characterization of the small RNA component of the transcriptome from grain and sweet sorghum stems**. *BMC Genomics* 2011, **12**:356.
47. Wang W, Messing J: **High-Throughput Sequencing of Three Lemnoideae (Duckweeds) Chloroplast Genomes from Total DNA**. *PLoS ONE* 2011, **6**(9):e24670.
48. Wang W, Wu Y, Messing J: **The Mitochondrial Genome of an Aquatic Plant, Spirodela polyrhiza**. *PLoS ONE* 2012, **7**(10):e46747.
49. Tang H, Bowers JE, Wang X, Paterson AH: **Angiosperm genome comparisons reveal early polyploidy in the monocot lineage**. *Proc Natl Acad Sci U S A* 2010, **107**(1):472-477.
50. Haas BJ, Delcher AL, Wortman JR, Salzberg SL: **DAGchainer: a tool for mining segmental genome duplications and synteny**. *Bioinformatics* 2004, **20**(18):3643-3646.
51. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends Genet* 2000, **16**(6):276-277.
52. Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J: **KaKs_Calculator: calculating Ka and Ks through model selection and model averaging**. *Genomics Proteomics Bioinformatics* 2006, **4**(4):259-263.
53. Shi X, Wang X, Li Z, Zhu Q, Tang W, Ge S, Luo J: **Nucleotide substitution pattern in rice paralogues: implication for negative correlation between the synonymous substitution rate and codon usage bias**. *Gene* 2006, **376**(2):199-206.
54. Gaut BS, Morton BR, McCaig BC, Clegg MT: **Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL**. *Proc Natl Acad Sci U S A* 1996, **93**(19):10274-10279.
55. Long M, Thornton K: **Gene duplication and evolution**. *Science* 2001, **293**(5535):1551.

56. Zhang L, Gaut BS, Vision TJ: **Gene duplication and evolution**. *Science* 2001, **293**(5535):1551.
57. Lynch M, Conery JS: **The evolutionary fate and consequences of duplicate genes**. *Science* 2000, **290**(5494):1151-1155.
58. D'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M *et al*: **The banana (Musa acuminata) genome and the evolution of monocotyledonous plants**. *Nature* 2012, **488**(7410):213-217.
59. Janssen T, Bremer K: **The age of major monocot groups inferred from 800+ rbcL sequences**. *Botanical Journal of the Linnean Society* 2004, **146**(4):385-398.
60. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C *et al*: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla**. *Nature* 2007, **449**(7161):463-467.
61. Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A *et al*: **The genome of black cottonwood, Populus trichocarpa (Torr. & Gray)**. *Science* 2006, **313**(5793):1596-1604.
62. The_Tomato_Genome_Consortium: **The tomato genome sequence provides insights into fleshy fruit evolution**. *Nature* 2012, **485**(7400):635-641.
63. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L *et al*: **The TIGR Rice Genome Annotation Resource: improvements and new features**. *Nucleic Acids Res* 2007, **35**(Database issue):D883-887.
64. Yu J, Wang J, Lin W, Li S, Li H, Zhou J, Ni P, Dong W, Hu S, Zeng C *et al*: **The Genomes of Oryza sativa: a history of duplications**. *PLoS Biol* 2005, **3**(2):e38.
65. Tanaka T, Antonio BA, Kikuchi S, Matsumoto T, Nagamura Y, Numa H, Sakai H, Wu J, Itoh T, Sasaki T *et al*: **The Rice Annotation Project Database (RAP-DB): 2008 update**. *Nucleic Acids Res* 2008, **36**(Database issue):D1028-1033.
66. Magrane M, Consortium U: **UniProt Knowledgebase: a hub of integrated protein data**. *Database (Oxford)* 2011, **2011**:bar009.
67. Zdobnov EM, Apweiler R: **InterProScan--an integration platform for the signature-recognition methods in InterPro**. *Bioinformatics* 2001, **17**(9):847-848.
68. Li L, Stoeckert CJ, Jr., Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes**. *Genome Res* 2003, **13**(9):2178-2189.
69. Falcon S, Gentleman R: **Using GOstats to test gene lists for GO term association**. *Bioinformatics* 2007, **23**(2):257-258.
70. Turlapati PV, Kim KW, Davin LB, Lewis NG: **The laccase multigene family in Arabidopsis thaliana: towards addressing the mystery of their gene function(s)**. *Planta* 2011, **233**(3):439-470.
71. Cosgrove DJ: **Loosening of plant cell walls by expansins**. *Nature* 2000, **407**(6802):321-326.
72. Gilbert HJ: **The biochemistry and structural biology of plant cell wall deconstruction**. *Plant Physiol* 2010, **153**(2):444-455.
73. Prasanna V, Prabha TN, Tharanathan RN: **Fruit ripening phenomena--an overview**. *Crit Rev Food Sci Nutr* 2007, **47**(1):1-19.

74. Wang KL, Li H, Ecker JR: **Ethylene biosynthesis and signaling networks**. *Plant Cell* 2002, **14 Suppl**:S131-151.
75. Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput**. *Nucleic Acids Res* 2004, **32**(5):1792-1797.
76. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis**. *Mol Biol Evol* 2000, **17**(4):540-552.
77. Price MN, Dehal PS, Arkin AP: **FastTree 2--approximately maximum-likelihood trees for large alignments**. *PLoS One* 2010, **5**(3):e9490.
78. Huerta-Cepas J, Dopazo J, Gabaldon T: **ETE: a python Environment for Tree Exploration**. *BMC Bioinformatics* 2010, **11**:24.
79. Letunic I, Bork P: **Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation**. *Bioinformatics* 2007, **23**(1):127-128.
80. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods**. *Mol Biol Evol* 2011, **28**(10):2731-2739.
81. Choi D, Cho H-T, Lee Y: **Expansins: expanding importance in plant growth and development**. *Physiologia Plantarum* 2006, **126**(4):511-518.
82. Cho H-T, Cosgrove DJ: **Regulation of root hair initiation and expansin gene expression in Arabidopsis**. *The Plant cell* 2002, **14**(12):3237-3253.
83. Yu Z, Bo K, He X, Lv S, Bai Y, Ding W, Chen M, Cho H-T, Wu P: **Root hair-specific expansins modulate root hair elongation in rice**. *The Plant Journal : for cell and molecular biology* 2011, **66**(5):725-734.
84. Bonawitz ND, Chapple C: **The genetics of lignin biosynthesis: connecting genotype to phenotype**. *Annual review of genetics* 2010, **44**:337-363.
85. Raes J, Rohde A, Christensen JH, Van de Peer Y, Boerjan W: **Genome-wide characterization of the lignification toolbox in Arabidopsis**. *Plant Physiol* 2003, **133**(3):1051-1071.
86. Xu Z, Zhang D, Hu J, Zhou X, Ye X, Reichel KL, Stewart NR, Syrenne RD, Yang X, Gao P *et al*: **Comparative genome analysis of lignin biosynthesis gene families across the plant kingdom**. *BMC Bioinformatics* 2009, **10 Suppl 11**:S3.
87. Pourcel L, Routaboul JM, Kerhoas L, Caboche M, Lepiniec L, Debeaujon I: **TRANSPARENT TESTA10 encodes a laccase-like enzyme involved in oxidative polymerization of flavonoids in Arabidopsis seed coat**. *The Plant cell* 2005, **17**(11):2966-2980.
88. Liang M, Davis E, Gardner D, Cai X, Wu Y: **Involvement of AtLAC15 in lignin synthesis in seeds and in root elongation of Arabidopsis**. *Planta* 2006, **224**(5):1185-1196.
89. Berthet S, Demont-Caulet N, Pollet B, Bidzinski P, Cezard L, Le Bris P, Borrega N, Herve J, Blondet E, Balzergue S *et al*: **Disruption of LACCASE4 and 17 results in tissue-specific alterations to lignification of Arabidopsis thaliana stems**. *The Plant cell* 2011, **23**(3):1124-1137.
90. Ohdan T, Francisco PB, Jr., Sawada T, Hirose T, Terao T, Satoh H, Nakamura Y: **Expression profiling of genes involved in starch synthesis in sink and source organs of rice**. *J Exp Bot* 2005, **56**(422):3229-3244.

91. Temple SJ, Bagga S, Sengupta-Gopalan C: **Down-regulation of specific members of the glutamine synthetase gene family in alfalfa by antisense RNA technology**. *Plant molecular biology* 1998, **37**(3):535-547.
92. Huijser P, Schmid M: **The control of developmental phase transitions in plants**. *Development* 2011, **138**(19):4117-4129.
93. Irish VF: **The flowering of Arabidopsis flower development**. *The Plant Journal* 2010, **61**(6):1014-1028.
94. Jarillo JA, Pineiro M: **Timing is everything in plant development. The central role of floral repressors**. *Plant Sci* 2011, **181**(4):364-378.
95. Wang JW, Czech B, Weigel D: **miR156-regulated SPL transcription factors define an endogenous flowering pathway in Arabidopsis thaliana**. *Cell* 2009, **138**(4):738-749.
96. Wang JW, Schwab R, Czech B, Mica E, Weigel D: **Dual effects of miR156-targeted SPL genes and CYP78A5/KLUH on plastochron length and organ size in Arabidopsis thaliana**. *Plant Cell* 2008, **20**(5):1231-1243.
97. Corbesier L, Coupland G: **The quest for florigen: a review of recent progress**. *J Exp Bot* 2006, **57**(13):3395-3403.
98. Dorca-Fornell C, Gregis V, Grandi V, Coupland G, Colombo L, Kater MM: **The Arabidopsis SOC1-like genes AGL42, AGL71 and AGL72 promote flowering in the shoot apical and axillary meristems**. *Plant J* 2011, **67**(6):1006-1017.
99. Ditta G, Pinyopich A, Robles P, Pelaz S, Yanofsky MF: **The SEP4 gene of Arabidopsis thaliana functions in floral organ and meristem identity**. *Curr Biol* 2004, **14**(21):1935-1940.
100. Gregis V, Sessa A, Colombo L, Kater MM: **AGL24, SHORT VEGETATIVE PHASE, and APETALA1 redundantly control AGAMOUS during early stages of flower development in Arabidopsis**. *Plant Cell* 2006, **18**(6):1373-1382.
101. Li D, Liu C, Shen L, Wu Y, Chen H, Robertson M, Helliwell CA, Ito T, Meyerowitz E, Yu H: **A repressor complex governs the integration of flowering signals in Arabidopsis**. *Dev Cell* 2008, **15**(1):110-120.
102. Ratcliffe OJ, Kumimoto RW, Wong BJ, Riechmann JL: **Analysis of the Arabidopsis MADS AFFECTING FLOWERING gene family: MAF2 prevents vernalization by short periods of cold**. *Plant Cell* 2003, **15**(5):1159-1169.
103. Hanzawa Y, Money T, Bradley D: **A single amino acid converts a repressor to an activator of flowering**. *Proc Natl Acad Sci U S A* 2005, **102**(21):7748-7753.
104. Xi W, Liu C, Hou X, Yu H: **MOTHER OF FT AND TFL1 regulates seed germination through a negative feedback loop modulating ABA signaling in Arabidopsis**. *Plant Cell* 2010, **22**(6):1733-1748.

7. Supplementary Figure and Table List

7.1 Figure list

7.2 Supplementary Figure list

7.3 Supplementary Table list

(Table S1) Total sequence input for the genome assembly.

(Table S2) Statistics of Spirodela genome assembly.

(Table S3) Size of assembled peusomolecules.

(Table S4) Length and distribution of chloroplast DNA fragment insertions on the Spirodela scaffolds.

(Table S5) Length and distribution of mitochondrial DNA insertions in the Spirodela scaffolds.

(Table S6) Accuracy of the assembly validated by fosmid sequence.

(Table S7) Gene characteristics.

(Table S8) Comparison of gene content with the AraCyc pathway.

(Table S9) Spirodela samples collected and pooled for Roche/454 EST sequencing

(Table S10) Number and fraction of ESTs aligned to pseudomolecules.

(Table S11) Tandem genes and clusters in 7 species.

(Table S12) Number of candidate miRNAs for each miRNA family of Spirodela.

(Table S13) Repeat composition of Spirodela compared to other plant genomes.

(Table S14) Genomic size and number of conserved paralogous gene pairs of duplicated chromosomal segments in Spirodela.

(Table S15) Average copy number per orthoMCL cluster for each species.

(Table S16) Overrepresented GO terms in division of AtMuOsTo except Spirodela.

(Table S17) Overrepresented functional categories of Spirodela-specific genes.

(Table S18) Comparative numbers of cell-wall gene families in Arabidopsis, rice and Spirodela.

(Table S19) Copy numbers of mono-lignin biosynthesis genes.

(Table S20) List of starch biosynthesis genes and their origins.

(Table S21) Copy numbers of selected genes involved in juvenile-to-adult transition and flowering.

Supplementary Figures

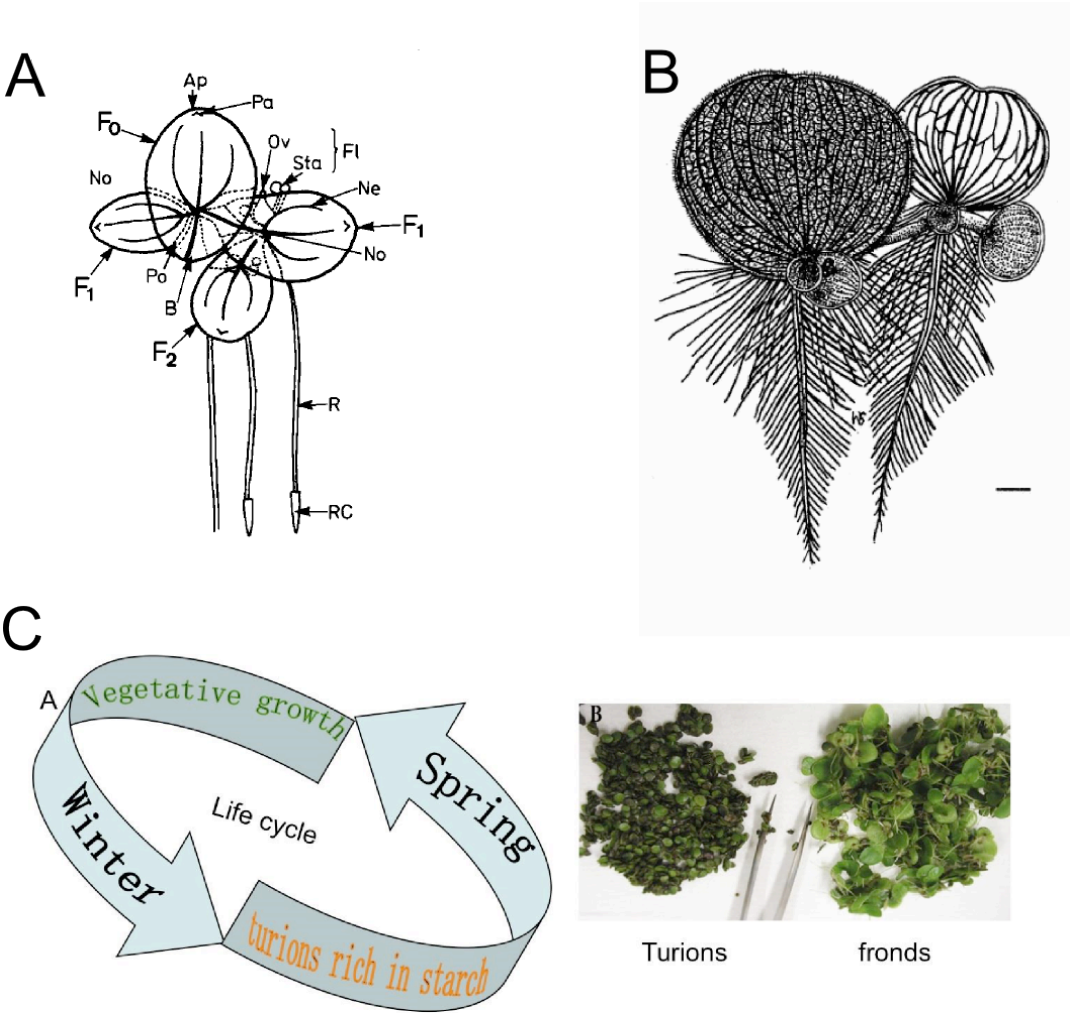


Figure S1: The morphology and life cycle of *Lemnoideae*.

In (A), a schematic drawing of *Lemna* taken from (Landolt 1986) is shown. F0, F1 and F2 represent mother and daughter fronds. Subfigure (B) shows the fossil Limnobiophyllum, an ancestor of the *Lemnoideae*. Subfigure C illustrates the vegetative life cycle of Spirodela, alternating between turions as dormant stages and fronds floating on the water surface.

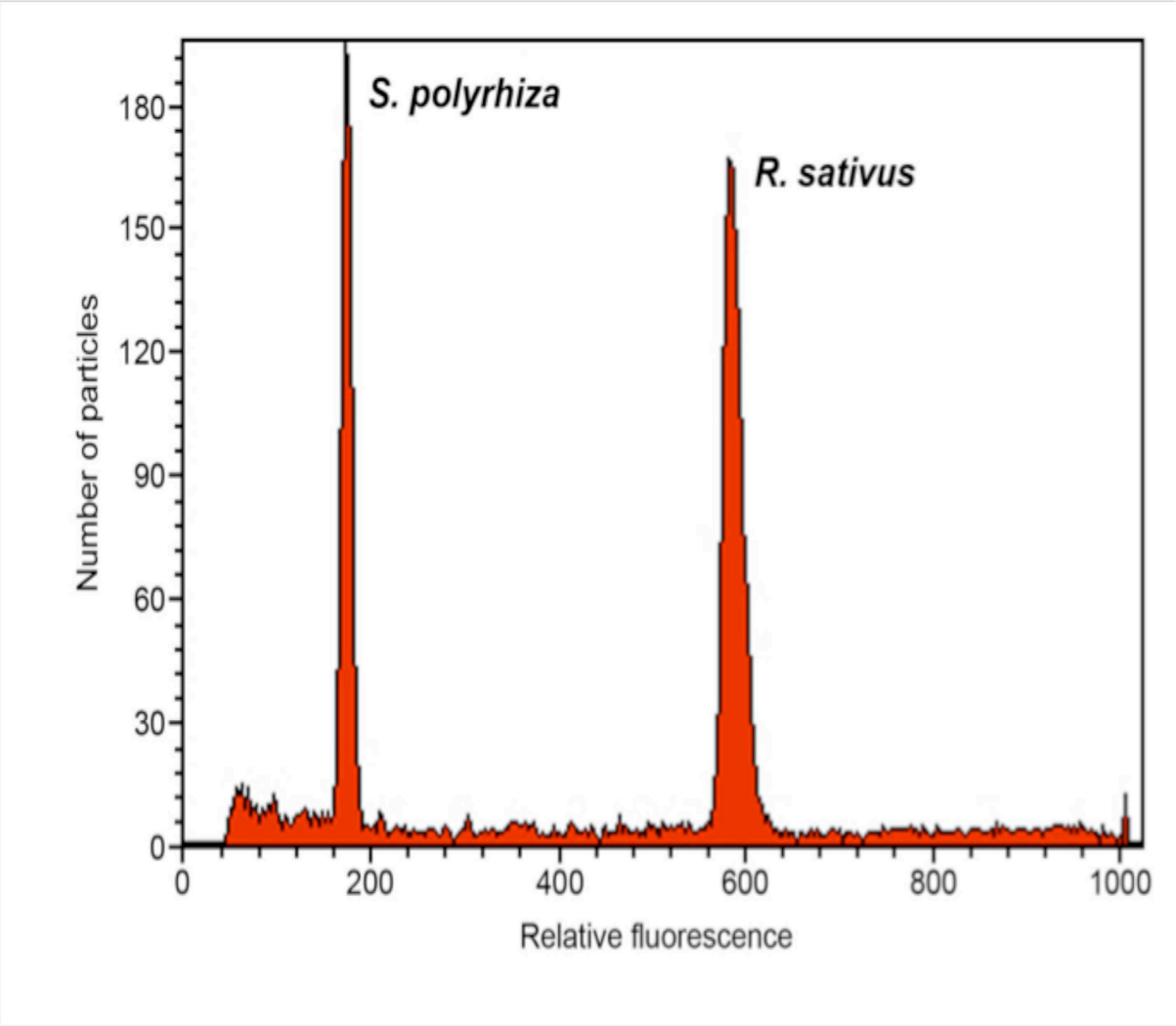


Figure S2: Genome size estimated from Flow cytometry (FCM) histogram.

The histogram shows the relative DNA content of *Spirodela polyrhiza* in relation to the internal reference standard *Raphanus sativus* (543 Mbp). Based on the G1 peak means the genome size of Spirodela (Sp) was estimated to be ~158 Mb.

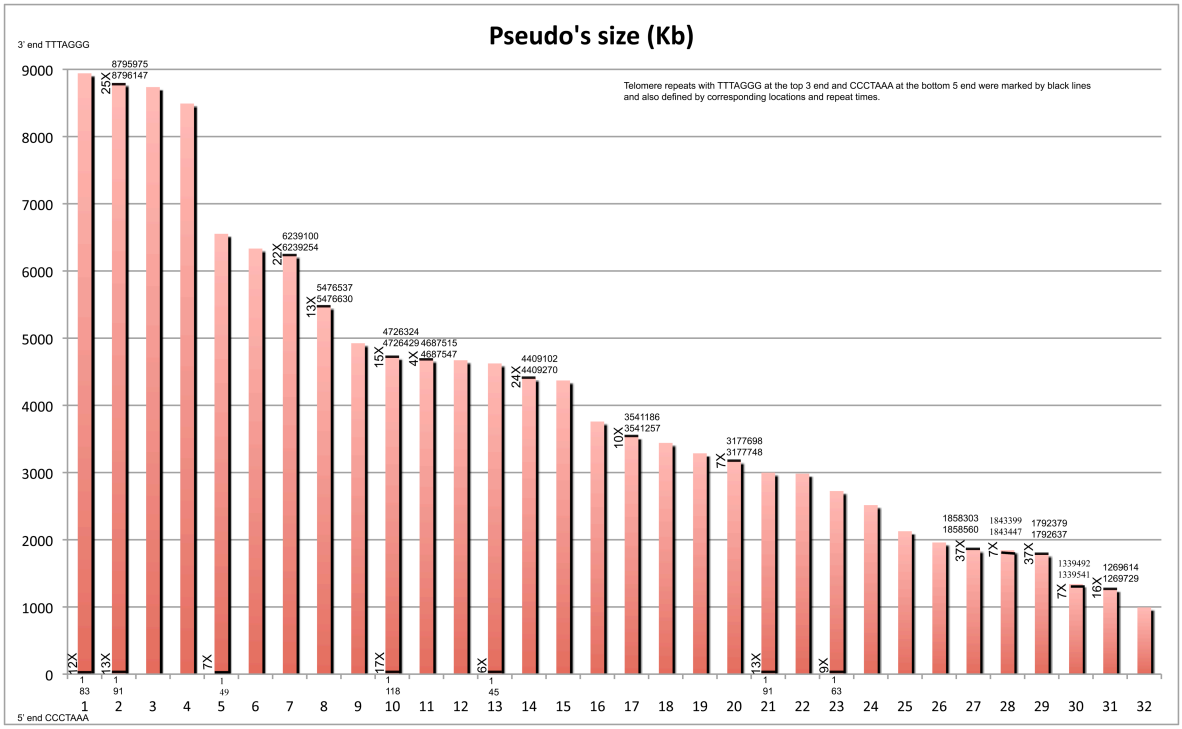


Figure S3: Alignment of telomeric sequences to the pseudomolecules.

Hexa- and Heptameric sequences characteristic for plant telomere sequences were mapped to the Spirodela pseudomolecules 1 to 32. Location and copy numbers of consecutive matches are indicated for each pseudomolecule.

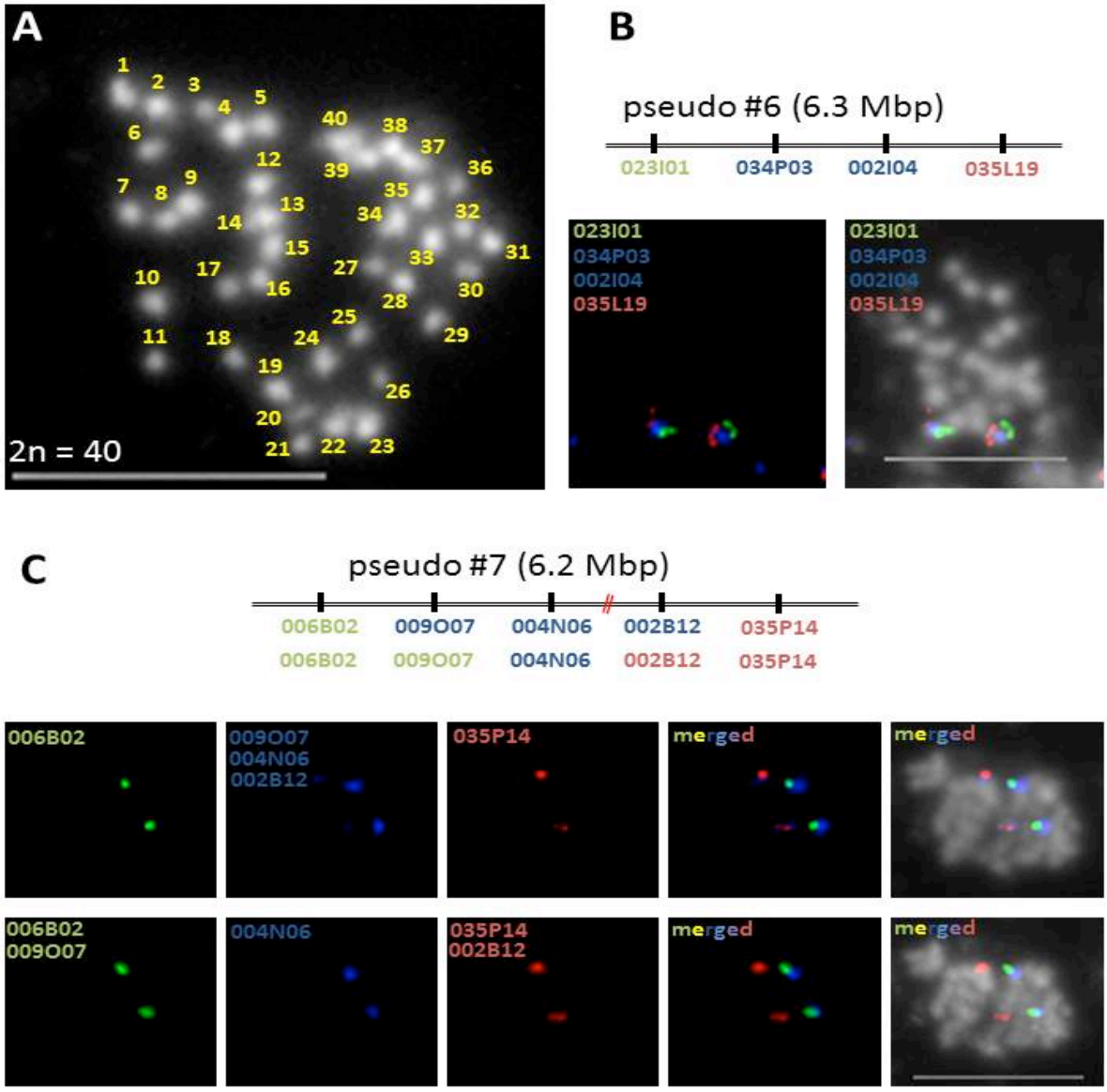


Figure S4: Cytogenetics by fluorescence in situ hybridization

A) Metaphase spread indicating a chromosome number of 2n = 40 for *Spirodela polyrhiza* 7498. B) Validation of BAC positions on a single chromosome pair by anchoring pseudo #6 of *S. polyrhiza* via multicolor FISH. C) The chimeric pseudo #7 was revealed by reprobing BACs with different fluorescence color schemes. Scale bars: 10 μm.

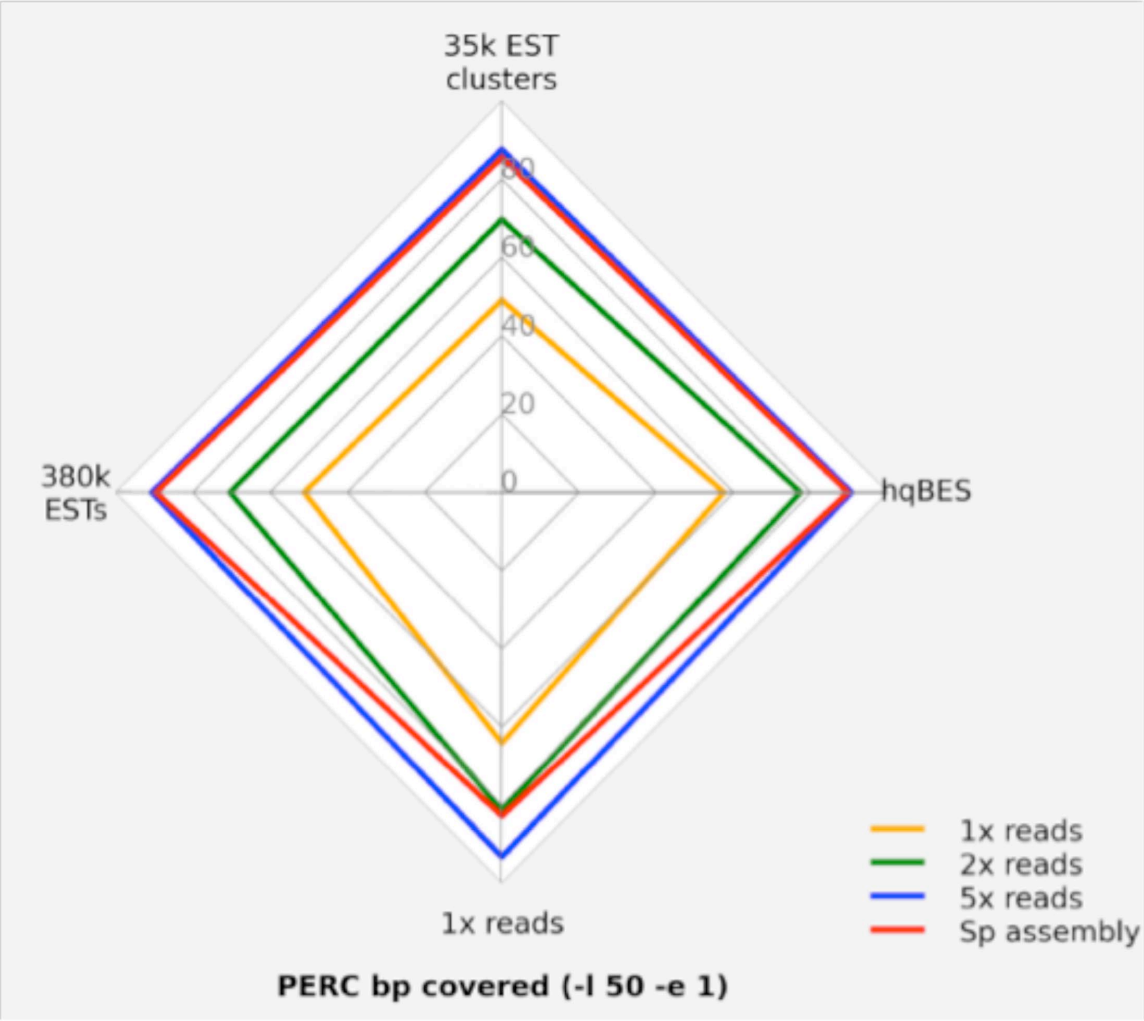


Figure S5: Sequence completeness of the Spirodela assembly compared to 454 reads set with different genome coverage.

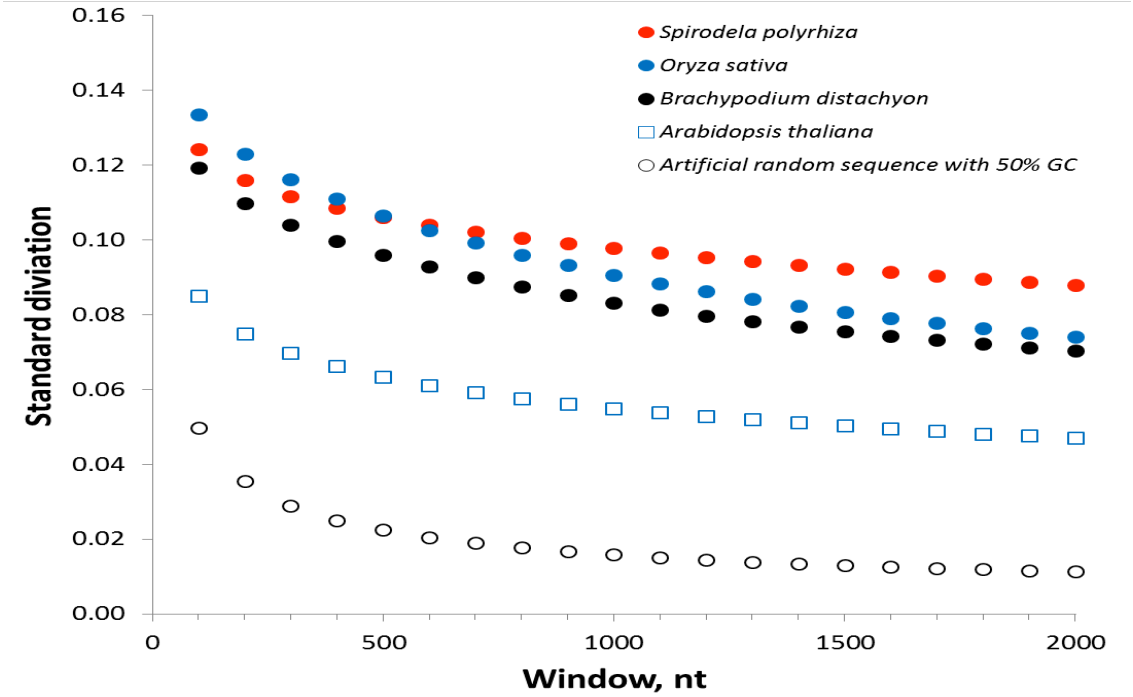


Figure S6: Comparison of genome compositional heterogeneity in several plants. Standard variation of GC content within a sequence fragment (window) is shown as a function of the window size.

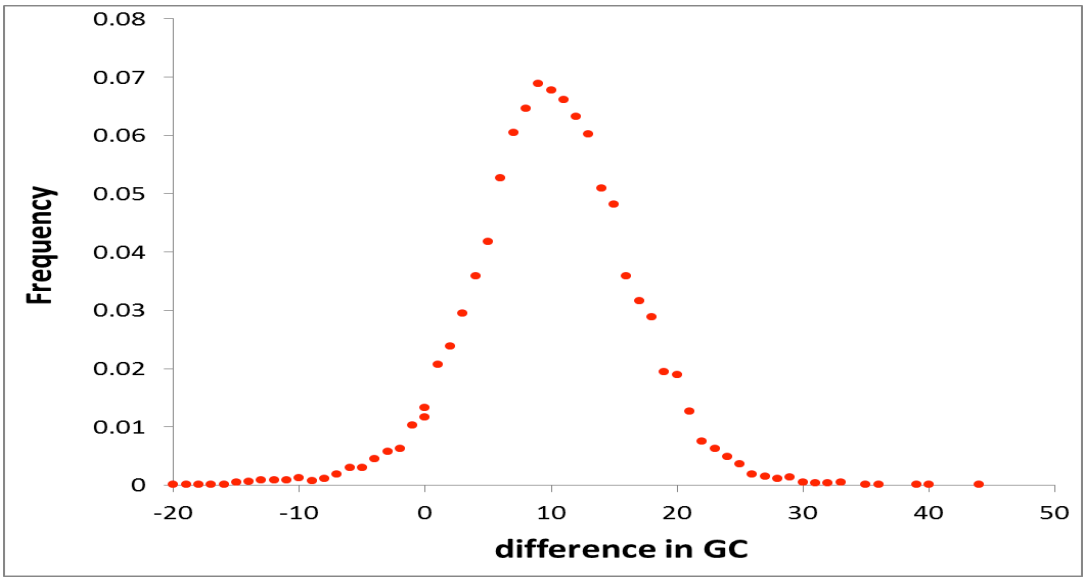


Figure S7: Distribution of difference between GC content of exons and introns in the same gene among the genes in Spirodela genome.

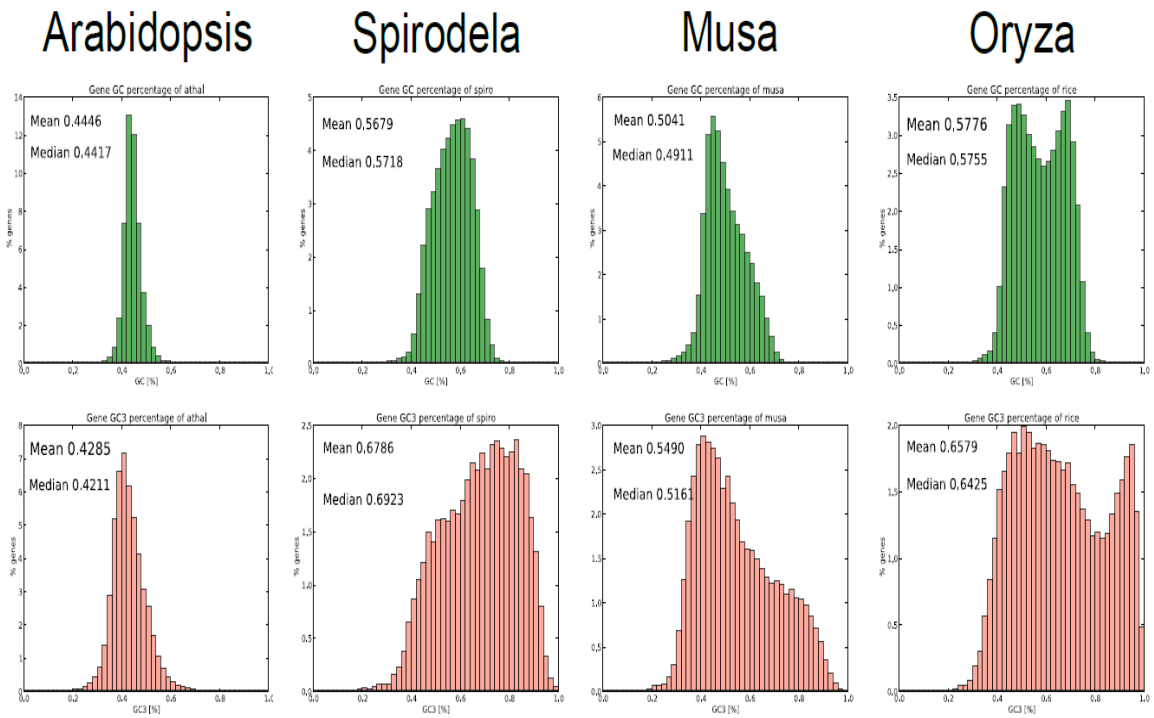


Figure S8. GC3 distributions in genes.

Distribution of GC (green histograms, top row) and GC3 (red histograms, bottom row) is unimodal for Arabidopsis, bimodal for Oryza, whereas Spirodela and Musa show skewed unimodal distributions with banana genes being more AT-rich while a higher GC3 composition is more frequent in Spirodela genes.

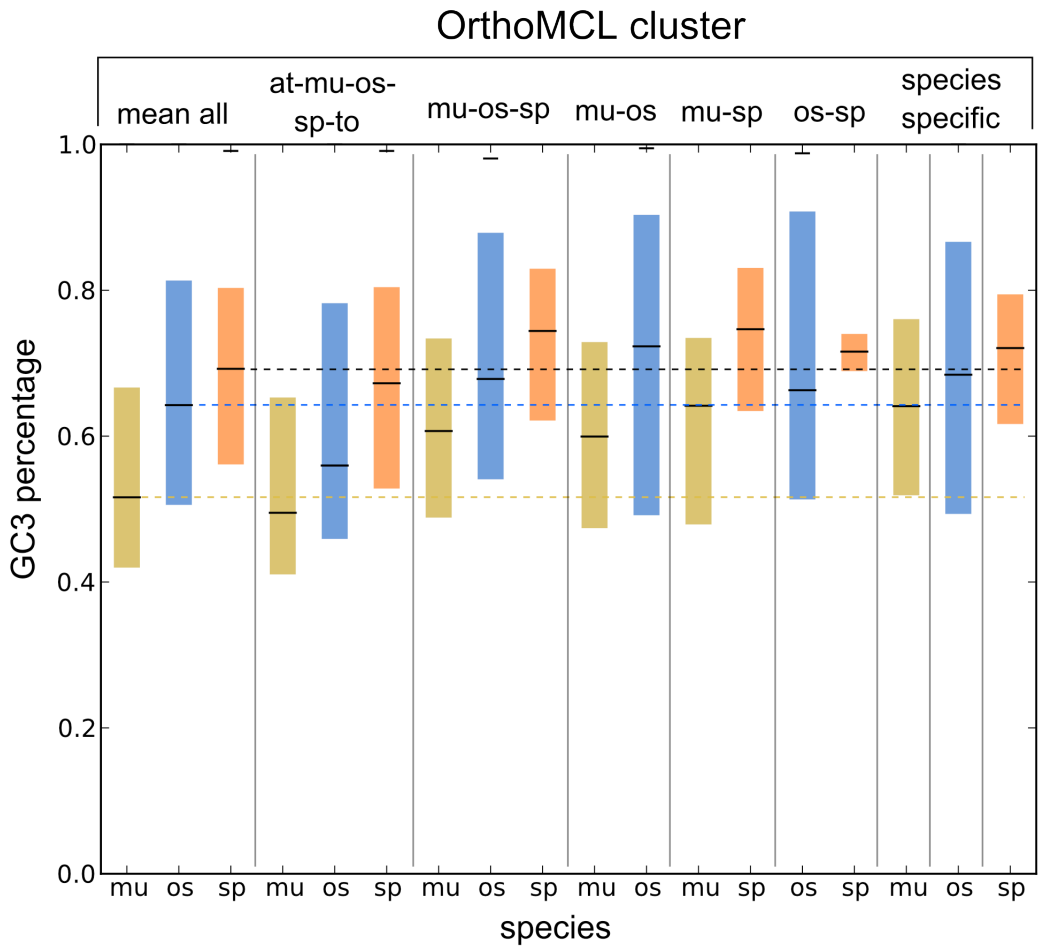


Figure S9: GC3 composition of orthoMCL divisions.

For each orthoMCL division of the Venn-diagramm (Figure S12, see also Supplement), GC3-content for each monocotyledonous species has been determined. 'At', 'To', 'Mu', 'Os' and 'Sp' indicate Arabidopsis, tomato, banana, rice and Spirodela genes, respectively. GC3 content for monocot genes is higher in monocot specific divisions compared to their genome-wide mean (left column) or genes shared in all five species. This trend is observed for all three monocot species indicating the evolution of a mutational GC3 bias in this group.

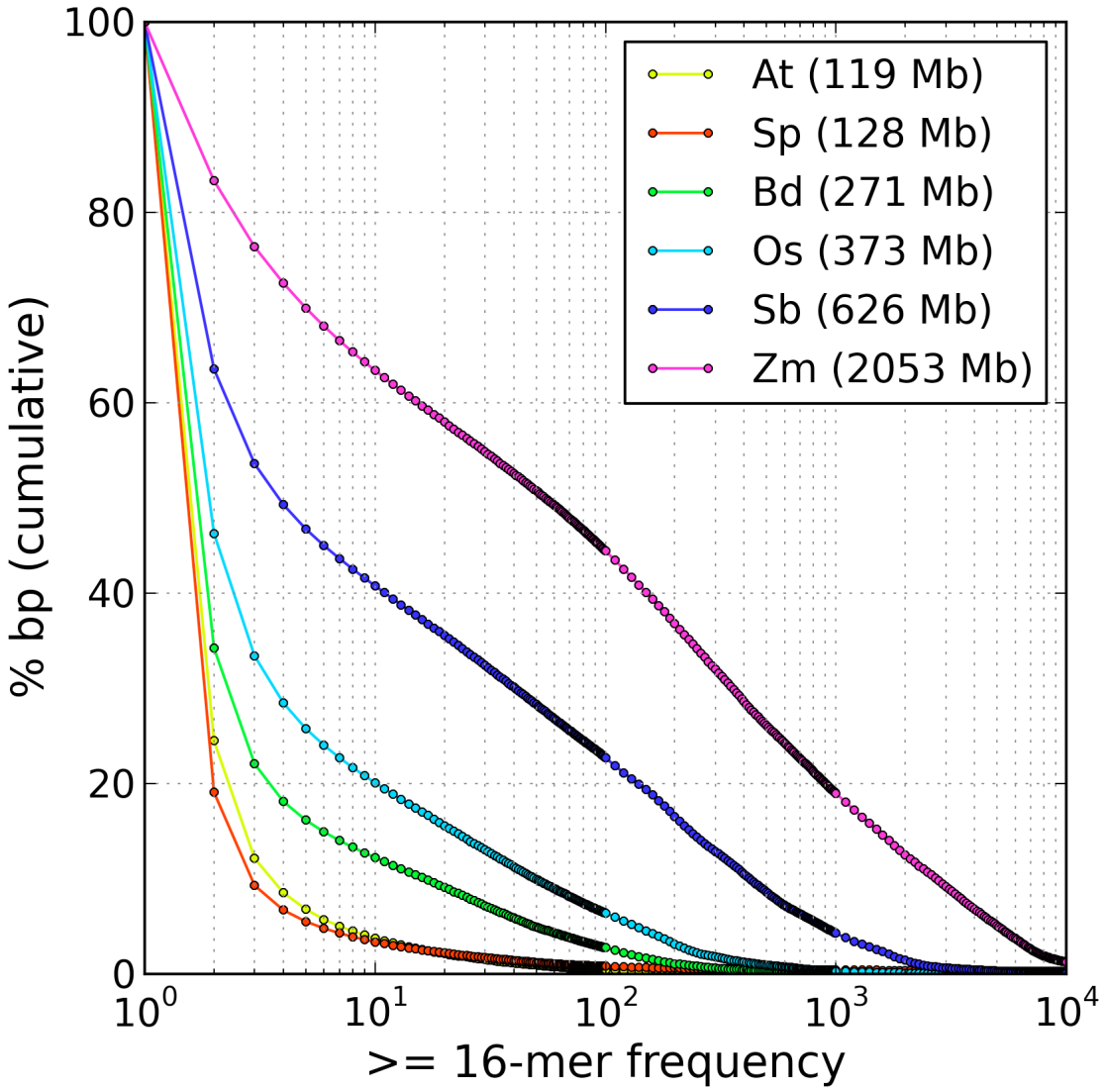


Figure S10: Repetitive elements of the Spirodela genome assembly in comparison to other plant genomes.

The figure is based on 16-mer counts and depicts the cumulative genome content versus 16-mer frequency cutoffs: e.g. in rice all 16mers occurring >= 10 times account for 20% of the total genome size.

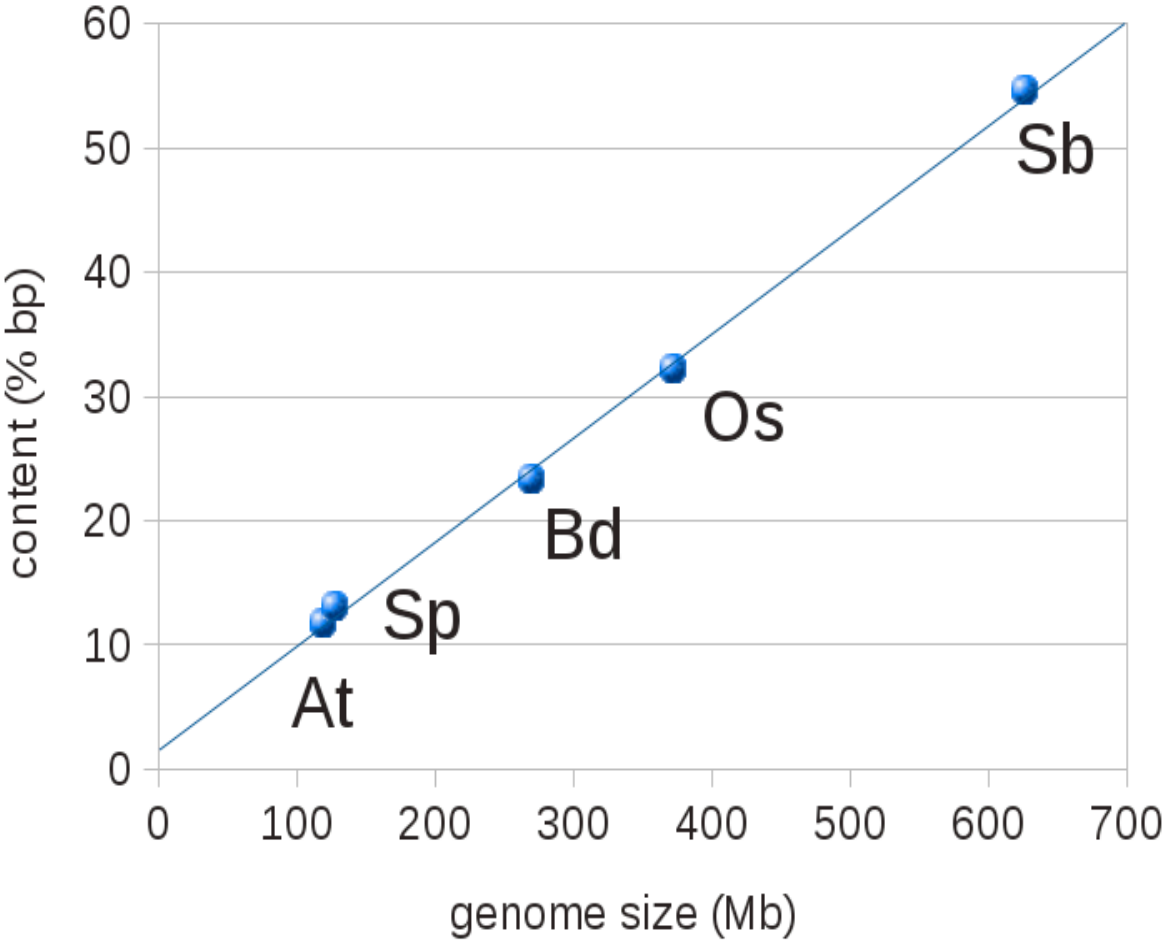


Figure S11: Linear dependency between genome size and LTR retrotransposon content for small sized (<< 1Gb) plant genomes.

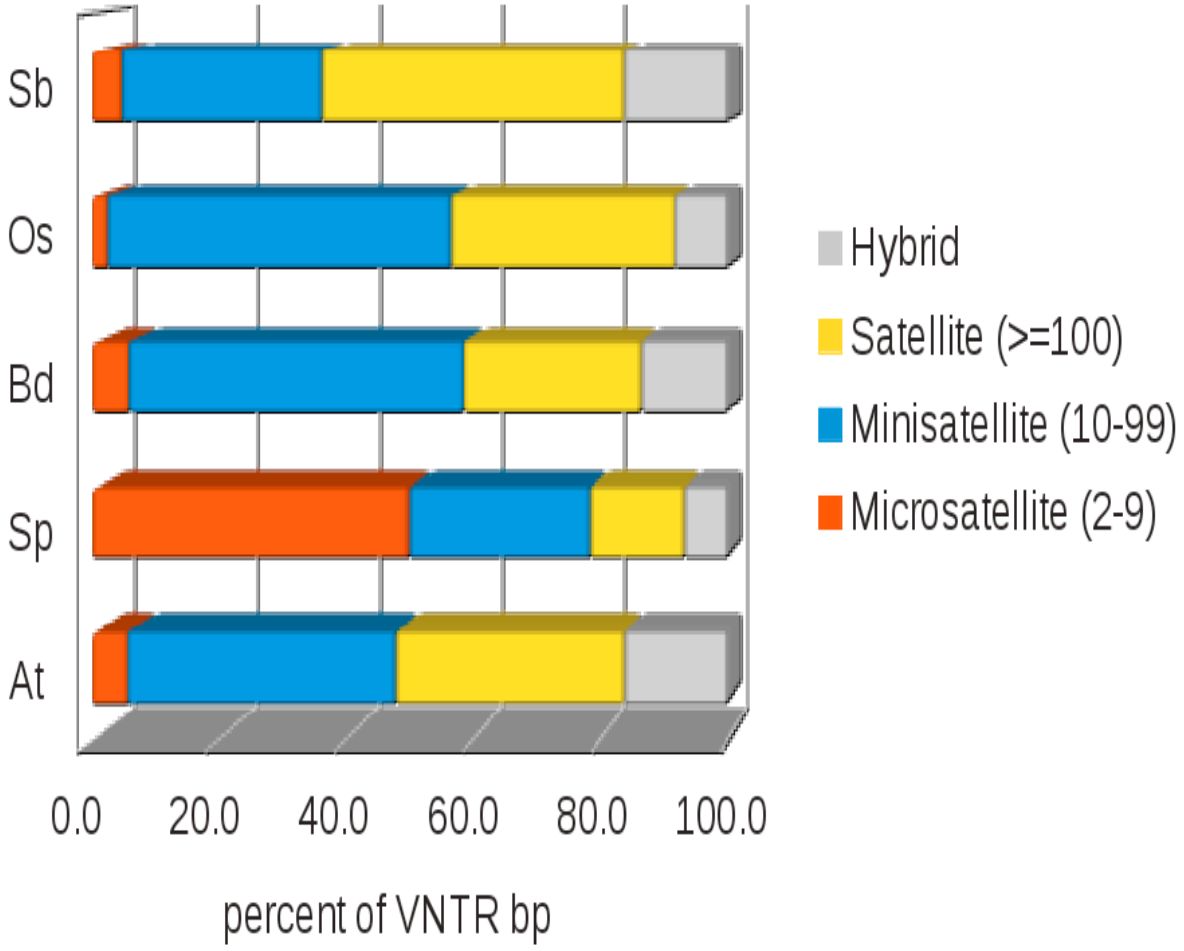


Figure S12: Composition of tandem repeat.

Comparative tandem repeat composition indicated that Spirodela has an exceptional high proportion of microsatellite tandem repeats: 50% versus 3 to 6% in four different reference genomes.

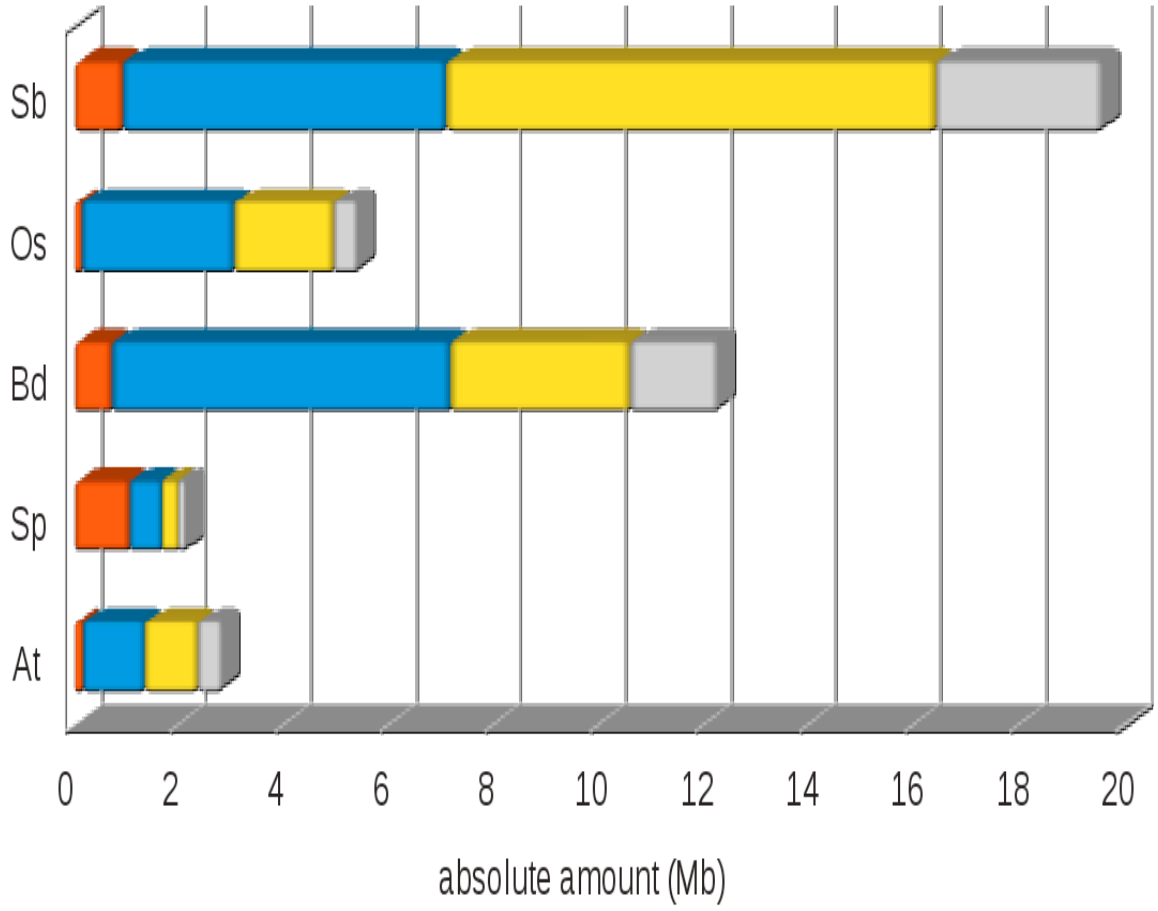


Figure S13: Amount of tandem repeats.

Comparison of tandem repeats indicated that Spirodela topped the list with 1 Mb over the much larger sorghum genome (0.9 Mb).

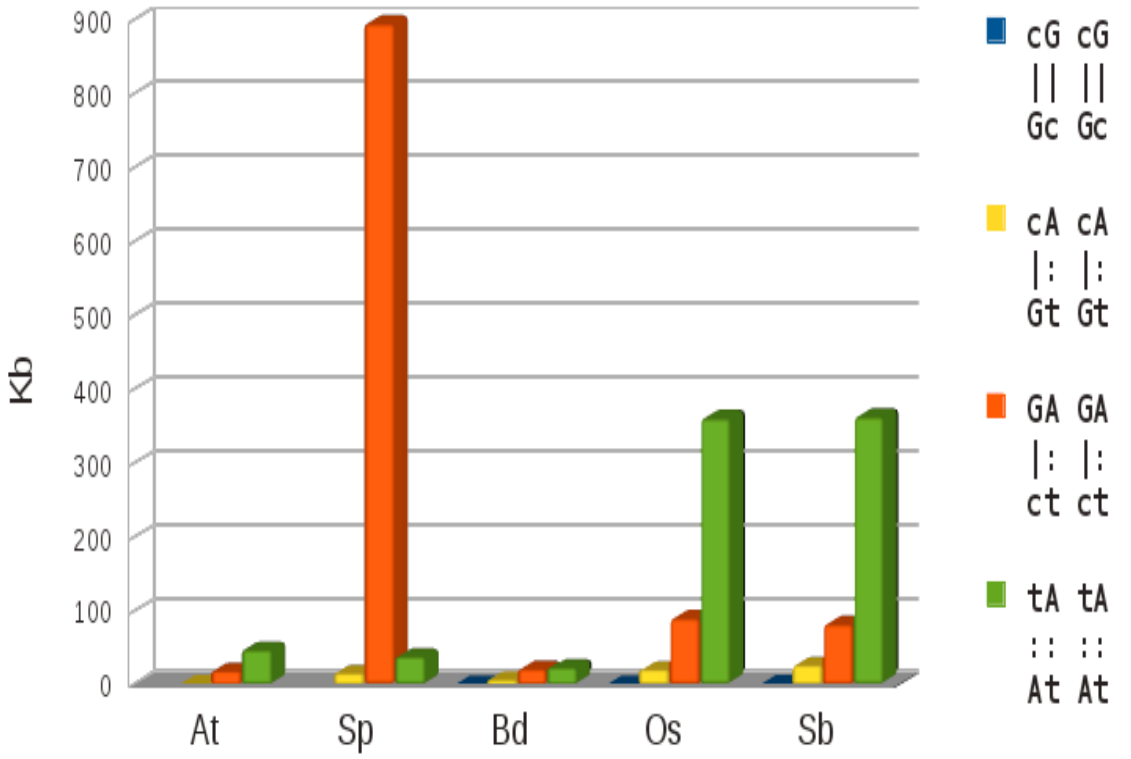


Figure S14: Distribution of the four possible types dinucleotide repeats.

The Spirodela genome assembly contains a disproportionally high amount (both in element number and total bp) of "GAGA" tandem repeats in comparison to other genomes. The dinucleotide repeat with the strongest binding (cG)n and 100% GC content is absent in At and Sp, the larger genomes contain only very small amounts of it. The usual order of abundance (cG)n, (cA)n, (GA)n and (tA)n is disrupted in Spirodela by the over 10-times increase of (GA)n, where one strand contains only purine the other only pyrimidine bases. lower case: pyrimidine bases, upper case: purine bases, |: 3 hydrogen bonds, |: 2 hydrogen bonds, the annotated dinucleotide repeats have a stretch length of at least 25 bp (default parameter of Tandem Repeats Finder).

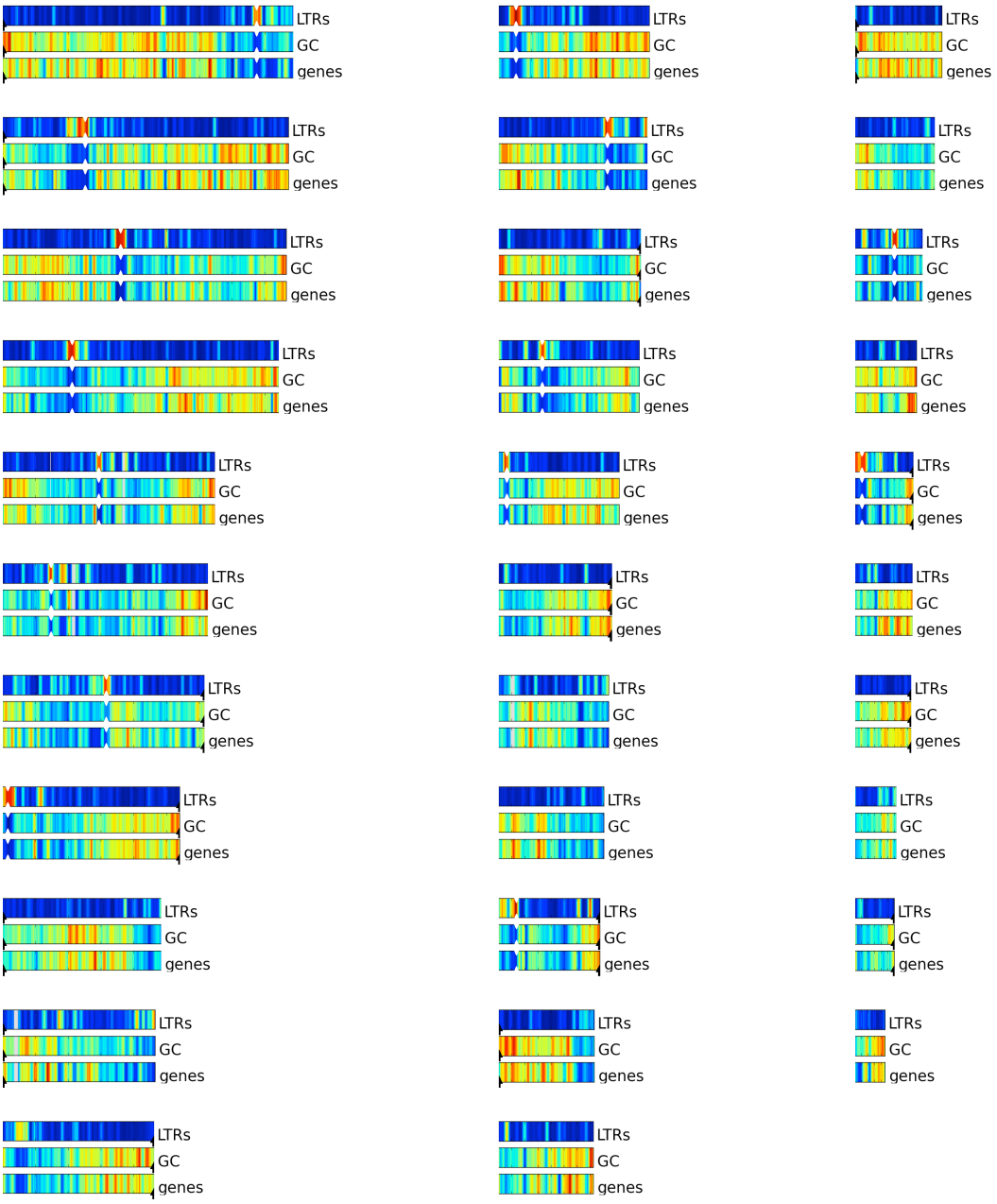


Figure S15: Heatmap for the 32 pseudomolecules.

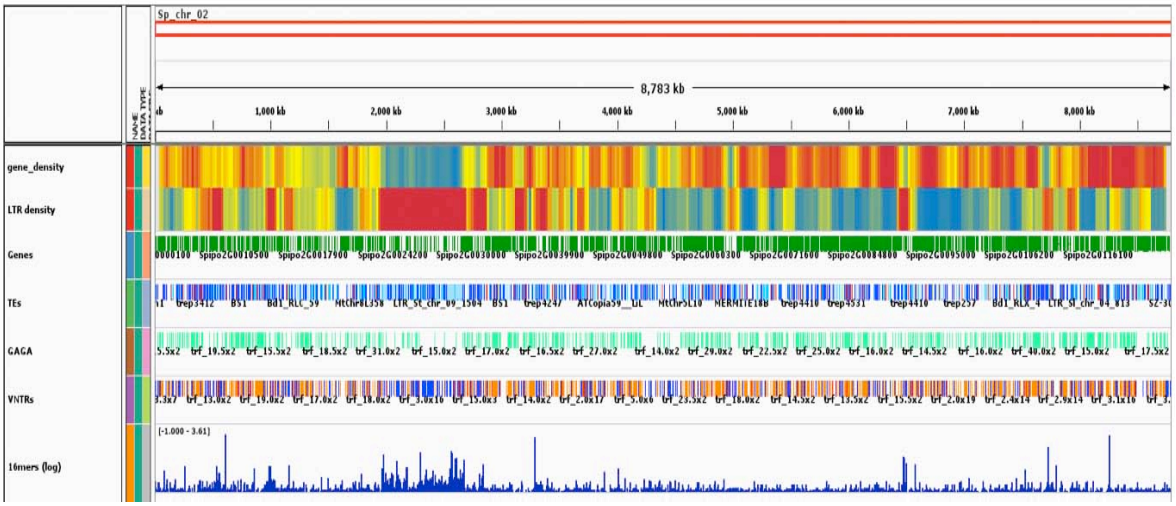


Figure S16: The overview of heatmap for pseudomolecule 2.

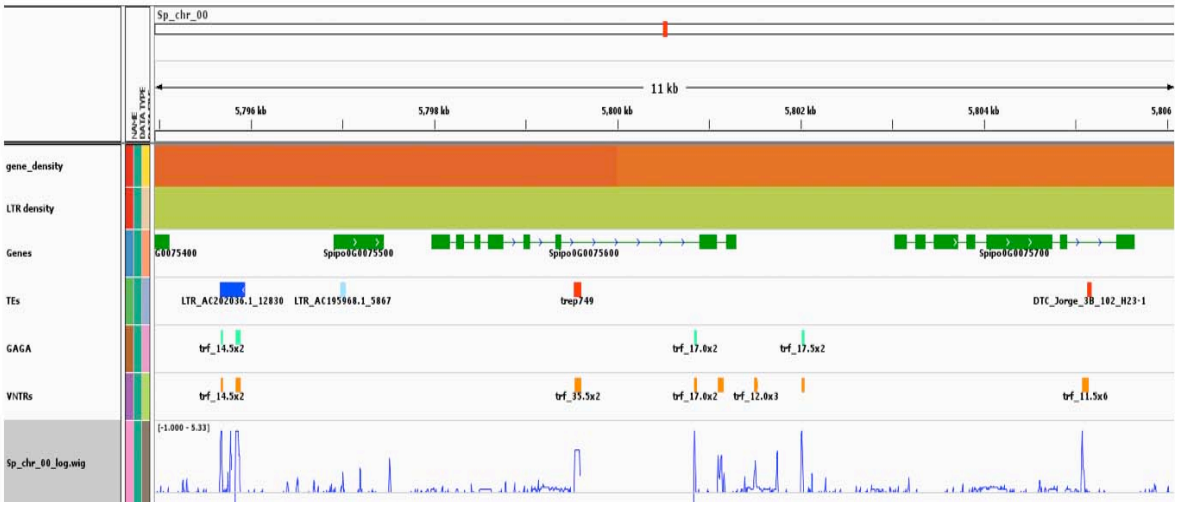


Figure S17: The overview of heatmap for zoomed in part of pseudomolecule 0.

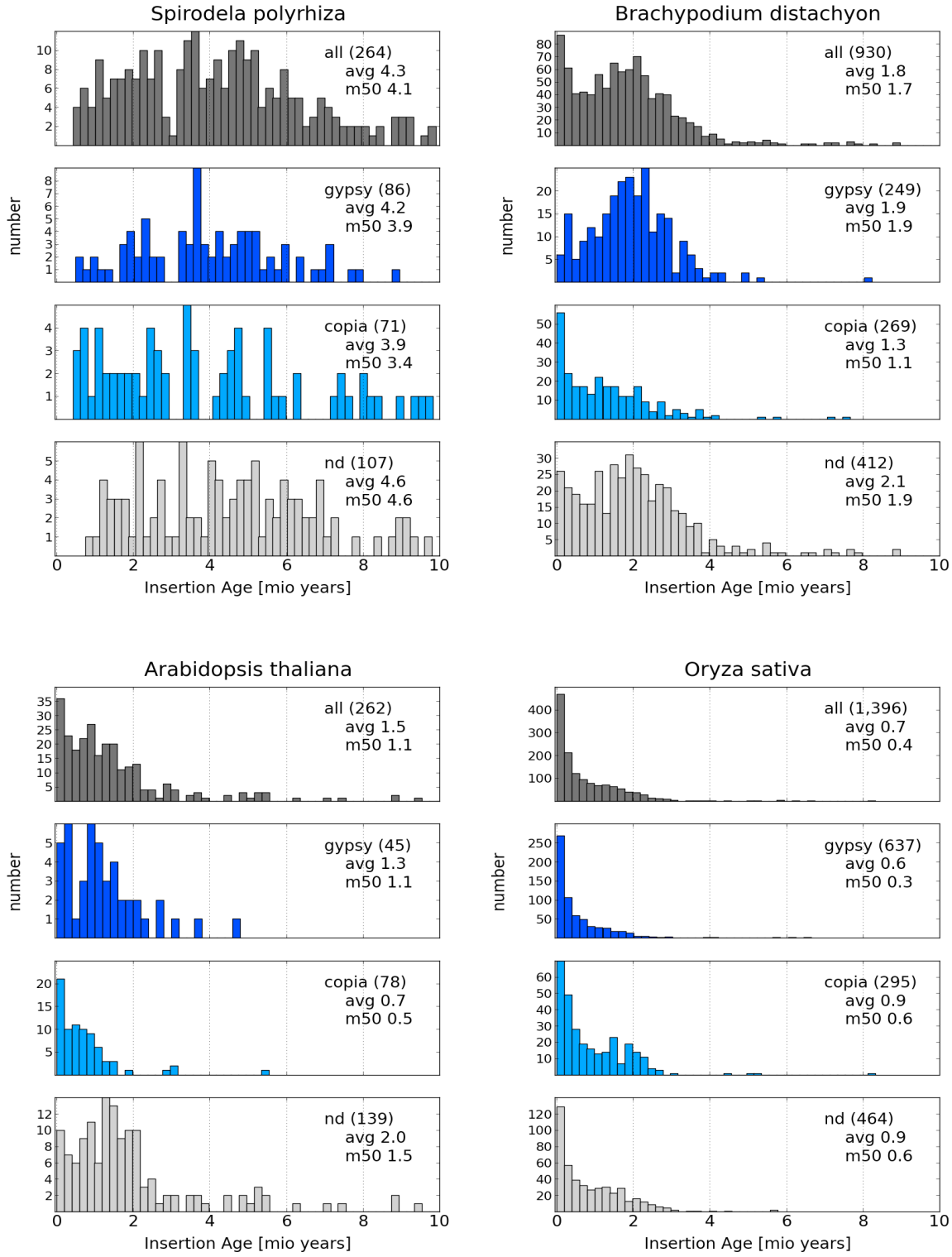


Figure S18: Comparison of LTR insertion age distributions.

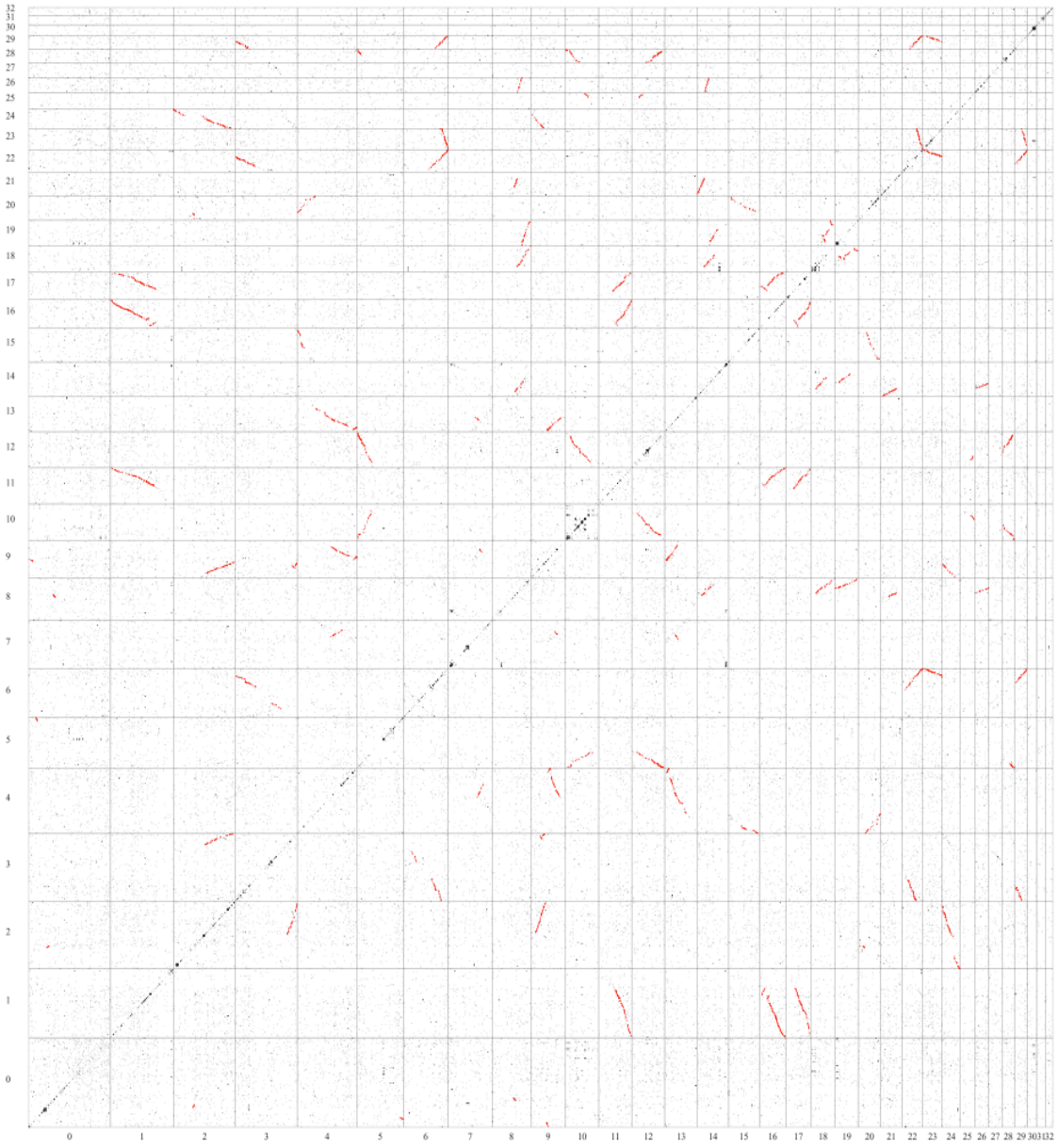


Figure S19: Dot plot of intragenomic duplications in Spirodela.

The top 20 blastp hits for each gene of an intra-genomic blastp comparison of Spirodela are plotted. Matches of selected duplicated regions are shown in red.

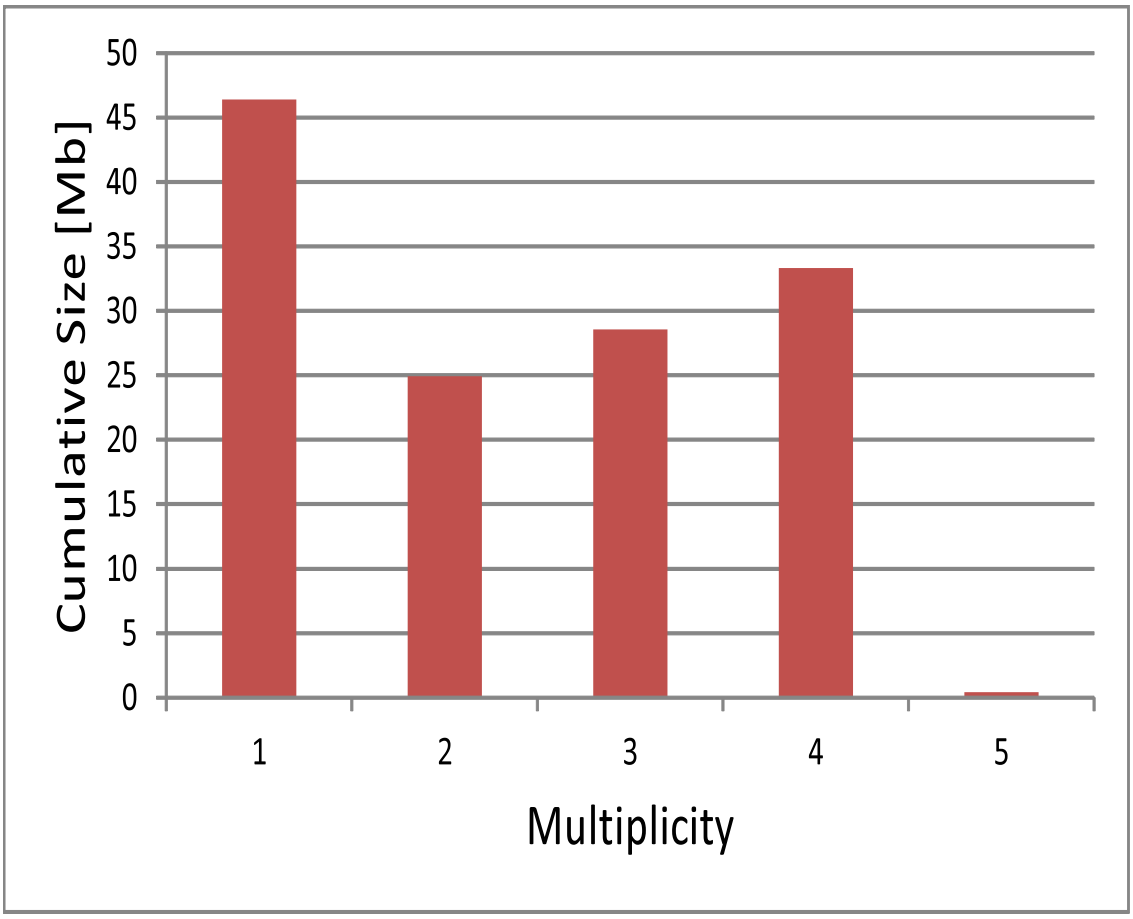


Figure S20: Distribution of copy numbers of duplicated chromosomal segments in Spirodela.

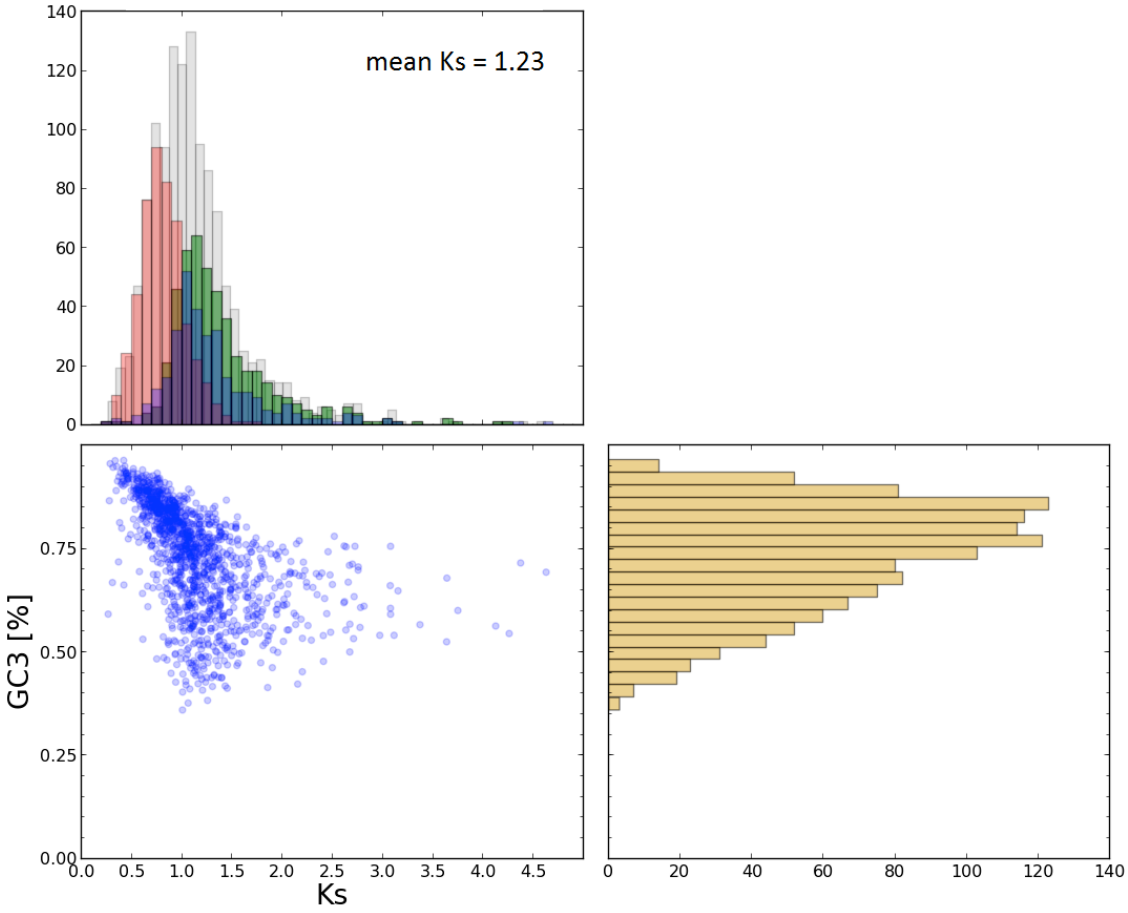


Figure S21: Distribution of Ks-values of syntenic gene pairs of duplicated segments in Spirodela and relation to GC3 composition.

Top figure shows the histogram of synonymous substation rates Ks of paralogous Spirodela gene pairs located in duplicated segments (grey bars), figure at the bottom right the histogram of the GC3 gene composition. A clear negative correlation is observed between GC3 and Ks in the figure left bottom. Ks-values for gene pairs were also separately plotted for GC3-high (GC3 > 75% for both genes, red bars), -medium (exactly one gene with GC3 > 75%, blue bars) and -low (GC3 ≤ 75% for both genes, green bars) pairs.

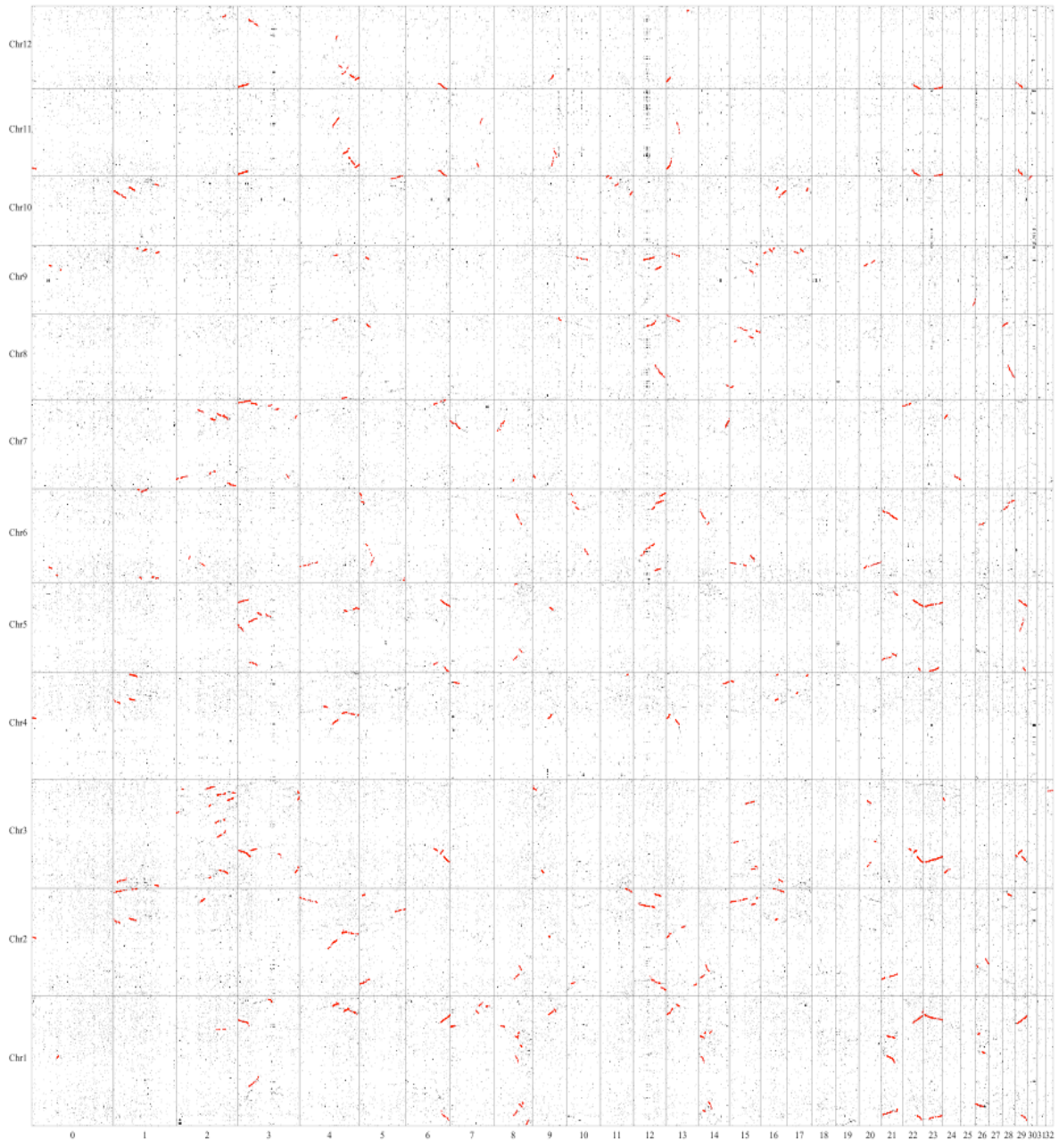


Figure S22: Dot plot analysis between Spirodela (horizontal axis) and Rice (vertical axis).

Genome and chromosome sizes of rice and Spirodela are normalized to equal axis sizes. Red dots show the reported syntenic segments between the two species.

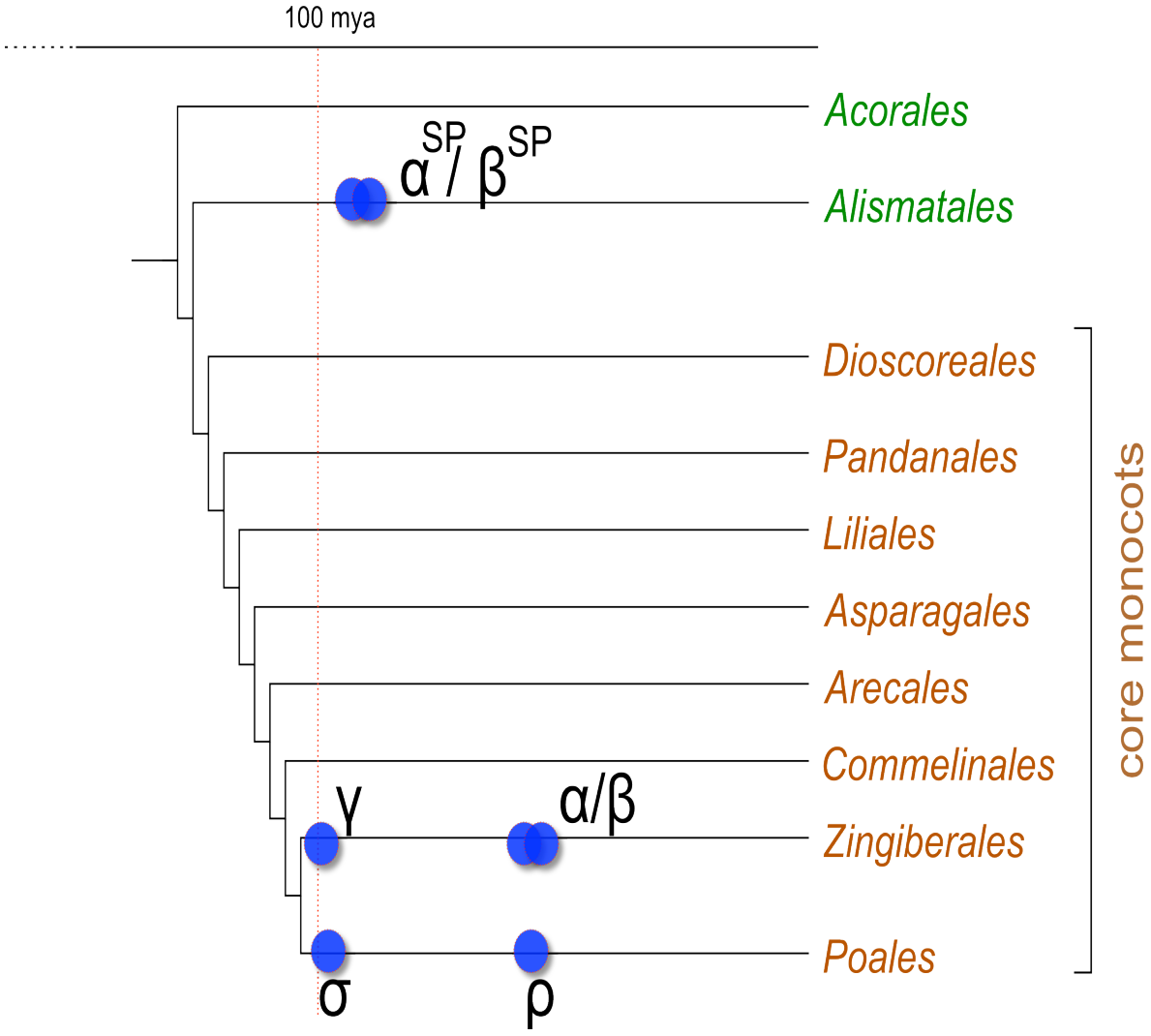


Figure S23: Phylogeny of monocotyledonous orders.

The dendrogram is a simplified version redrawn from (Janssen and Bremer 2004). Core monocots are shown in brown, known WGDs are shown as blue circles. Details about their placements are given in the supplementary text (4.3).

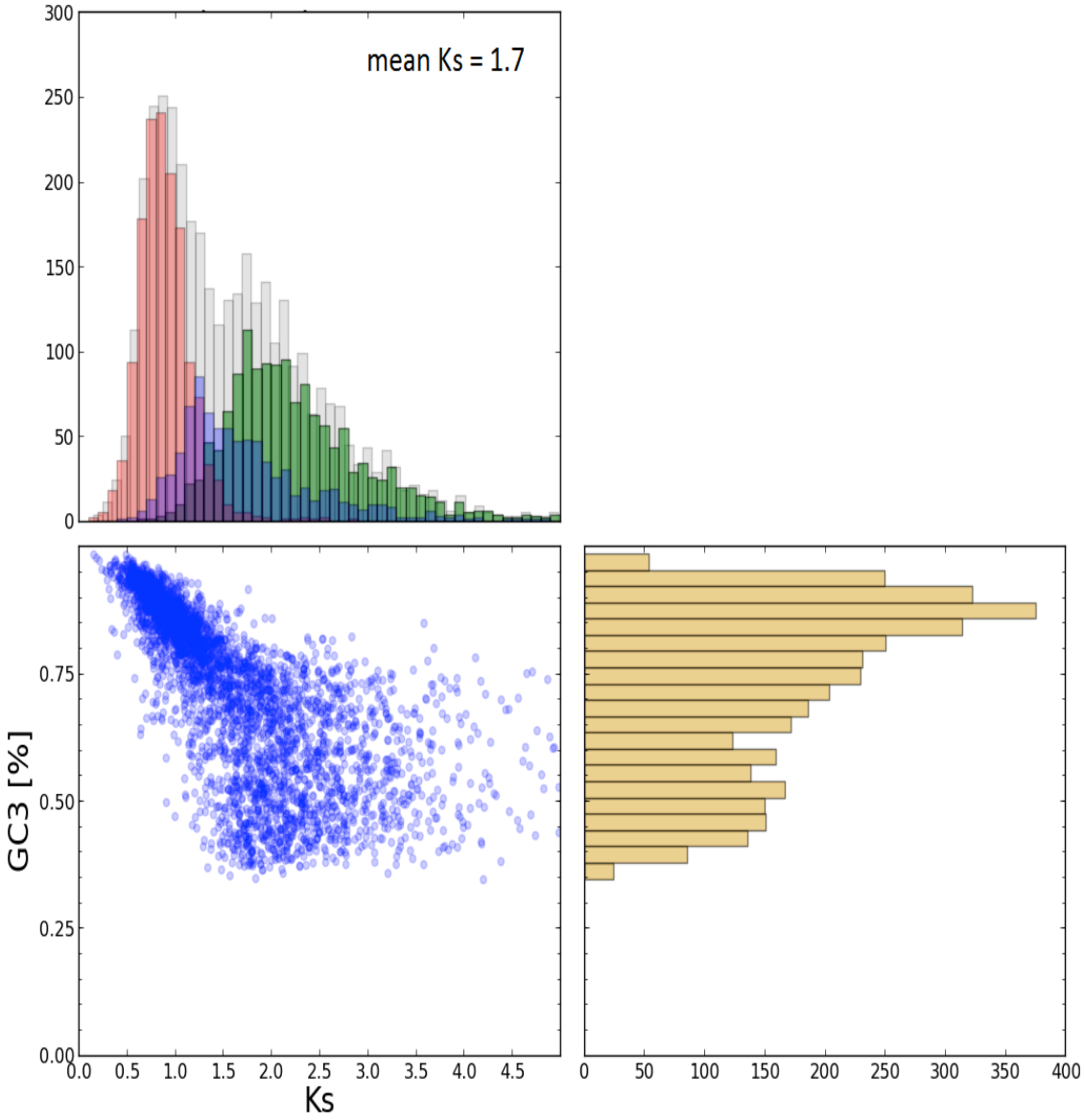


Figure S24: Ks values of syntenic gene pairs between rice and Spirodela and their dependency on GC3 composition.

Synonymous substiution rates of syntenic Spirodela-rice gene pairs. The bimodal Ks-distribution can explain the negative correlation of GC3 and Ks.

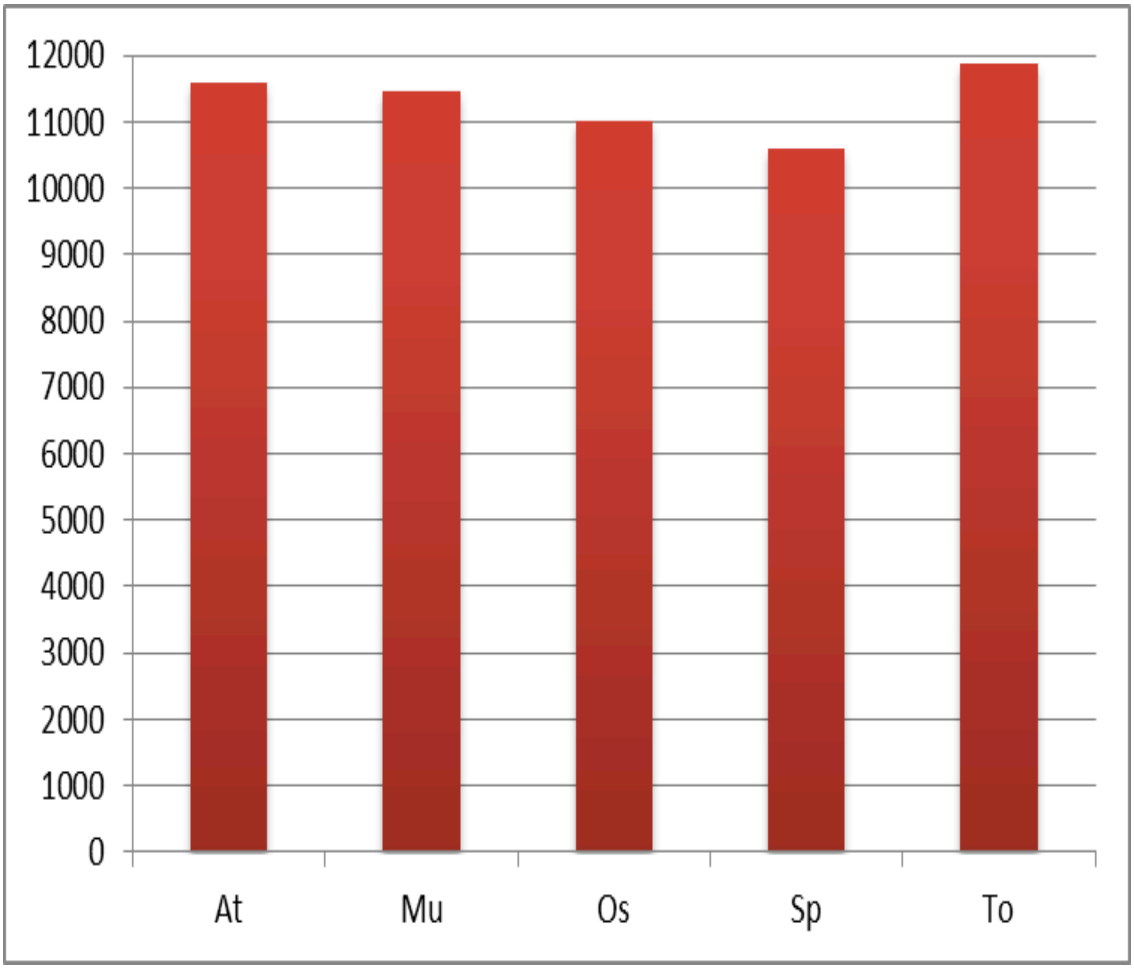


Figure S25: orthoMCL cluster per species.

Figure shows for each species (At: Arabidopsis, Mu: banana, Os: rice, Sp: Spirodela, To: tomato) the number of orthoMCL families that contain at least one gene of the species. Cluster class 1- the species-specific divisions, were excluded from counting.

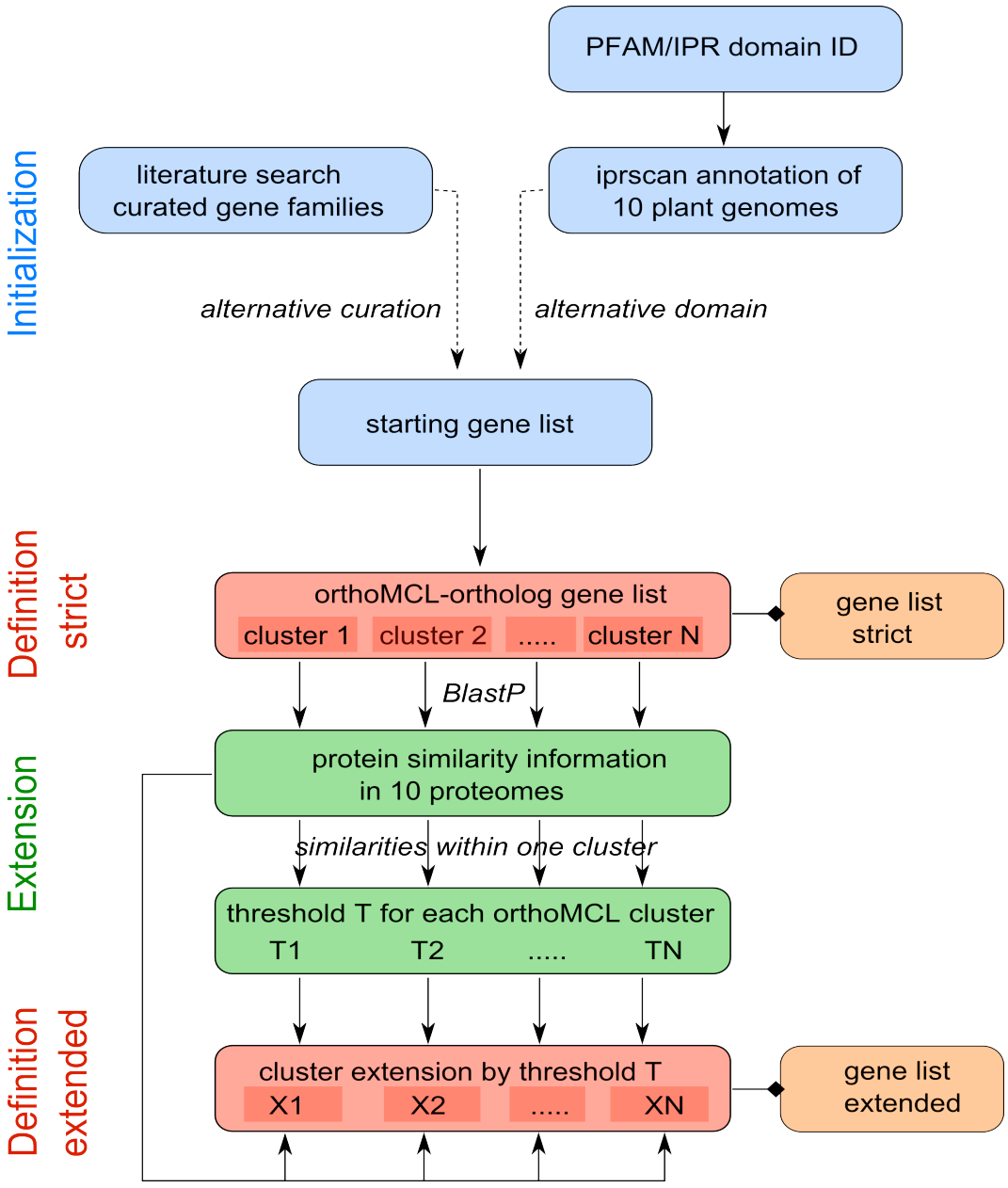


Figure S26: Schema for determination of gene families.

Starting from 2 alternative selection modes, an initial list of gene identifiers is defined either by a published data set or by the occurrence of a specific domain identifier. The starting set was adjusted by comparison to orthoMCL clusters and putatively extended by a blast search.

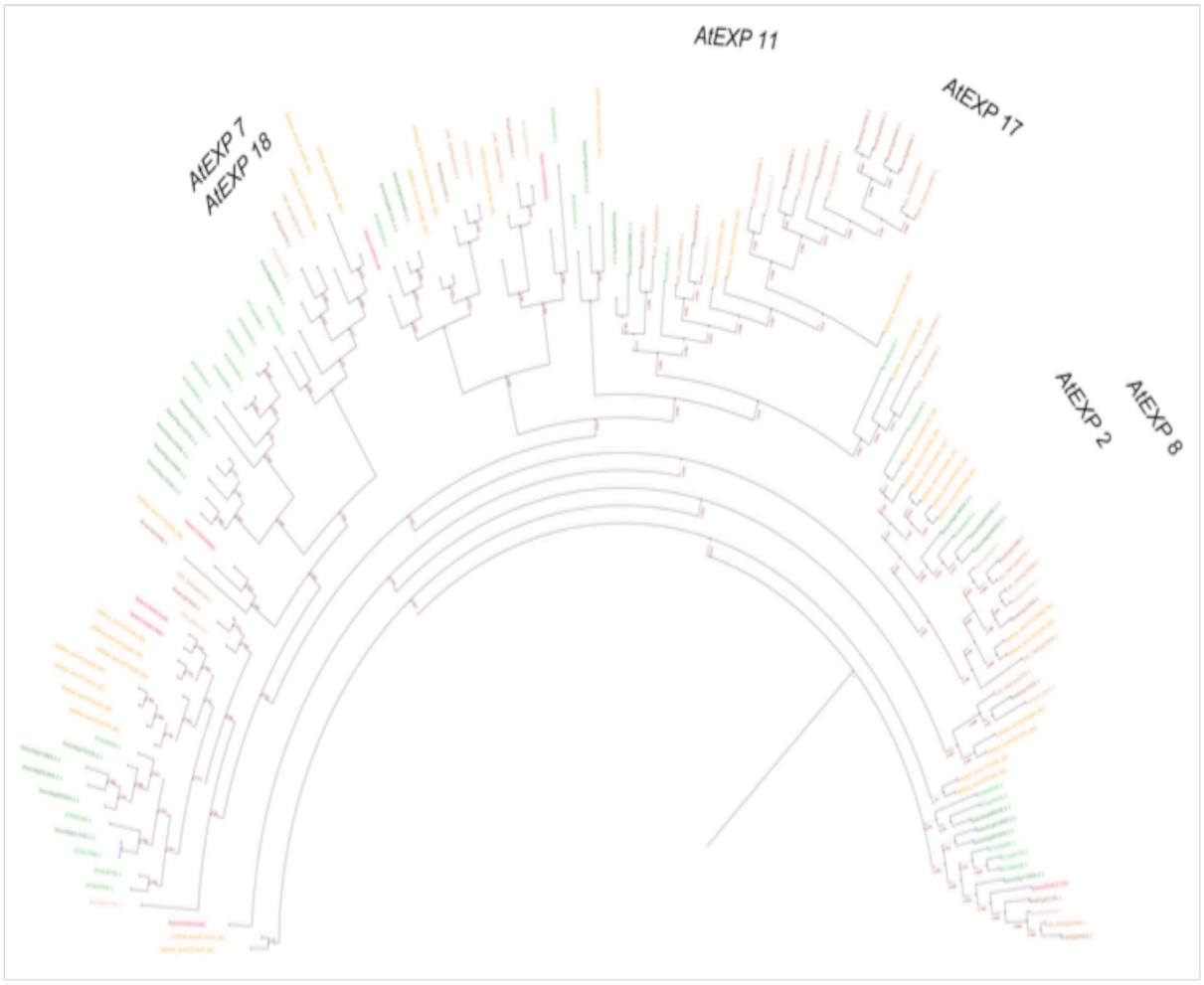


Figure S27: Phylogenetic tree of the α-expansin family.

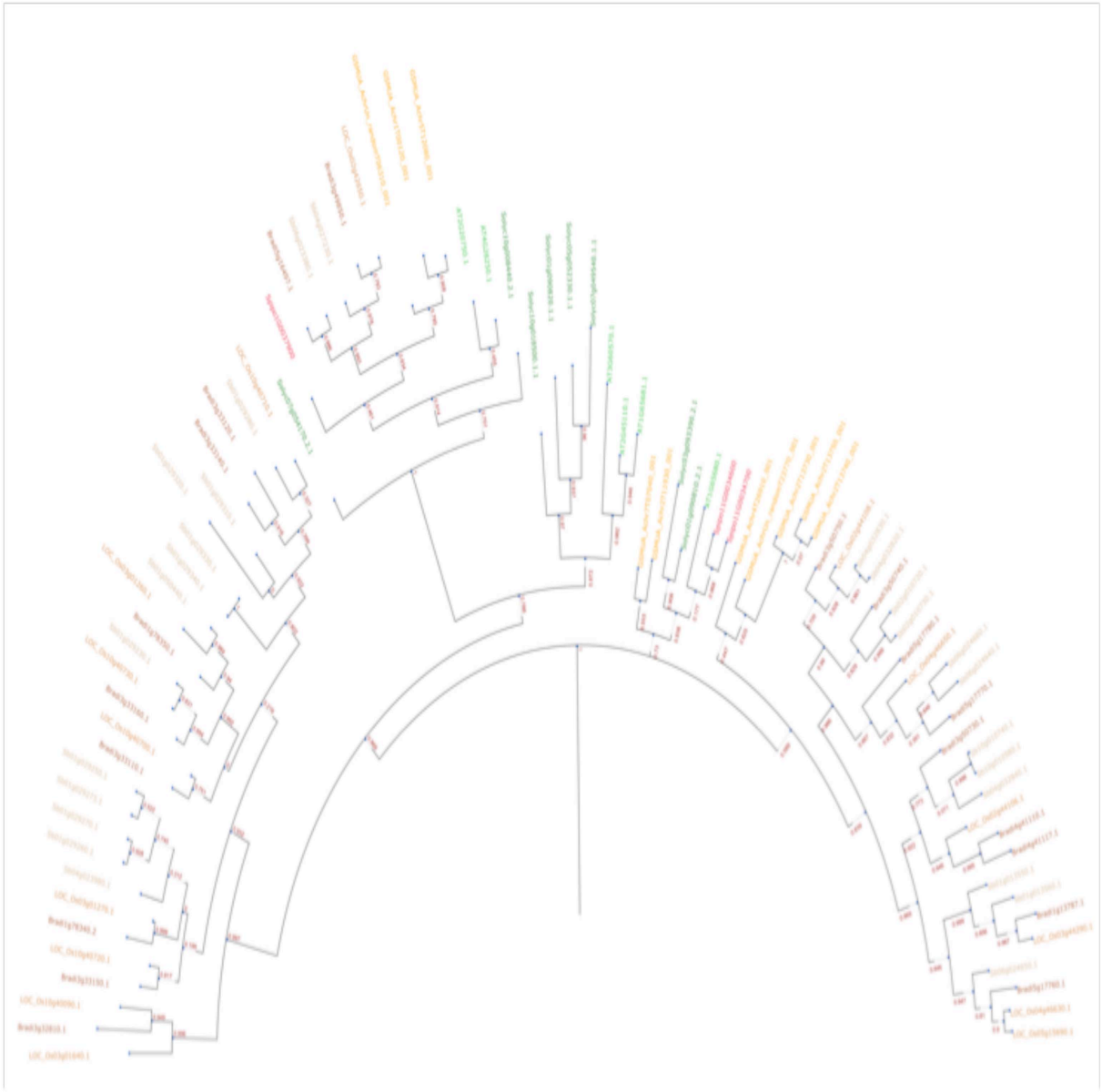


Figure S28: Phylogenetic tree of the β-expansin family.

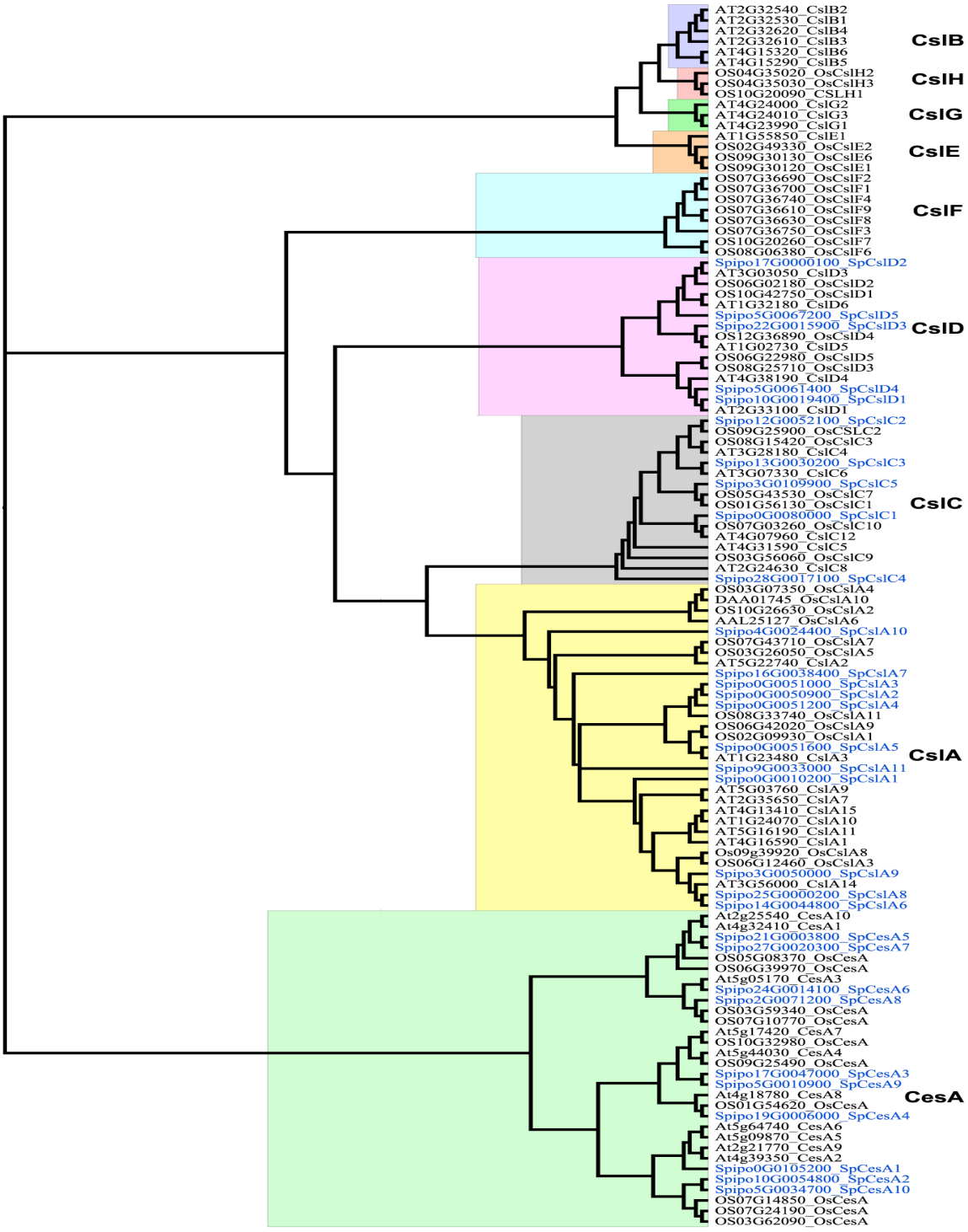


Figure S29: Phylogenetic tree for the cellulose synthase consisting of CesA and Csl sequences from Arabidopsis, rice and Spirodela.

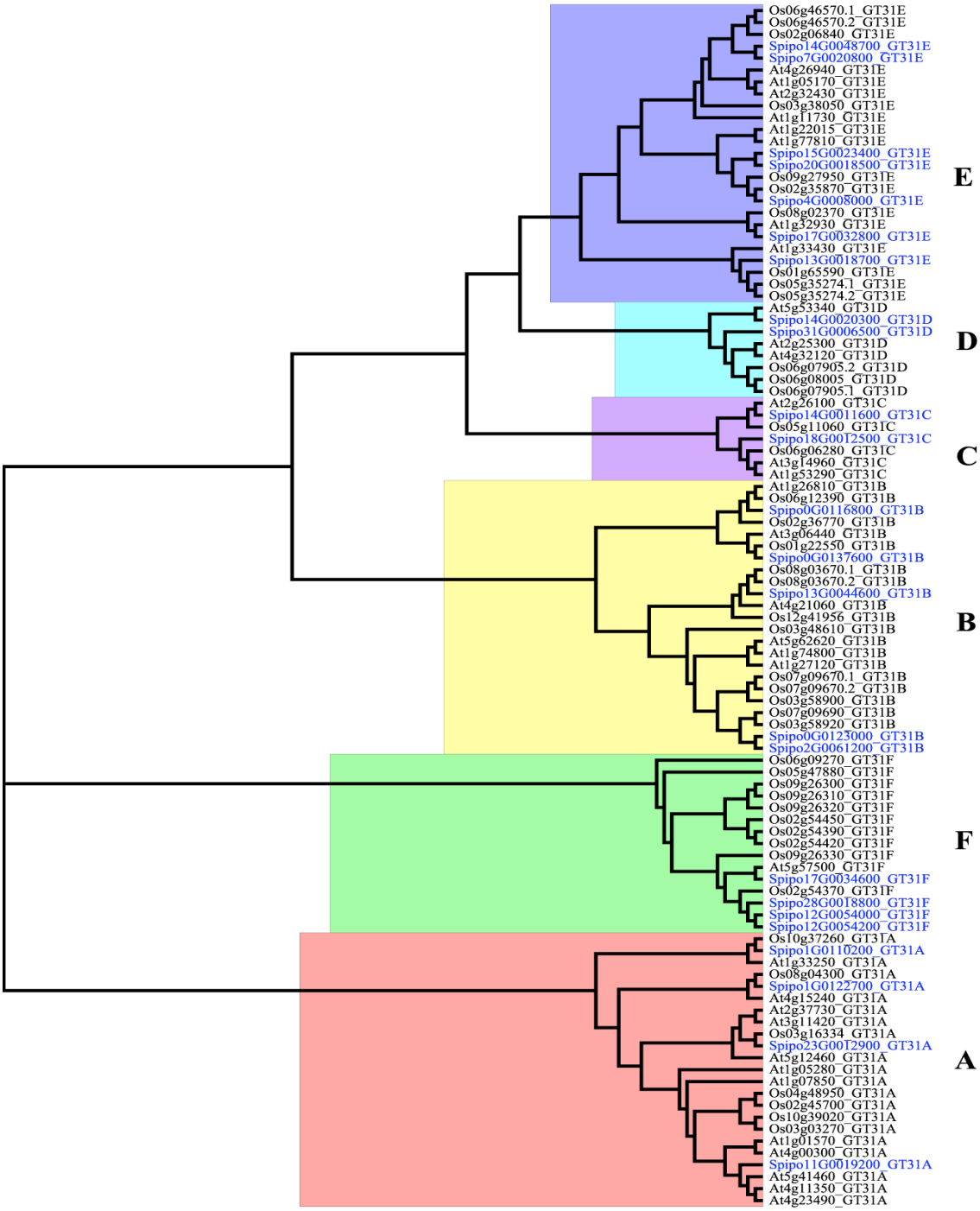


Figure S30: Phylogenetic tree for the family GT31 from Arabidopsis, rice and Spirodela.

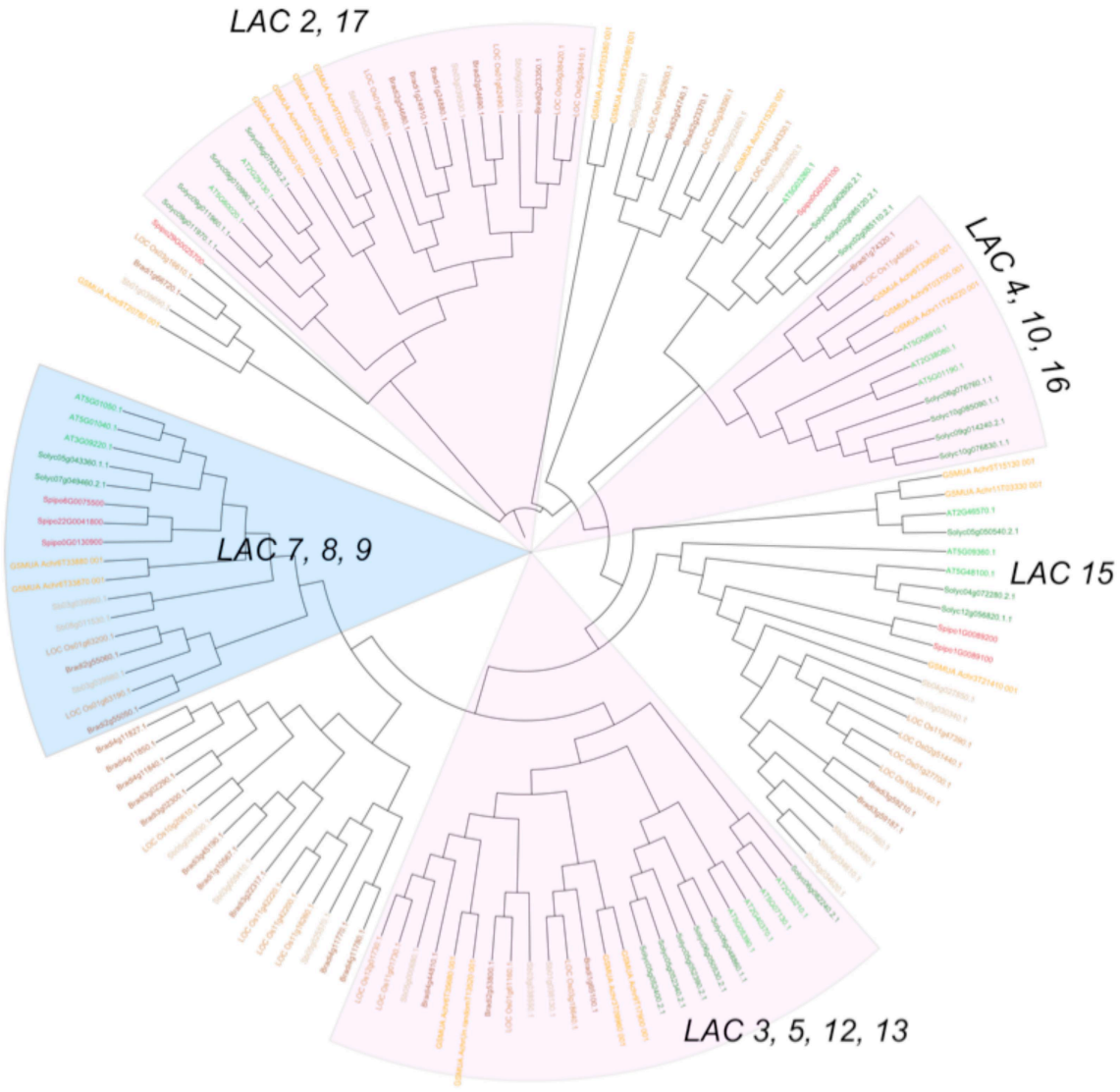


Figure S31: Laccase Gene Family.

Dendrogram of the laccase family in Spirodela (red), banana (orange), *Poaceae* (brown shades: Sorghum, *Brachypodium* and Rice), *Arabidopsis* (light green) and tomato (dark green). LAC1 is missing in this analysis because it represents a highly divergent clade in the Arabidopsis laccase family (Turlapati, Kim et al. 2011). Clade of AT5G48100 *(TT10/LAC15*) is shown at the right of the tree indicating doubling in Spirodela and higher expansion in *Poaceae*. Clade comprising LAC7/8/9 (highlighted in blue) shows an expansion of Spirodela genes relative to other monocotyledonous genes. Three other groups indicating loss or strong reduction of Spirodela members are highlighted in red.

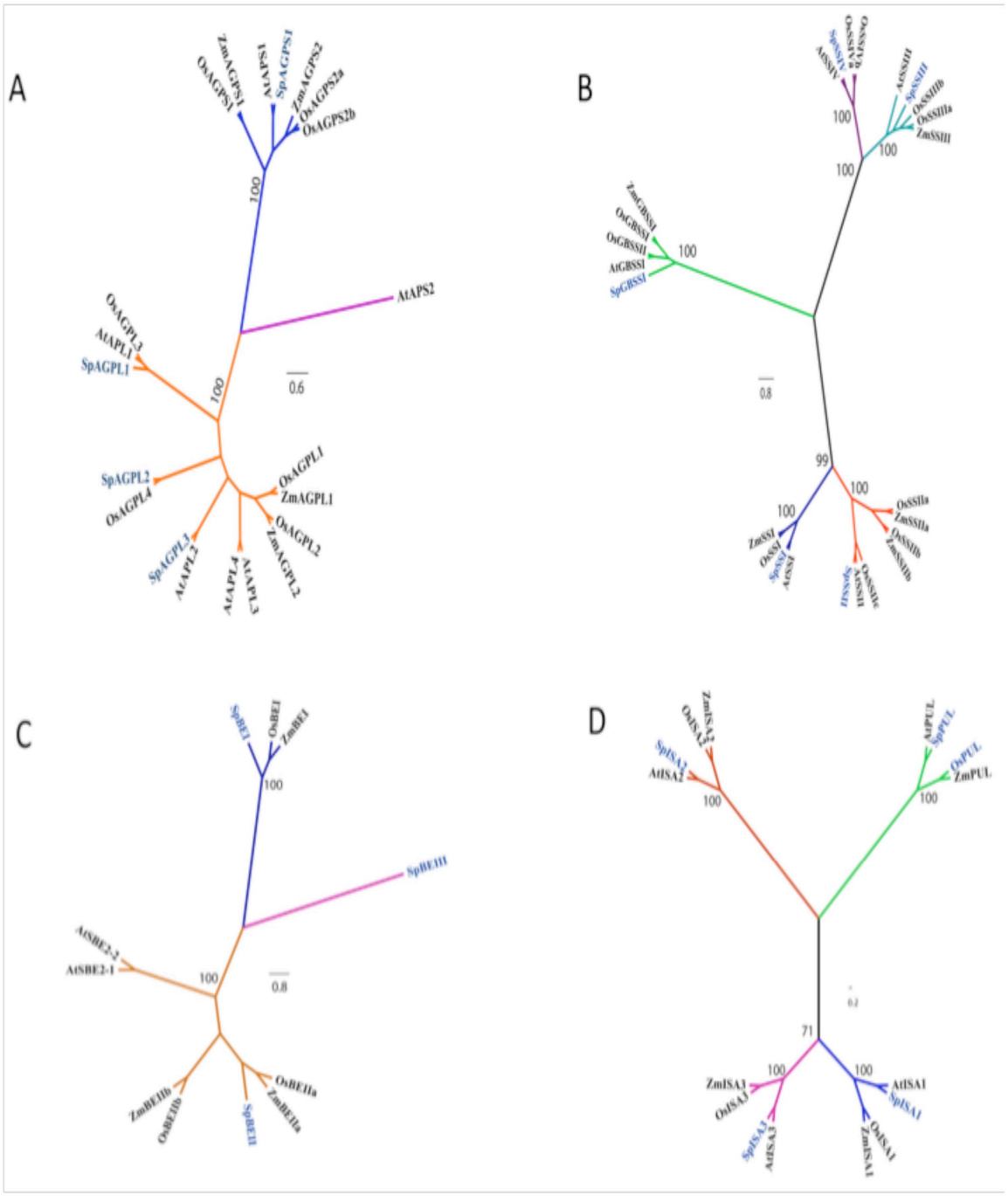


Figure S32: Starch biosynthesis genes in Spirodela and their orthologs in maize, rice and Arabidopsis.

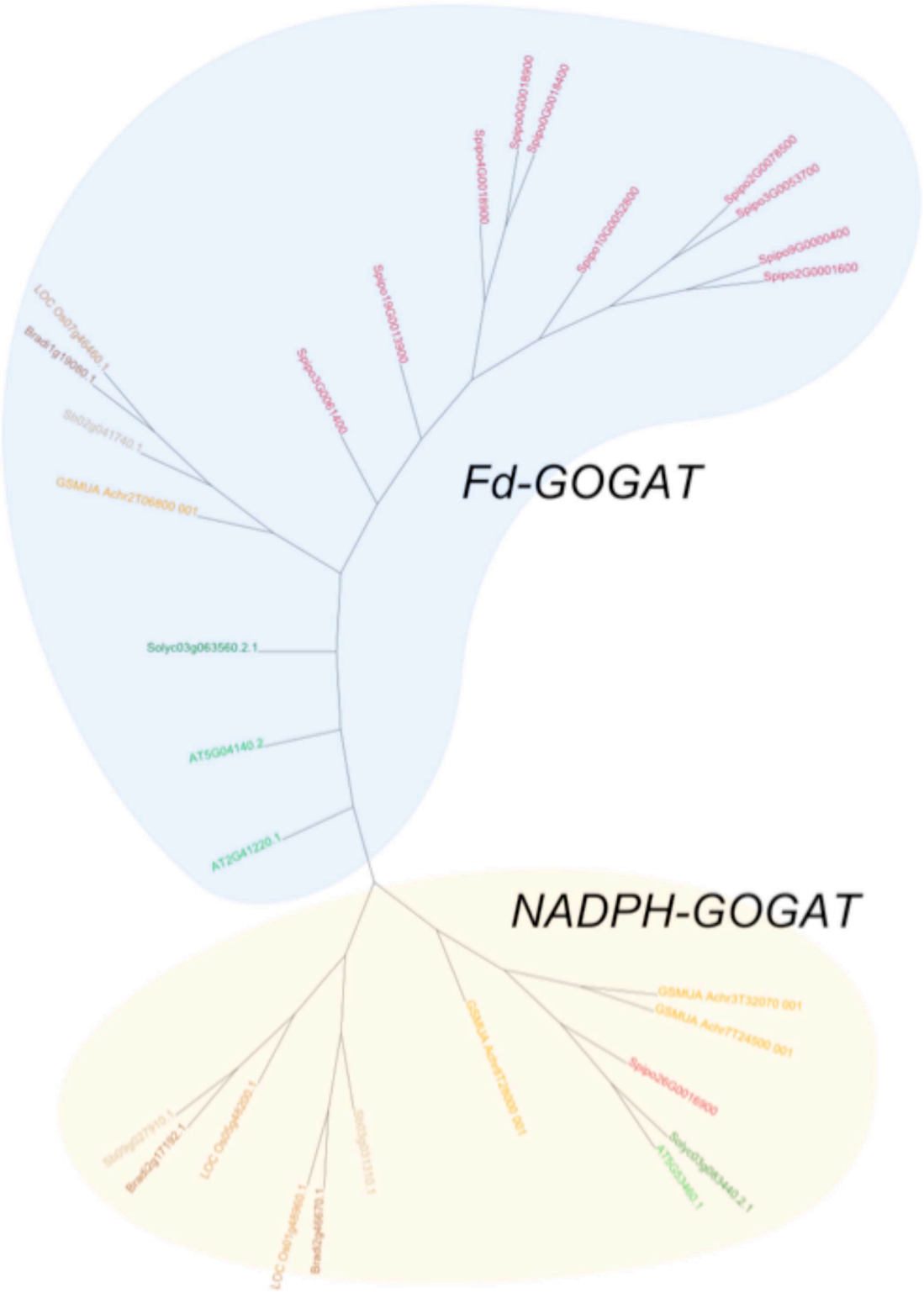


Figure S33: Dendrogram of glutamate synthase isoforms in rice, Brachypodium, sorghum, banana, tomato, Arabidopsis and Spirodela.

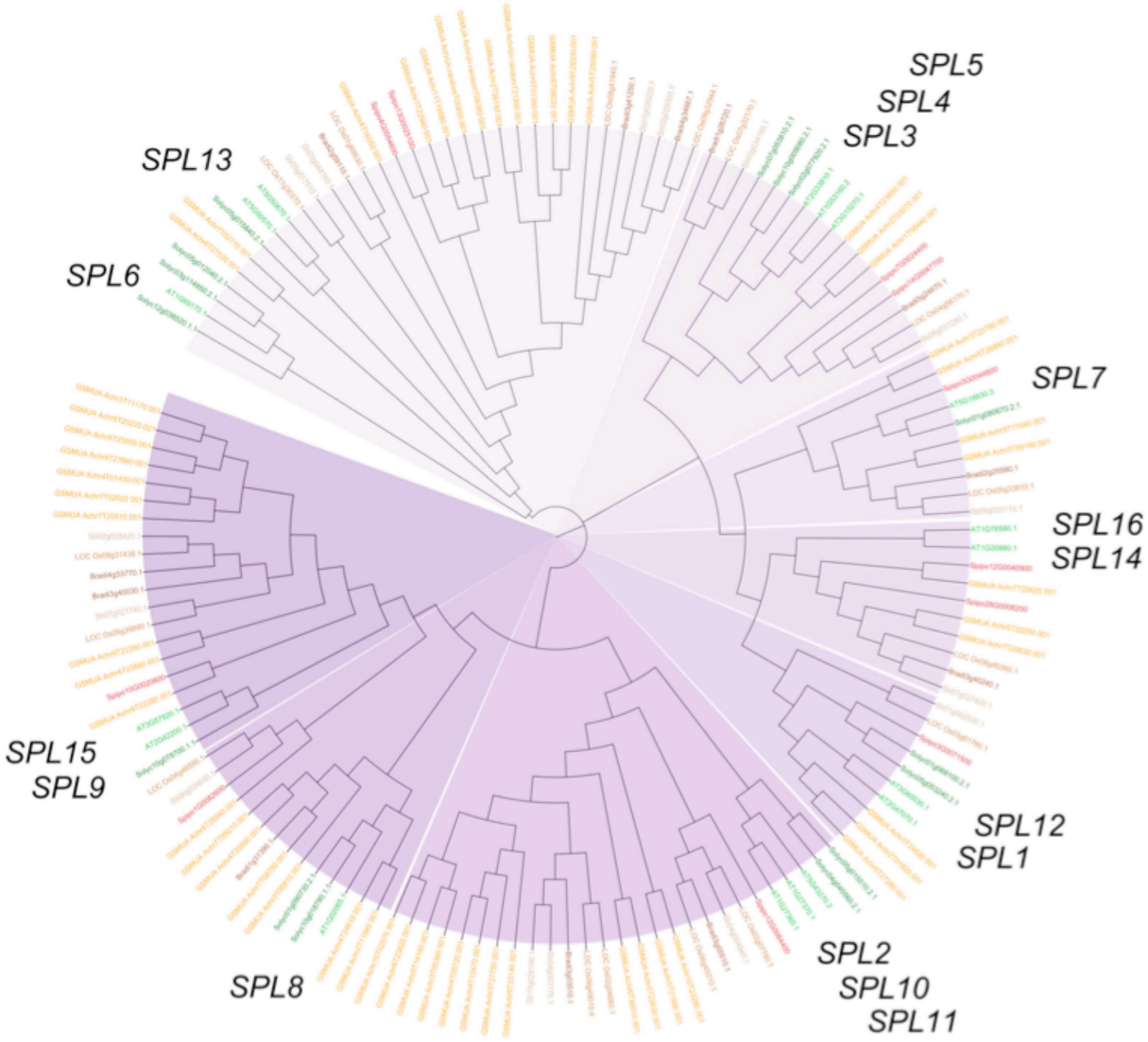


Figure S34: The SBP-like protein family.

For SPL9 and 15 as well as the clade containing SPL2, 10 and 11 we found only one Spirodela member while all other species contain at least two genes of each subfamily. However, for another subfamily reported to promote the transition from the juvenile to adult phase, SPL 3, 4 and 5, Spirodela had no reduced gene count: the two dicots and banana contained each 3 members while Spirodela had – like the three grass species – two copies in the genome.

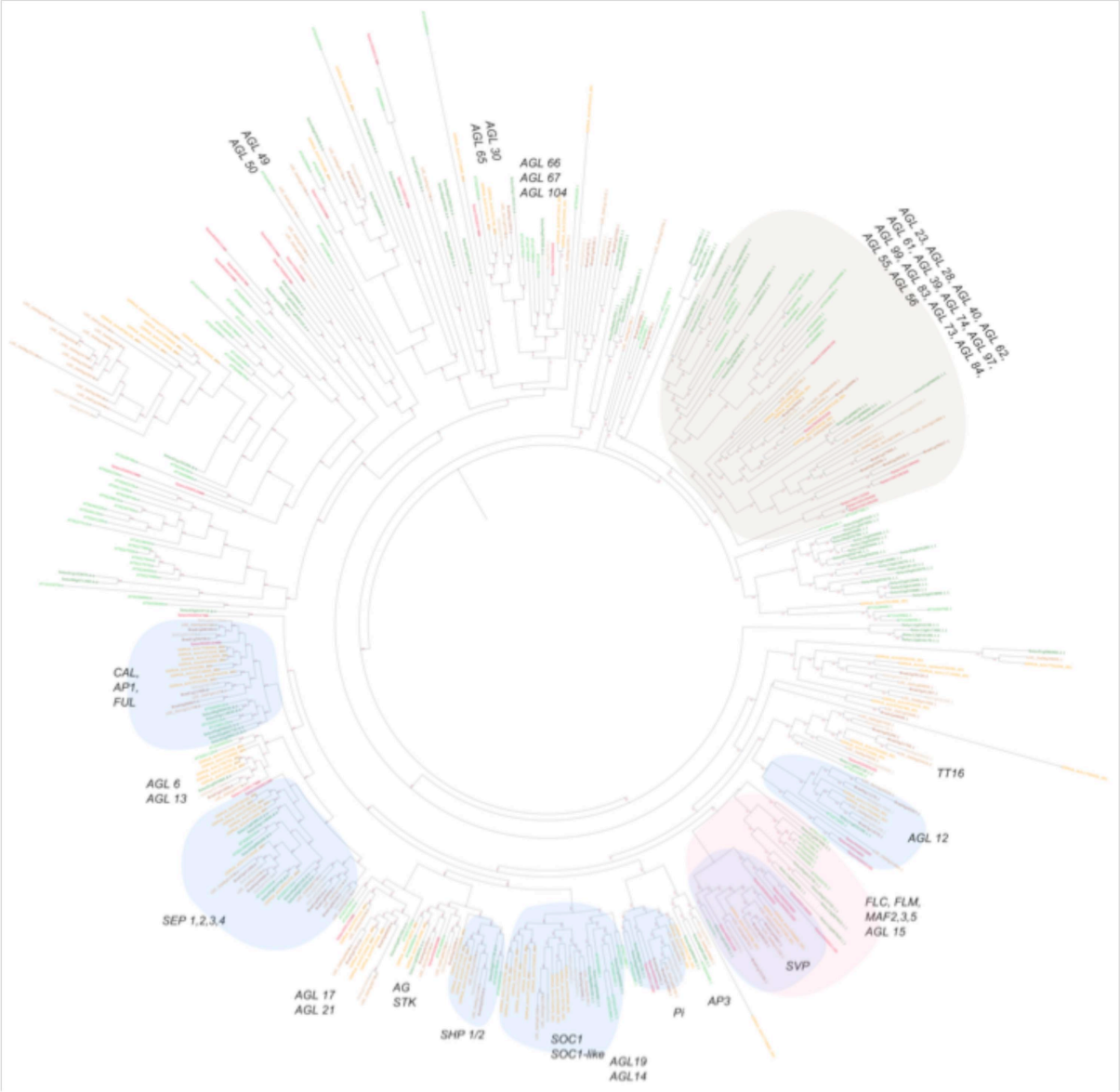


Figure S35: The MADS-box gene family.

A dendrogram of the MADS-box family of seven species: Arabidopsis (light green), tomato (dark green), rice (light brown), sorghum (tan), Brachypodium (dark brown), banana (yellow) and Spirodela (red). A detailed description of selected subfamilies in Spirodela in comparison to other species is provided in Supplement section 5.4.6.

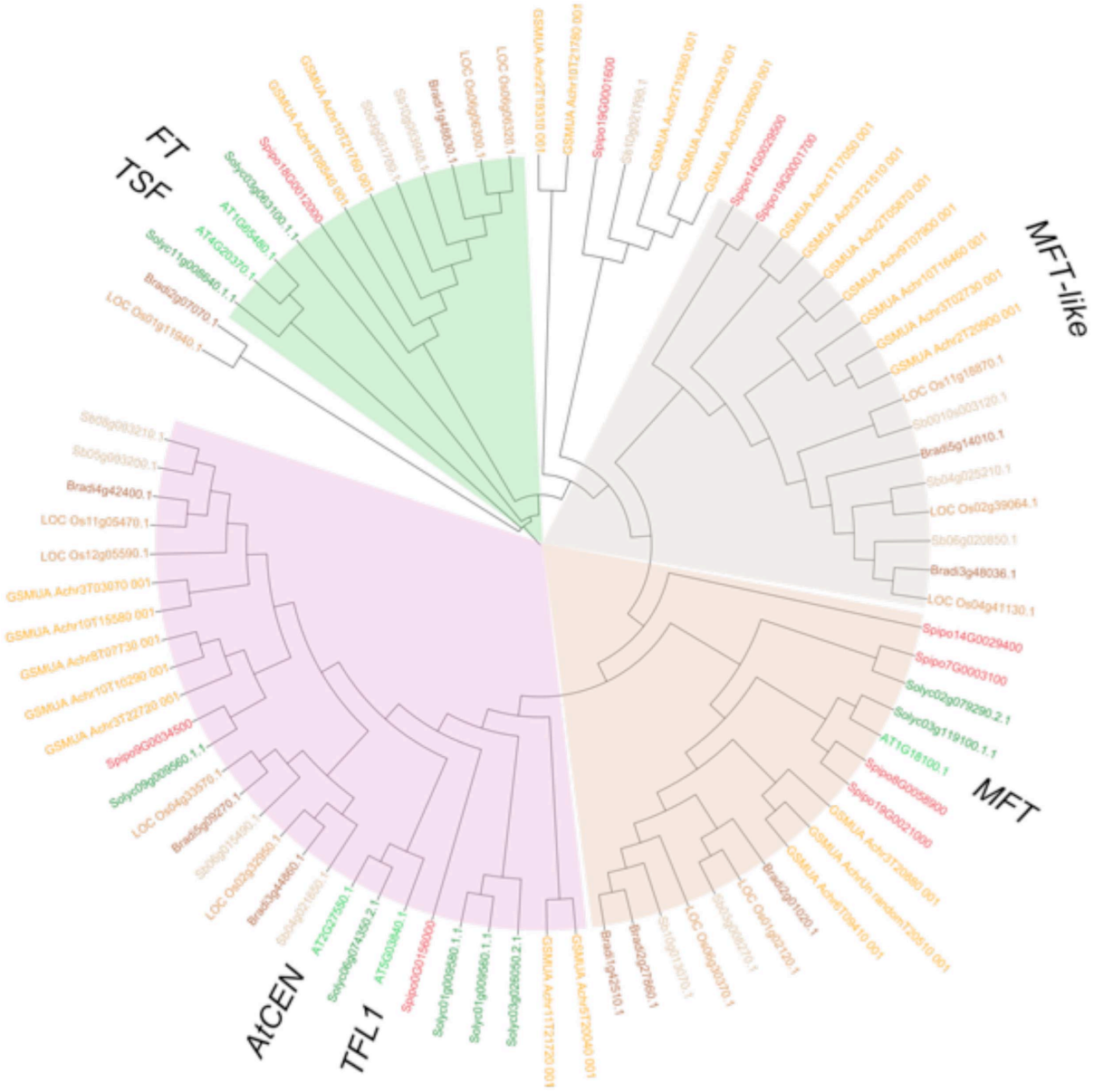


Figure S36: The Phosphatidylethanolamin-binding protein (PEBP)-family.

Supplementary Tables

Sequencing method	Library	Insert size (bp)	Sequence depth[a]	Average length (bp)	# Reads
Sanger reads	BACs	80K-120K	0.18x	102,085 +/- 25,521	29,410
Sanger reads	Fosmid	40 K	0.57x	38,026 +/- 9,506	119,482
LS454	GZWP.pairs	5 K	0.93x	4,771 +/- 1193	1,361,517
LS454	GZUG.linear	-	10.61x	n/a	4,652,570
LS454	GZUH.linear	-	10.24x	n/a	4,356,540
Total			22.53x		10,519,519

Table S1: Total sequence input for the genome assembly.

Assembly	
Number of scaffolds	1,071
Number of bases	145,095,613
Average scaffold size	135,476
N50 scaffold size	3,759,109
Number of contigs	16,055
Number of bases	132,026,954
Average contig size	8,221
Max contig size	124,558
N50 contig size	14,532
Min contig size	1,000

Mimimum scaffold length	Scaffolds #	Total scaffold length	Contigs #	Total contig length
All	1,071	145,095,613	16,055	132,026,954
1kb	1,071	145,095,613	16,055	132,026,954
2.5kb	699	144,272,137	11,294	124,350,168
5kb	517	143,680,478	7,846	111,798,090
10kb	252	141,803,455	4,217	85,815,699
25kb	71	139,269,830	937	35,263,298
50kb	62	138,956,030	150	9,250,503
100kb	60	138,826,889	2	227,753
250kb	53	137,782,056		
500kb	50	136,619,593		
1mb	38	128,614,578		
2.5mb	24	103,288,229		
5mb	9	53,130,157		

Table S2: Statistics of Spirodela genome assembly.

The unit of scaffold and contig size is base pair (bp).

Pseudo	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	0
Size (Kb)	8,941	8,796	8,736	8,492	6,553	6,333	6,239	5,477	4,925	4,726	4,688	4,671	4,624	4,409	4,370	3,759	11,607
Pseudo	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	Total
Size (Kb)	3,541	3,441	3,286	3,178	2,998	2,983	2,725	2,515	2,127	1,959	1,859	1,843	1,793	1,340	1,270	994	145,199

Table S3: Size of assembled pseudomolecules.

Range (bp)	Insertion frequency	Insertion size (bp)
<500	1,320	158,325
>500 & <1000	34	23,184
>1000 & <2000	21	29,098
>2000	10	29,625
Total	1,385	240,232

Table S4: Length and distribution of chloroplast DNA insertions in Spirodela scaffolds.

Range (bp)	Insertion frequency	Insertion size (bp)
<500	1554	180,540
>500 & <1000	31	20,834
>1000 & <2000	3	4,152
>2000	1	2,185
Total	1589	207,711

Table S5: Length and distribution of mitochondrial DNA insertions in Spirodela scaffolds.

Fosmid query	Size	Start	Stop	Direction	Assembly pseudo	Start	Stop	Identity	Error	Miss	Accuracy
13491	37,210	1	37,210	-	6	5,625,554	5,663,330	37,188	7	22	99.98
13492	43,619	8,143	42,580	-	21	49,600	92,500	32,110	2	11,509	99.99
13493	41,079	1	41,079	+	20	2,913,695	2,956,496	38,221	1	2,858	100.00
13494	43,229	1	43,229	+	6	3,259,087	3,302,374	39,318	2	3,911	99.99
13495	33,630	1	33,630	+	13	2,820,000	2,851,981	31,254	2	2,376	99.99
13496	35,790	1	35,790	-	26	1,515,084	1,550,953	35,790	2	0	99.99
13497	26,996	1,372	26,996	-	0	6,465,107	8,949,428	23,553	5	3,443	99.98
13497	13,414	27,211	40,410	-	10	1,756,936	1,781,965	13,245	71	169	99.46
13498	40,229	1	40,229	-	26	411,417	449,753	29,414	0	10,815	100.00
13499	36,170	1	36,170	-	17	3,406,476	3,443,411	36,124	2	46	99.99
13500	40,104	1	40,104	+	3	1,452,666	1,491,636	37,958	2	2,146	99.99
13501	40,089	1	40,089	+	3	1,452,666	1,491,621	37,943	2	2,146	99.99
13502	35,382	1	35,382	+	JN160603	52,704	88,082	35,375	5	7	99.99
13503	43,217	1	43,217	+	15	463,213	510,205	35,680	1	7,537	100.00
13504	45,509	1	25,237	-	5	3,552,208	3,577,529	23,975	33	21,534	99.86
13505	40,939	1	40,939	+	13	212,307	252,386	37,127	3	3,812	99.99
13506	36,086	1	36,086	-	17	1,509,134	1,545,821	36,085	8	1	99.98
13507	32,343	1	32,343	+	1	5	32,295	31,811	4	532	99.99
13508	36,629	1	36,629	-	23	942,714	980,301	34,021	0	2,608	100.00
13509	40,969	1	40,969	-	14	3,247,540	3,288,708	40,970	6	0	99.99
13510	35,341	1	35,156	-	2	4,143,463	4,175,364	30,804	4	4,537	99.99
13511	44,030	1	44,030	+	3	626,332	670,546	41,990	17	2,040	99.96
13512	44,501	1	44,501	-	17	1,589,440	1,634,037	44,331	14	170	99.97
13513	36,687	1	35,199	-	15	3,375,544	3,410,486	34,637	0	2,050	100.00
13514	40,097	2,122	40,097	+	4	3,237,920	3,275,575	27,381	541	12,716	98.22

Table S6: Accuracy of the assembly validated by fosmid sequence.

Species	Spirodela	Rice	Musa	Tomato	Arabidopsis
# genes	19,623	39,049	36,542	34,727	27,416
Mean gene size	3,458	2,330	3,596	2,942	1,869
Median gene size	2,245	1,654	2,268	1,872	1,559
Mean CDS size	1,108	1,064	1,038	1,036	1,218
Median CDS size	903	849	861	822	1,047
Mean exon size	213	259	192	229	238
Median exon size	121	139	128	134	134
Mean exon#/gene	5.2	4.1	5.4	4.5	5.1
Median exon#/gene	4	3	4	3	3
Mean intron size	560	407	581	541	159
Median intron size	178	170	148	215	99

Table S7: Gene characteristics*.

*Table shows statistics of gene features for 3 monocotyledonous species (Spirodela, Rice and Banana) and two dicotylodonous species (Tomato and Arabidopsis). For each species, alternative splice variants were not considered for the statistics and either the representative model for one locus – if available – or the longest transcript of each locus was used. CDS describes the coding sequence from start to stop codon without introns, gene the genomic sequence from start to stop codon including intronic sequences. All sizes are shown in [bp].

Species	Spirodela	Rice	Musa	Tomato
Genes found	2,270	2,398	2,365	2,546
% Genes found	83.2	87.9	86.7	93.4

Table S8: Comparison of gene content with the AraCyc pathway.

Non-redundant (i.e. excluding alternative splice variants) gene sets of four species were compared to the Arabidopsis genes of the AraCyc-pathways by blastp. The last row lists the number of AraCyC genes with high similarity to genes of each species.

#	Treatment	Samples	Replication	Total
1	Short day 2 day time course	12	3	36
2	Intermediate day2 day time course	12	3	36
3	Long day 2 day time course	12	3	36
4	Continous light 2 day time course	12	3	36
5	Short day 2 day time course with sucrose	12	3	36
6	37°C (bathwater)	6	3	18
7	0° C (ice water)	6	3	18
8	Desiccation (agar plates)	6	3	18
9	NaCl	6	3	18
10	pH (KOH, 9)	6	3	18
11	Control	6	3	18
12	Darkness	6	3	18
13	UV (15 min exposure)	6	3	18
14	Cu (CuCl2,20mg/l)	6	3	18
15	N (KNO3,300mg/l)	6	3	18
16	ABA (250nM)	6	3	18
17	Kinetin (10uM)	6	3	18
18	Mannitol (300mM)	6	3	18
	Total	138	54	414

Table S9: Spirodela samples pooled for Roche/454 EST sequencing.

Dataset	Number	Mapped number	Percent
total ESTs	379,502	363,065	95.7%
>100 bp	271,542	261,026	96.1%
>300 bp	141622	136,984	96.7%
>500 bp	34609	33,507	96.8%

Table S10: Number of ESTs aligned to pseudomolecules.

Species	Arabidopsis	Tomato	Sorghum	Rice	Brachypodium	Banana	Spirodela
Cluster	1,938	2,340	2,118	2,602	1,836	1,084	948
Genes	5,366	7,091	6,240	8,249	5,038	2,755	2,977
Genes [%]	19.6	20.4	22.6	21.1	19	7.5	15.2

Table S11: Tandem genes and clusters in 7 species*.

*Tandem cluster and genes are determined as described. The last row shows the percentage of tandemly repeated genes in relation to the entire number of gene loci in each genome.

miRNA family	count	miRNA family	Count
miR1029	1	miR414	9
miR1030	1	miR414/5658	11
miR1031	1	miR435	1
miR1048	1	miR4414	1
miR1511	2	miR477	1
miR1522	1	miR482	27
miR1530	1	miR482/2118	1
miR1533	46	miR4993	6
miR156	24	miR5021	45
miR156/157	7	miR5041	2
miR156/159/319	1	miR5054	4
miR160	6	miR5075	4
miR164	3	miR5079	1
miR165	1	miR5083	2
miR165/166	2	miR5137	1
miR166	1	miR5139	4
miR167	4	miR5142	2
miR168	4	miR5238	1
miR169	9	miR5239	1
miR170/171	3	miR5244	1
miR171	7	miR529	13
miR172	5	miR530	1
miR1848	1	miR5384	14
miR2105	2	miR5386	1
miR2607	1	miR5523	1
miR2628	1	miR5559	1
miR2662	1	miR5565	2
miR2663	1	miR5658	25
miR2673	4	miR5712	1
miR2911	1	miR5809	2
miR2919	13	miR5819	1
miR2925	1	miR6108	1
miR2931	1	miR6145	1
miR319	2	miR6180	1
miR3437	1	miR6187	1
miR3522	1	miR6190	1
miR390	2	miR6196	5
miR393	3	miR6214	1
miR394	2	miR6249	1
miR395	2	miR6300	2
miR396	11	miR6429	2
miR397	1	miR821	1
miR3979	2	miR845	4
miR398	1	miR847	1
miR399	4	miR854	1
miR400	1	miR859	1
miR408	3	miR902	8

Table S12: Number of candidate miRNAs for each miRNA family of Spirodela.

	At	Sp	Bd	Os	Sb
genome size (N free)	119 Mb	128 Mb	270 Mb	372 Mb	626 Mb
Mobile Element (TXX)	17.3	n.a.	28.1	42.5	63.5
Class I: Retroelement (RXX)	**11.7**	**13.06**	**23.3**	**32.1**	**54.5**
LTR Retrotransposon (RLX)	10.79	13.06	21.39	30.85	54.47
copia (RLC)	1.65	1.72	5.13	3.32	5.18
gypsy (RLG)	2.16	6.06	13.46	9.06	19.00
gypsy/copia ratio	*1.3*	*3.5*	*2.6*	*2.7*	*3.7*
unclassified LTR (RLX)	6.98	5.27	2.80	18.46	30.28
non-LTR Retrotransposon (RXX)	0.89	n.a.	1.94	1.24	0.06
Class II: DNA Transposon (DXX)	**5.4**	n.a.	**4.8**	**10.1**	**7.5**
Unclassified Element (TXX)	0.25	n.a.	0.00	0.26	1.51
VNTR (variable number tandem repeat)	**2.35**	**1.66**	**3.29**	**1.99**	**3.13**
Microsatellite (2-9 bp unit)	0.13	**0.83**	0.19	0.05	0.15
Minisatellite (10-99 bp unit)	0.99	0.47	1.73	1.07	0.99
Satellite (>=100 bp unit)	0.85	0.25	0.92	0.70	1.49
Hybrid	0.38	0.11	0.45	0.17	0.51

Table S13: Repeat composition of Spirodela compared to other plant genomes.

BlockID	Contig A	Size A [bp]	# A-genes	Contig B	Size B [bp]	# B-genes	# Syn-genes	% syntenic
14	**pseudo1**	5986615	965	**pseudo11**	2236636	381	120	8.9
183	**pseudo1**	791536	127	**pseudo16**	379360	45	18	10.5
281	**pseudo1**	4635134	752	**pseudo16**	2371725	415	124	10.6
282	**pseudo1**	347243	59	**pseudo16**	208867	32	16	17.6
24	**pseudo1**	5880219	954	**pseudo17**	2039428	328	104	8.1
273	**pseudo10**	2832058	386	**pseudo12**	3247954	497	72	8.2
254	**pseudo10**	516725	62	**pseudo25**	454149	39	14	13.9
31	**pseudo10**	1886712	263	**pseudo28**	1541139	229	44	8.9
33	**pseudo10**	3472044	446	**pseudo5**	1870304	253	42	6
35	**pseudo11**	2194230	373	**pseudo16**	2970758	474	66	7.8
37	**pseudo11**	2472099	408	**pseudo17**	2359627	364	56	7.3
255	**pseudo12**	408417	48	**pseudo25**	303340	28	12	15.8
300	**pseudo12**	2209151	369	**pseudo28**	1409408	207	64	11.1
43	**pseudo12**	3702086	562	**pseudo5**	1938825	263	100	12.1
47	**pseudo13**	294683	51	**pseudo4**	564026	94	26	17.9
49	**pseudo13**	2241845	300	**pseudo4**	4556218	678	88	9
237	**pseudo13**	300374	49	**pseudo7**	455376	63	14	12.5
175	**pseudo13**	148169	33	**pseudo9**	140184	28	16	26.2
298	**pseudo13**	1381629	208	**pseudo9**	1756848	249	42	9.2
292	**pseudo14**	1423510	215	**pseudo18**	1445093	218	32	7.4
290	**pseudo14**	1010524	122	**pseudo19**	1471477	197	34	10.7
263	**pseudo14**	950434	162	**pseudo21**	2022141	328	48	9.8
289	**pseudo14**	604627	105	**pseudo26**	1712145	270	30	8
179	**pseudo14**	1272318	211	**pseudo8**	1105829	187	32	8
262	**pseudo15**	3332950	359	**pseudo20**	1772754	205	42	7.4
75	**pseudo15**	686890	102	**pseudo4**	359664	44	16	11
231	**pseudo15**	688296	68	**pseudo4**	428570	38	18	17
81	**pseudo16**	840491	89	**pseudo17**	570327	75	14	8.5
297	**pseudo16**	2178919	377	**pseudo17**	1645919	268	60	9.3
275	**pseudo18**	328476	32	**pseudo19**	549903	46	14	17.9
276	**pseudo18**	404230	39	**pseudo19**	581251	70	14	12.8
279	**pseudo18**	354253	56	**pseudo19**	786229	114	28	16.5
87	**pseudo18**	2636340	327	**pseudo8**	1887433	309	46	7.2
89	**pseudo19**	3022807	370	**pseudo8**	1216093	201	52	9.1
94	**pseudo2**	246461	37	**pseudo20**	662994	69	14	13.2
96	**pseudo2**	1523560	187	**pseudo24**	767306	88	38	13.8
98	**pseudo2**	3878913	629	**pseudo24**	1532881	214	82	9.7
99	**pseudo2**	3897393	646	**pseudo3**	1393329	201	58	6.8
101	**pseudo2**	3880852	647	**pseudo9**	1357788	207	80	9.4
229	**pseudo20**	2074670	254	**pseudo4**	2444847	236	34	6.9
107	**pseudo21**	988765	161	**pseudo8**	416552	76	30	12.7
110	**pseudo22**	848599	164	**pseudo23**	2496506	418	74	12.7
287	**pseudo22**	1723210	297	**pseudo29**	1698748	290	52	8.9
283	**pseudo22**	1172956	180	**pseudo3**	2612309	425	70	11.6
285	**pseudo22**	2496312	399	**pseudo6**	2675149	393	62	7.8
288	**pseudo23**	2043438	340	**pseudo29**	668082	123	44	9.5
286	**pseudo23**	2530083	421	**pseudo6**	1157301	206	80	12.8
174	**pseudo24**	1676808	228	**pseudo9**	1568196	223	30	6.7
133	**pseudo26**	1675331	261	**pseudo8**	617307	99	26	7.2
137	**pseudo28**	607075	109	**pseudo5**	464627	80	26	13.8
201	**pseudo29**	142864	16	**pseudo3**	225652	41	14	24.6
202	**pseudo29**	667010	114	**pseudo3**	1331870	217	32	9.7
143	**pseudo29**	1551521	258	**pseudo6**	1728419	272	48	9.1
199	**pseudo3**	2780539	450	**pseudo6**	1387647	181	54	8.6
200	**pseudo3**	1351389	154	**pseudo6**	722719	83	16	6.8
153	**pseudo3**	203072	35	**pseudo9**	197690	34	12	17.4
155	**pseudo3**	176603	35	**pseudo9**	327013	56	16	17.6
158	**pseudo4**	1587985	267	**pseudo7**	854422	92	30	8.4
163	**pseudo4**	415304	70	**pseudo9**	210421	31	18	17.8
171	**pseudo4**	2448699	431	**pseudo9**	1098817	161	56	9.5
173	**pseudo7**	285316	32	**pseudo9**	285667	43	14	18.7

Table S14: Genomic size and number of conserved paralogous gene pairs of duplicated chromosomal segments in Spirodela.

Class	Spirodela	banana	rice	Arabidopsis	tomato
all	1.34	2.02	1.88	1.77	1.87
At				4.06	
Mu		2.72			
Os			3.96		
Sp	3.99				
To					4.24
At, Mu		1.80		1.42	
At, Os			2.14	2.15	
At, Sp	1.02			1.20	
At, To				1.60	1.86
Mu, Os		1.93	1.75		
Mu, Sp	1.12	1.53			
Mu, To		1.37			1.20
Os, Sp	2.24		1.72		
Os, To			2.43		2.20
Sp, To	1.50				2.13
At, Mu, Os		1.76	1.50	1.56	
At, Mu, Sp	1.09	1.64		1.35	
At, Mu, To		1.69		1.44	1.82
At, Os, Sp	1.05		1.40	1.18	
At, Os, To			1.86	2.30	2.12
At, Sp, To	1.25			1.32	1.43
Mu, Os, Sp	1.17	2.00	1.38		
Mu, Os, To		1.56	1.67		1.50
Mu, Sp, To	1.19	2.00			1.36
Os, Sp, To	1.27		2.09		3.45
At, Mu, Sp, To	1.14	1.78		1.37	1.46
At, Os, Sp, To	1.16		1.72	1.50	1.52
Mu, Os, Sp, To	1.31	1.78	1.76		1.72
At, Mu, Os, Sp	1.25	1.70	1.22	1.29	
At, Mu, Os, To		1.96	1.47	1.60	1.53
At, Mu, Os, Sp, To	1.23	2.03	1.51	1.57	1.57

Table S15: Average copy number per orthoMCL cluster for each species.

Mean copy number of paralogous and co-orthologous genes per orthoMCL-cluster was determined for each species and each division of the Venn diagram as described in Supplement 5.2. The set of species defining a particular division is provided by abbreviations: At: Arabidopsis, Mu: Banana, Os: Rice, Sp: Spirodela and To: Tomato. Row ‘all’ shows the average of all clusters. Species with the lowest copy number in each division are shown in red.

GO-ID	p-value	GO term description	GO term description
GO:0015976	1,20E-15	carbon utilization	carbon utilization
GO:0006979	1,03E-13	response to oxidative stress	response to oxidative stress
GO:0051707	3,95E-13	response to other organism	response to other organism
GO:0009698	6,52E-11	phenylpropanoid metabolic process	phenylpropanoid metabolic process
GO:0055114	3,15E-10	oxidation reduction	oxidation reduction
GO:0005975	1,72E-09	carbohydrate metabolic process	carbohydrate metabolic process
GO:0009692	1,95E-09	ethylene metabolic process	ethylene metabolic process
GO:0046274	2,37E-09	lignin catabolic process	lignin catabolic process
GO:0006555	6,27E-09	methionine metabolic process	methionine metabolic process
GO:0030036	1,11E-08	actin cytoskeleton organization	actin cytoskeleton organization
GO:0030259	1,39E-08	lipid glycosylation	lipid glycosylation
GO:0071266	8,06E-08	de novo' L-met biosynthetic process	de novo' L-met biosynthetic process
GO:0010044	9,00E-08	response to aluminum ion	response to aluminum ion
GO:0006857	2,70E-07	oligopeptide transport	oligopeptide transport
GO:0006813	2,94E-07	potassium ion transport	potassium ion transport
GO:0006629	3,42E-07	lipid metabolic process	lipid metabolic process
GO:0046836	4,24E-07	glycolipid transport	glycolipid transport
GO:0009773	2,11E-06	photosynthetic electron transport in PS I	photosynthetic electron transport in PS I
GO:0050896	2,63E-06	response to stimulus	response to stimulus
GO:0030001	3,03E-06	metal ion transport	metal ion transport
GO:0006855	1,01E-05	multidrug transport	multidrug transport
GO:0007067	5,79E-05	mitosis	mitosis
GO:0006597	2,28E-04	spermine biosynthetic process	spermine biosynthetic process
GO:0007264	2,54E-04	small GTPase mediated signaltransduction	small GTPase mediated signaltransduction
GO:0015986	3,40E-04	ATP synthesis coupled proton transport	ATP synthesis coupled proton transport
GO:0009664	9,19E-04	plant-type cell wall organization	plant-type cell wall organization
GO:0006811	2,26E-03	ion transport	ion transport
GO:0008299	2,70E-03	isoprenoid biosynthetic process	isoprenoid biosynthetic process
GO:0006098	7,99E-03	pentose-phosphate shunt	pentose-phosphate shunt
GO:0007049	2,14E-02	cell cycle	cell cycle
GO:0009094	2,33E-02	L-phenylalanine biosynthetic process	L-phenylalanine biosynthetic process
GO:0006833	2,92E-02	water transport	water transport

Table S16: Overrepresented GO terms in division of AtMuOsTo except Spirodela.

Table shows overrepresented GO terms for the orthoMCL family cluster present in four genomes but missing in Spirodela. Further explanation for some columns and GO identifiers is shown in Supplement paragraph 5.3.

GO-ID	pvalue	GO short description	identified proteins
GO:0045926	1,32E-66	negative regulation of growth	antimicrobial peptide
GO:0006952	6,29E-21	defense response	antimicrobial peptide, NBS-LRR
GO:0006468	1,40E+00	protein amino acid phosphorylation	protein kinase
GO:0043086	3,58E+00	negative regulation of catalytic activity	subtilisin
GO:0006915	6,92E+01	apoptosis	NBS-LRR
GO:0009856	3,99E+02	pollination	S-locus receptor kinase
GO:0016998	3,19E+03	cell wall macromolecule catabolic process	GPI-anchored protein
GO:0006869	3,67E+03	lipid transport	bifunctional lipid transfer protein
GO:0006979	4,35E+03	response to oxidative stress	peroxidases
GO:0005985	7,23E+03	sucrose metabolic process	sucrose synthase
GO:0006508	8,38E+03	proteolysis	proteinases
GO:0071554	2,81E+04	cell wall organization or biogenesis	GPI-anchored protein
GO:0006629	4,16E+04	lipid metabolic process	mixed

Table S17: Overrepresented functional categories of Spirodela-specific genes.

First column shows the overrepresented GO identifier, second the p-value, third a short description of the GO term. The forth column lists a description of molecular class(es) of proteins found in majority in the cluster with the respective GO term. 'Mixed' indicates that there was no major contribution by one class.

Family name	Sub-Family	Arabidopsis	Rice	Spirodela
CesA	Total	10	10	10
CSL	A	9	11	11
	B	6	0	0
	C	5	6	5
	D	5	5	5
	E	1	3	0
	F	0	8	0
	G	3	0	0
	H	0	3	0
	Total	29	36	21
GT31	A	12	7	4
	B	6	10	2
	C	8	8	2
	D	3	2	2
	E	3	2	7
	F	1	10	4
	Total	33	39	21

Table S18: Comparative numbers of cellulose biosynthesis gene families in Arabidopsis, rice and Spirodela.

Species	CAD	CCoAMT	4CL	CCR	PAL	C4H	HCT	COMT	C3H	F5H	Total
Arabidopsis	9	7	13	7	4	1	1	16	3	2	63
Medicago	21	4	10	18	4	1	6	26	1	3	94
Poplar	21	7	22	40	6	3	7	35	4	4	149
Sorghum	14	7	15	44	8	3	4	41	2	3	141
Rice	5	11	16	55	14	4	9	38	1	3	156
Spirodela	4	1	9	21	3	3	20	5	1	3	70
Total	74	37	85	185	39	15	47	161	12	18	673

Table S19: Copy numbers of mono-lignin biosynthesis genes.

Copy number of each gene family from Arabidopsis, Medicago, Poplar, Sorghum, Rice and Spirodela for biosynthesis of mono-lignols is shown.

Gene family	Clade	Arabidopsis	Spirodela	Oryza	Zea
ADP-glucose pyrophosphorylase	small subunit (AGPS)	AtAPS1: U70616	Sp_AGPS1: Spipo28G0001400	OsAGPS1: AK073146	ZmAGPS1: AY032604
		AtAPS2: NM_100441		OsAGPS2a: AK071826	ZmAGPS2: AF334959
				OsAGPS2b: AK103906	
	large subunit (AGPL)	AtAPL1: U72290	Sp_AGPL1: Spipo3G0049000	OsAGPL1: D50317	ZmAGPL1: Z38111
		AtAPL2: AC012375	Sp_AGPL2: Spipo6G0024200	OsAGPL2: U66041	ZmAGPL2: M81603
		AtAPL3: Y18432	Sp_AGPL3: Spipo18G0038500	OsAGPL3: AK069296	
		AtAPL4: NM_127730		OsAGPL4: AK121036	
Starch synthase	Soluble starch synthase I (SSI)	AtSSI: AY128273	SpSSI: Spipo26G0026900	OsSSI: D16202	ZmSSI: AF036891
	Soluble starch synthase II (SSII)	AtSSII: BT002555	SpSSII: Spipo0G0050800	OsSSIIa: AF419099	ZmSSIIa: AF019296
				OsSSIIb: AF395537	ZmSSIIb: AF019297
				OsSSIIc: AF383878	
	Soluble starch synthase III (SSIII)	AtSSIII: AC007296	SpSSIII: Spipo14G0048800	OsSSIIIa: AY100469	ZmSSIII: AF023159
				OsSSIIIb: AF432915	
	Soluble starch synthase IV (SSIV)	AtSSIV: AL161548	SpSSIV: Spipo14G0042000	OsSSIVa: AY100470	
				OsSSIVb: AY373258	
	Granule-bound starch synthase (GBSS)	AtGBSSI: AY149948	SpGBSSI: Spipo1G0057900	OsGBSSI: X62134	ZmGBSSI: X03935
				OsGBSSII: AY069940	
Starch branching enzyme	Starch branching enzyme I (BEI)		SpBEI: Spipo1G0057400	OsBEI: D11082	ZmBEI: AF072724
	Starch branching enzyme II (BEII)	AtSBE2-1: NM_129196	SpBEII: Spipo0G0008100	OsBEIIa: AB023498	ZmBEIIa: U65948
		AtSBE2-2: NM_120446		OsBEIIb: D16201	ZmBEIIb: AF072725
	Starch branching enzyme II (BEIII)		SpBEIII: Spipo19G0037800		
Starch debranching enzyme	Isoamylase (ISA)	AtISA1: BT000443	SpISA1: Spipo12G0062400	OsISA1: AB093426	ZmISA1: AF030882
		AtISA2: AY139980	SpISA2: Spipo3G0051400	OsISA2: AC132483	ZmISA2: AY172633
		AtISA3: AY091058	SpISA3: Spipo20G0022100	OsISA3: AP005574	ZmISA3: AY172634
	Pullulanase (PUL)	AtPUL: BT002411	SpPUL: Spipo1G0018500	OsPUL: AB012915	ZmPUL: AF080567
Total member #		17	15	24	16

Table S20: List of starch biosynthesis genes and their origins.

Subfamily	Arabidopsis	Rice	Spirodela
FT/TSF	2	2	1
TFL1/CEN	2	2	2
MFT1	1	2	4
AP2, TOE1-3, SNZ, SMZ	5	n.d.	5
SPL9/15	2	3	1
SPL2/10/11	3	4	1
SPL3/4/5	3	2	2
SOC1, SOC-like	5	3	0
SVP, FLC, FLM, MAFs	8	2	6
SEP	4	5	1
AP1, CAL, FUL	3	3	1

Table S21: Copy numbers of selected genes involved in juvenile-to-adult transition and flowering.

Repressors of the adult phase are highlighted in red, promoters in green. For a description of the genes and gene names, see Supplement section 5.4.6.

Hindawi Publishing Corporation
Journal of Botany
Volume 2011, Article ID 570319, 9 pages
doi:10.1155/2011/570319

Research Article

Evolution of Genome Size in Duckweeds (*Lemnaceae*)

Wenqin Wang,[1] Randall A. Kerstetter,[1,2] and Todd P. Michael[1,2]

[1] *Rutgers, Department of Plant Biology and Pathology, The State University of New Jersey, The Waksman Institute of Microbiology, Piscataway, NJ 08854, USA*
[2] *Monsanto Company, 800 North Lindbergh Boulevard, Creve Coeur, MO 63167, USA*

Correspondence should be addressed to Todd P. Michael, todd.p.michael@monsanto.com

Received 4 February 2011; Revised 5 April 2011; Accepted 19 May 2011

Academic Editor: Johann Greilhuber

To extensively estimate the DNA content and to provide a basic reference for duckweed genome sequence research, the nuclear DNA content for 115 different accessions of 23 duckweed species was measured by flow cytometry (FCM) stained with propidium iodide as DNA stain. The 1C-value of DNA content in duckweed family varied nearly thirteen-fold, ranging from 150 megabases (Mbp) in *Spirodela polyrhiza* to 1,881 Mbp in *Wolffia arrhiza*. There is a continuous increase of DNA content in *Spirodela, Landoltia, Lemna, Wolffiella*, and *Wolffia* that parallels a morphological reduction in size. There is a significant intraspecific variation in the genus *Lemna*. However, no such variation was found in other studied species with multiple accessions of genera *Spirodela, Landoltia, Wolffiella,* and *Wolffia*.

1. Introduction

The *Lemnaceae*, commonly known as duckweeds, are the smallest, fastest-growing, and simplest of flowering plants. In this globally distributed aquatic monocot family (Figure 1(a)), there are 33 species representing five genera: *Spirodela*, *Landoltia*, *Lemna*, *Wolffiella,* and *Wolffia*. Among them, *Spirodela* is the most ancestral, while *Wolffia* is the most derived [1]. The individual plants range in size from 1.5 cm long (*Spirodela polyrhiza*) to less than one millimeter (*Wolffia globosa*). Therefore, there is a successive reduction of morphological structures in parallel with evolutionary advancement within the family (Figure 1(b)). Duckweeds are not simply miniature versions of larger angiosperms; they represent a highly modified structural organization that resulted from the alteration, simplification, or loss of many morphological and anatomical features [2]. The biomass doubling time of the fastest-growing duckweeds in optimal growth conditions is less than 30 hours, nearly twice as fast as other "fast" growing flowering plants and more than double that of conventional crops [3].

Before the days of *Arabidopsis*, duckweeds, and more specifically *Lemna,* were an important model system for plant biology [4]. Since duckweeds are small, morphologically reduced (although with root and leaf-like structure), fast growing, easily cultivated under aseptic conditions (Figure 1(c)), transformable, crossable, and particularly suited to biochemical studies (direct contact with media), it is an ideal system for biological research [5]. Much of what we know about photoperiodic flowering responses comes from fundamental research conducted on *Lemna* by the preeminent plant biologist Dr. William Hillman at the Brookhaven National Laboratories [6]. Some of the current uses of *Lemnaceae* are a testimony to its scientific, commercial, and biomass utility: basic research and evolutionary model system [7], toxicity testing organism [8], biotech protein factories [9], wastewater remediation [10], high protein animal feed, carbon cycling [5], and biofuel potential candidates [11].

The advent of high-throughput sequencing technologies has enabled a new generation of model plant systems [12]. In an effort to initiate duckweed genomic research, we endeavoured to identify species with small genomes that would be ideal for sequencing. First, we queried the Kew plant genome database (http://data.kew.org/cvalues/) and found that there were only 6 duckweed accessions that

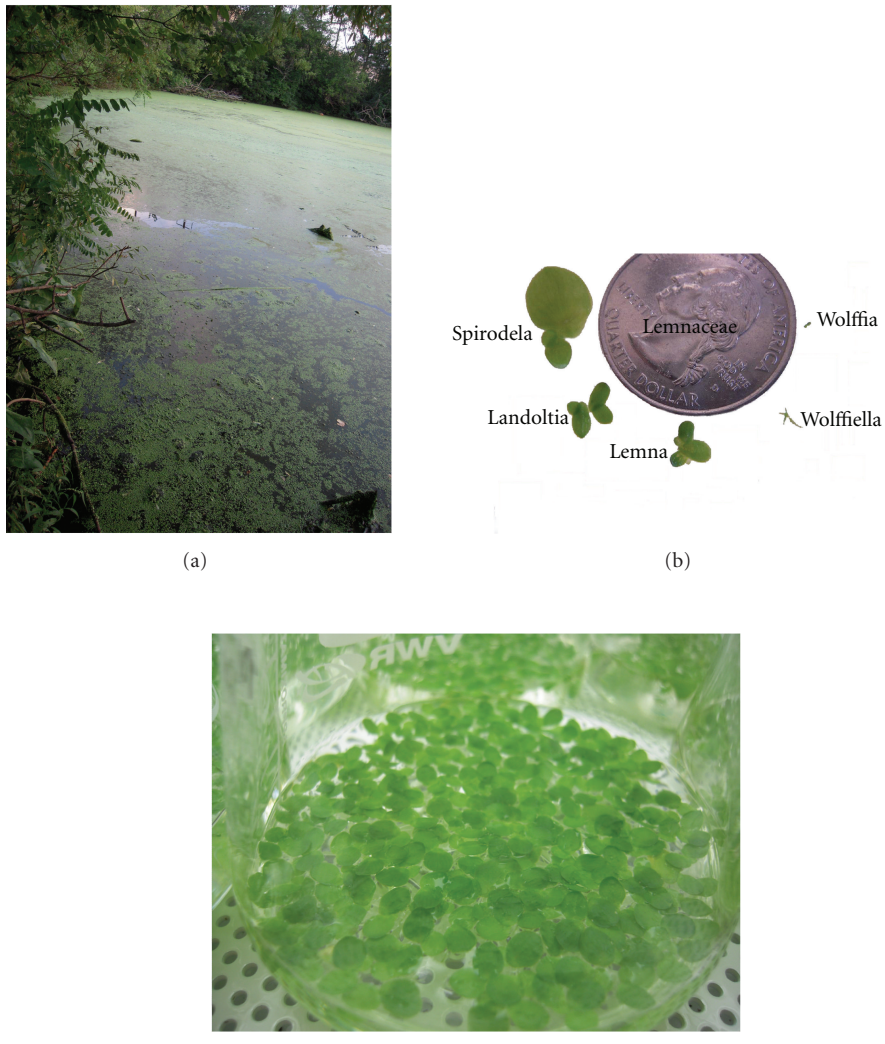


Figure 1: Duckweeds are small aquatic plants that are widely distributed in nature and amenable to culturing in the lab. (a) Duckweeds growing in the Raritan Canal River, Piscataway, NJ, USA. This population of duckweed includes *Wolffia*, *Spirodela,* and *Lemna*. (b) The relative size of *Spirodela, Landoltia, Lemna, Wolffiella,* and *Wolffia* in the order of phylogeny as compared to an American Quarter. (c) Sterile *Spirodela polyrhiza* grown in the Schenk and Hildebrandt basal salt medium.

had been measured by the Feulgen method [13, 14]. DNA content of single species from each genus was determined and showed obvious difference. Due to it being laborious and time consuming, the popularity of Feulgen technique has waned. Feulgen has been largely replaced by flow cytometry (FCM) [15], a faster, easier, and more accurate method and the current preferred technique for genome size estimations and DNA ploidy analyses in plants [16].

In order to find the smallest duckweed genome for sequencing and also explore previous observations about genome complexity in duckweeds, we estimated the genome size of all of the five duckweed genera using FCM. These genome size measurements will form the foundation for future work in sequencing duckweed genome, and enabling duckweeds as a model and applied system.

2. Materials and Methods

2.1. Plant Materials. 115 accessions of 23 duckweed species representing all 5 genera were measured in this study (Table S1 in Supplementary Materials available online at doi:10.1155/2011/570319) (They provide details of sample collection and results of nuclear DNA content measurements for *Lemnaceae.*). Elias Landolt collected most of the duckweed accessions described in this work over the past 50 years (Landolt Duckweed Collection) [2]. Accessions were either obtained directly from Elias Landolt, BIOLEX (NC, USA) or The University of Toronto Culture Collection of Algae and Cyanobacteria (UTCC). Currently, the Landolt Duckweed Collection has been moved to Rutgers University. Additional lines were collected from lakes and wastewater

ponds by TPM and WW (NJ, USA). Plants were grown aseptically for 2 weeks with 1/2 full concentration of Schenk and Hildebrandt Basal Salt mixture (Sigma, USA) liquid culture medium under short day growth condition (8 h light and 16 h darkness with constant temperature 23°C). We bar-coded all the determined and undetermined species by identification of polymorphisms of chloroplast *atpF-atpH* noncoding spacer [17].

2.2. Isolation and Staining of Nuclei. To estimate nuclear DNA contents with flow cytometry (FCM), sample tissue nuclei were stained with propidium iodide (PI) [18]. Briefly, 10 mg of fresh duckweed tissue and the same amount of the internal standard were chopped simultaneously with new razor blades and isolation buffer in a plastic Petri dish [19]. Isolates were filtered through a 30-μm nylon mesh into an Eppendorf tube. The suspensions of nuclei were stained with 50 μg mL^{-1} PI mixed with 50 μg ml^{-1} RNase (R4875, Sigma). The samples were incubated on ice for a few minutes before estimation by FCM.

2.3. Analysis of Nuclear DNA Content by FCM. PI-stained nuclei were analyzed for DNA content with a Coulter Cytomics FC500 Flow Cytometer (Beckman Coulter, Inc., Miami, Florida, USA). In all experiments, the fluorescence of at least 3000 G1-phase nuclei was measured. DNA content of each target sample was calculated by comparing its mean nuclear fluorescence with that of an internal standard (Figure 2(a)). We utilized internal controls that closely match the duckweed genome sizes being measured to ensure accuracy. The internal standard is a *Brachypodium distachyon* line, (Bd21, 300 Mbp) [16], *Arabidopsis thaliana* Columbia., (At, 147 Mbp) [20], and *Physcomitrella patens* ssp patens, (Pp, 480 Mbp) [21]. The numbers in bracket were generated by our flow cytometry equipment and our methods. Therefore, the validated genome sizes are not exactly the same but very close to cited references. Both duckweed and internal standards have very little secondary compounds, which will interfere with quantitative DNA staining. The absolute DNA content of a sample is calculated based on the values of the G1 peak means:

$$\text{Sample 1C DNA content} = \left[\frac{(\text{sample G1 peak mean})}{(\text{standard G1 peak mean})}\right] \times \text{standard 1C DNA content (Mbp)}. \quad (1)$$

At least, three independent biological replicates for each sample were analyzed on different days to estimate the mean DNA content. The transformation factor from pg to Mbp is: 1 pg = 978 Mbp [22].

2.4. Statistical Analysis. Data on intraspecies variation of genome size were analysed by ANOVA: single factor test. To test whether genome size variation was correlated with geographic location or altitude of populations, the Spearman correlation coefficient (r) was used.

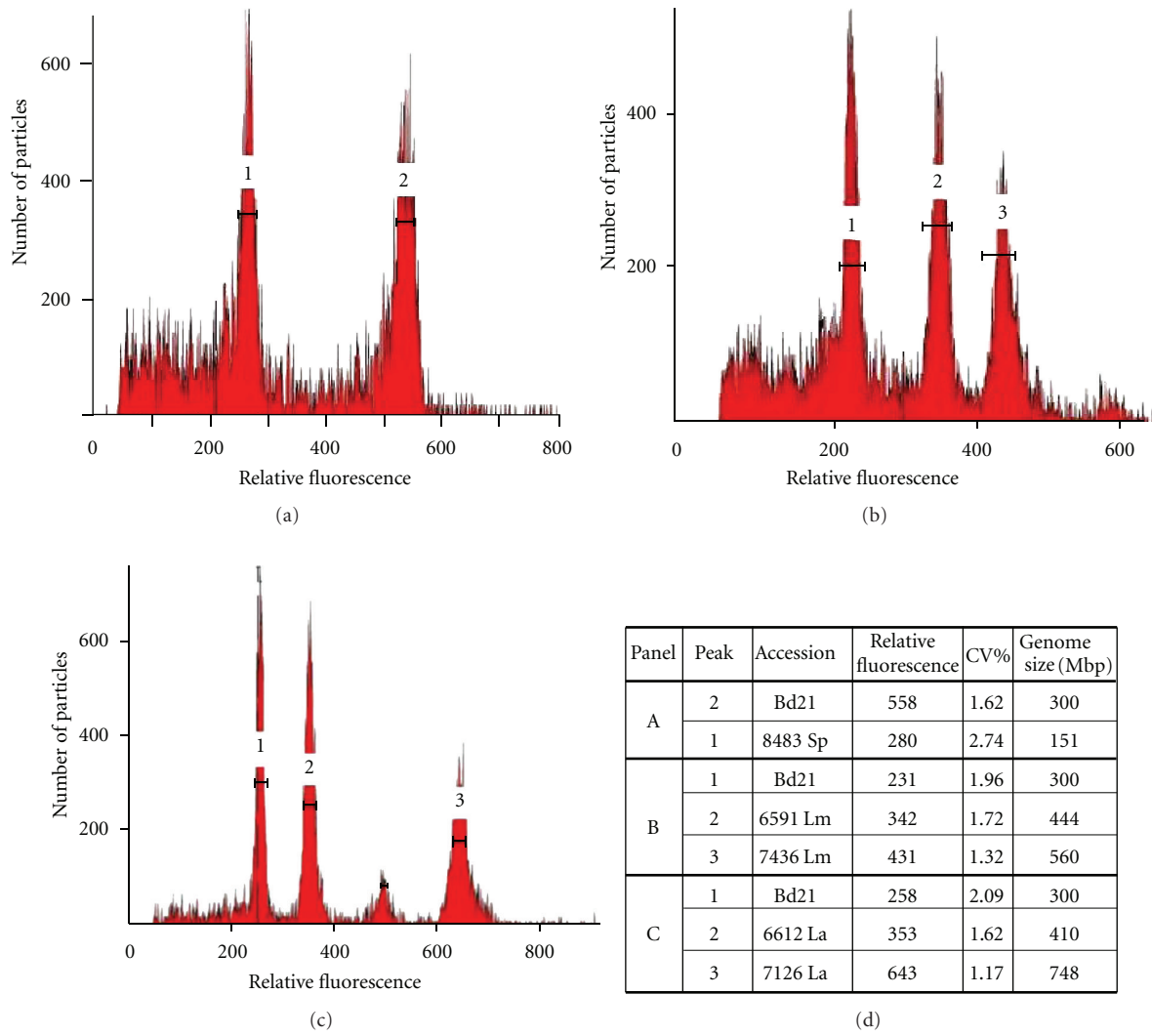
3. Results

3.1. Intra- and Interspecies Variations of Genome Sizes. The genome sizes of 115 accessions from 23 species representing 5 genera were estimated by FCM (Table S1). The DNA content estimates varied nearly thirteen-fold, ranging from 150 Mbp in *Spirodela polyrhiza* to 1,881 Mbp in *Wolffia arrhiza*. We superimposed the estimated 1C-value on a phylogenetic tree for *Lemnaceae* based on combination of morphological, flavonoid, allozyme, and DNA sequence analysis [1] and found that there is a continuous increase of DNA content in order of *Spirodela*, *Landoltia, Lemna*, *Wolffiella,* and *Wolffia*, which correlates well with the morphological reduction within the family (Figures 3(a) and 3(b)).

In the genus *Spirodela*, we measured genome size for 34 accessions and found that the 1C DNA content only varies from 150 to 167 Mbp (Figure 3(a); Table S1). The analysis of variance (ANOVA: single factor test) revealed that there was not a significant difference in *Spirodela polyrhiza* genome sizes ($P > 0.05$). Similarly, the 1C DNA content for 19 accessions of *Landoltia punctata* from 372 to 397 Mbp did not show significant variation (Figure 3(a); Table S1). In the genus *Wolffiella,* the genome sizes range from 623 Mbp to 973 Mbp (Figure 3(a)), which is almost as 4–6 times large as *Spirodela polyrhiza*. Like *Spirodela polyrhiza* and *Landoltia punctata*, there are no obvious intraspecific genome size variations in *Wolffiella hyalina* and *Wolffiella lingulata*. In the genus *Wolffia*, we measured 11 species and found that they have the largest genome sizes on average among the duckweed family (Figure 3(a)). 5.3-fold difference was observed from *Wolffia australiana* (357 Mbp) to *Wolffia arrhiza* (1,881 Mbp).

In the genus *Lemna*, 7 species were investigated. There is a large amount of genome size variation in this genus. *Lemna valdiviana* has the smallest genome size (323 Mbp), while *Lemna aequinoctialis* has the biggest (760 Mbp). Surprisingly, intraspecific genome-size fluctuations are also impressive. For *Lemna minor,* 26 accessions have genome sizes ranging from 356 to 604 Mbp with up to 69.6% of the intraspecific DNA content variance. We confirmed the intraspecific difference of them by randomly choosing 2 *Lemna minor* with simultaneous measurement of both accessions (26.0% difference between 6591 Lm and 7436 Lm, Figure 2(b)). Statistical analyses revealed significant differences among the *Lemna minor* accessions ($P < 0.01$). As well, *Lemna aequinoctialis* (424–760 Mbp, 79.2%) (Figure 2(c)), *Lemna trisulca* (446–709 Mbp, 59.0%), and *Lemna japonica* (426–600 Mbp, 40.8%) all show intraspecific difference, indicating a drastically uneven evolution of intraspecific genome expansion in *Lemna*.

3.2. 1C-Value and Latitude, Longitude, and Altitude. To investigate whether there is a correlation between genome-size variations and the geographic distribution in the duckweed, we compared genome size estimates with the latitude, longitude, and altitude of recorded collection. However, genome size variation was not correlated with



Panel	Peak	Accession	Relative fluorescence	CV%	Genome size (Mbp)
A	2	Bd21	558	1.62	300
	1	8483 Sp	280	2.74	151
B	1	Bd21	231	1.96	300
	2	6591 Lm	342	1.72	444
	3	7436 Lm	431	1.32	560
C	1	Bd21	258	2.09	300
	2	6612 La	353	1.62	410
	3	7126 La	643	1.17	748

(d)

Figure 2: Flow cytometry (FCM) histograms showing relative nuclear DNA content of Duckweed. (a) Histogram showing relative DNA content of *Spirodela polyrhiza* (1, 151 Mbp) and internal standard *Brachypodium distachyon* Bd21 (2, 300 Mbp) based on relative PI fluorescent intensity (channel number). Linear PI fluorescence intensity of G1 nuclei was used for the calculation of DNA content (3500 particles were counted); (b) Difference in relative DNA content of two simultaneously measured *Lemna minor* accessions (2, Lm6591, 444 Mbp; 3, Lm7436, 560 Mbp) with internal standard Bd21 (1); 5000 particles were counted. (c) Difference in relative DNA content of two simultaneously measured *Lemna aequinoctialis* accessions (2, La6612, 410 Mbp; 3, La7126, 748 Mbp) with internal standard Bd21 (1); 5000 particles were counted. (d) Summary of Panel a, b, and c and genome size corresponding to each peak.

latitude by Pearson coefficient (r-value: *Spirodela* = −0.05, *Landoltia* = 0.17, *Lemna* = −0.07, *Wolffiella* = −0.17, *Wolffia* = 0.34) (Figure 4(b)), nor with longitude (r-value: *Spirodela* = 0.17, *Landoltia* = 0.04, *Lemna* = 0.26, and *Wolffia* = −0.41) (Figure 4(c)) except *Wolffiella* with a high r-value −0.86 possibly due to limited accessions (n = 8). No correlation was found between C-values and altitude, either (r-value: *Spirodela* = 0.13, *Landoltia* = −0.25, *Lemna* = −0.33, *Wolffiella* = −0.41, and *Wolffia* = 0.13) (Figure 4(d)). It is interesting we found that most of *Spirodela, Landoltia, Wolffiella*, and *Wolffia* were collected from a similar geographic range between 0° to 45° and preferred to localize above 600 m to 1200 m of altitude. In contrast, most of *Lemna* species were collected between 30° to 60° and preferred to distribute below 600 m. However, this most likely represents a sampling bias and could also explain the absence of a relationship between genome size and the environment in duckweed.

4. Discussion

4.1. Genome Evolution in Duckweeds. In the phylogeny of *Lemnaceae*, there is a strong relationship observed between genome size evolution and morphological progression. We found that the ancestral genus *Spirodela* has the smallest genome size, while the most advanced genus *Wolffia* contains biggest genome size (Figure 3; Table S1), which correlates with the morphological reduction rather than organism

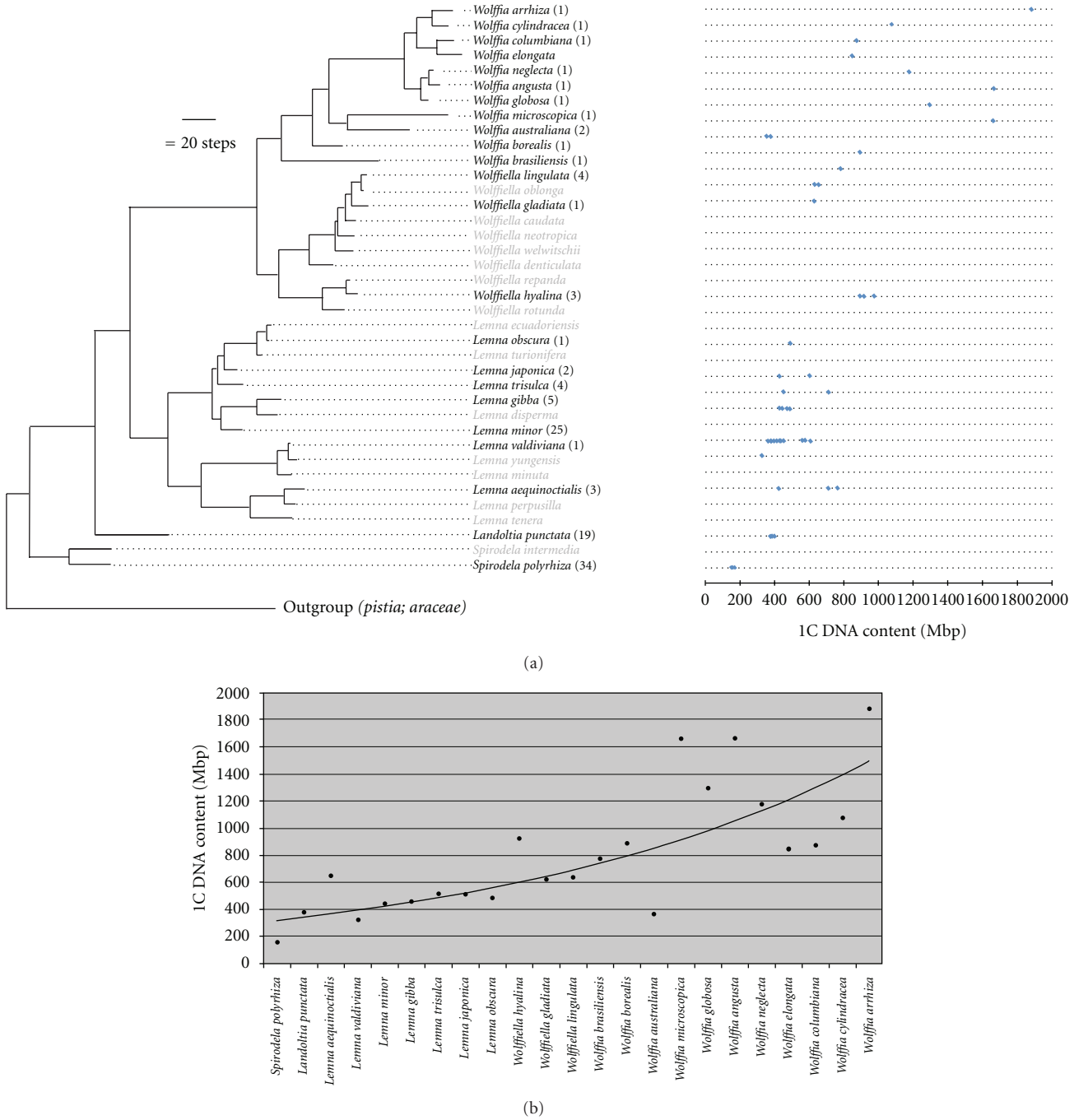


Figure 3: Genome size variation across the duckweeds. Estimated 1C-value superimposed on a phylogenetic tree for *Lemnaceae* based on combination of morphological, flavonoid, allozyme, and DNA sequence analysis [1]. The species in black were what we tested, and the species in the grey were the ones we did not examine in this experiment. In the bracket is the number of different accessions we tested. (b) Average genome sizes (y-axis) of duckweed species negatively parallel with degree of primitivity (x-axis). Duckweed species are arranged on the x-axis from lower to higher evolutionary status, which deduced from primitive and derived morphological traits [13].

complexity within the family. This result is consistent with Geber's finding, which showed that there was a relationship between DNA content and degree of primitivity [23].

Genome doubling has been a pervasive force in plant evolution, which has occurred repeatedly [24]. Even the smaller genome of *Arabidopsis thaliana* has been impacted by genome duplication [25]. Cytological variation by counting the chromosomes was extensively investigated within duckweed. They concluded that polyploidy ($2n$ = 20, 30, 40, 50, 60, and 80) is the main intrapopulational variation [2], which means polyploidization was very active and occurred in the duckweeds for multiple rounds in the past.

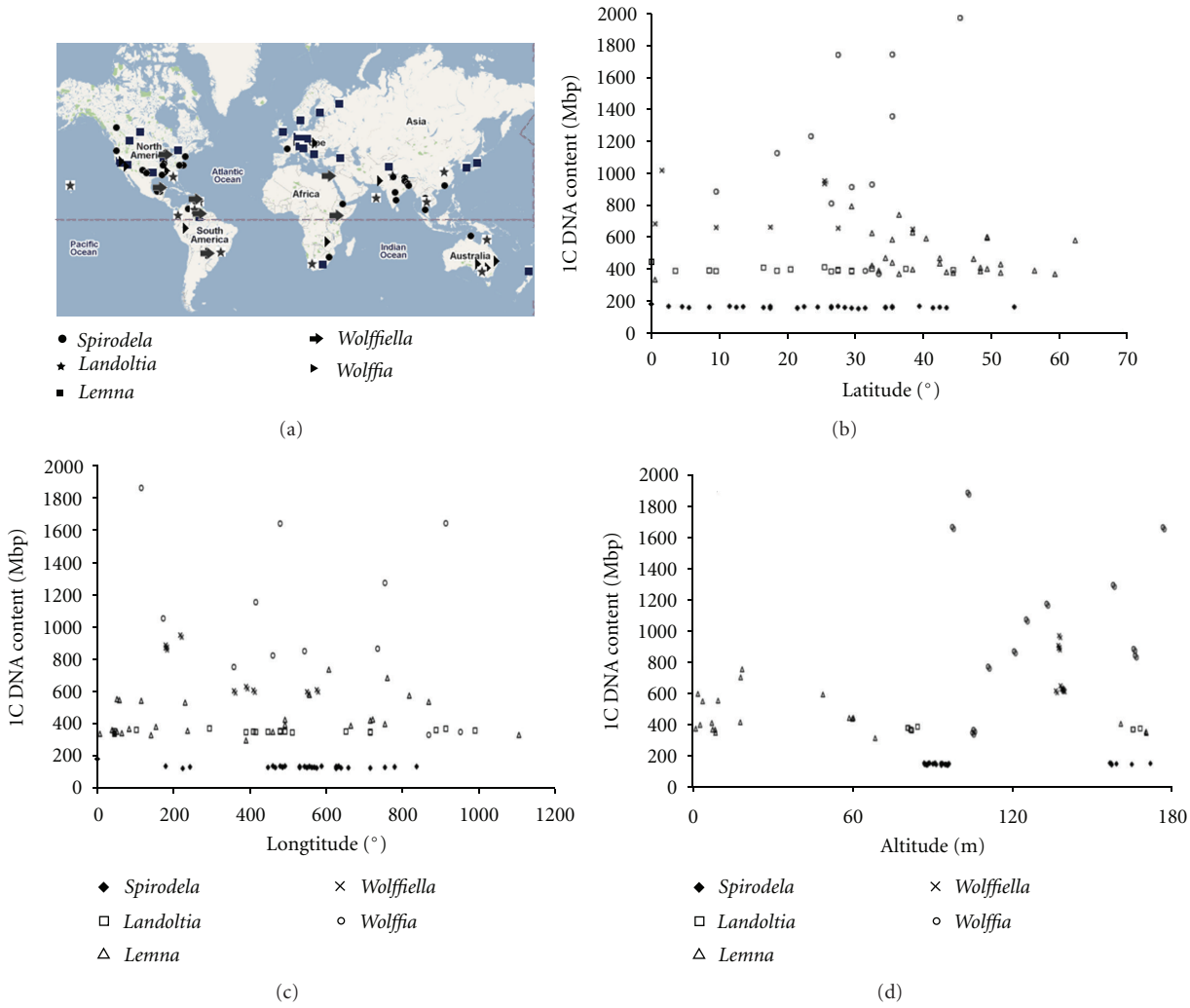


Figure 4: The relation of 1C DNA content with geographical coordinates and altitude. (a) Geographical origin of the duckweed accessions analyzed; (b) Latitude and 1C DNA content; (c) Longitude and 1C DNA content; (d) Altitude and 1C DNA content. http://maps.google.com/maps/ms?ie=UTF&msa=0&msid=208117868393582853927.0004a4942c7109c95133a.

After polyploidization, transposable element mobility, insertions, deletion, and epigenome restructuring contribute to the successful development of a new species and also genome size changes [26]. Changes in genome structure could lead to differential gene loss, extensive changes in gene expression [27], and have immediate effects on the phenotype and fitness of an individual [28]. It is likely polyploidy might drive the divergence during duckweed evolution.

4.2. Geographic Distribution and Genome Size Variation. It was suggested that variation in DNA content has adaptive significance and is correlated with the environmental traits of species [29]. The environmental conditions of plants are to a large extent determined by latitude, longitude, and altitude. Previous studies have indicated a positive correlation between genome size and latitude (associated with the length of sun light with the growing season and the temperature) and also altitude (associated with the temperatures) among plant species. For example, the increase of DNA content corresponded with the increasing latitude found in the *Pinaceae* family [30] and with increasing altitude observed in *Zea mays* [31]. Duckweeds are distributed broadly around the world (Figure 4(a)). Our result shows that there is no significant overall correlation of genome size with latitude, longitude, and altitude (Figure 4). The same result was found in *Vicia faba* [32], *Sesleria albicans* [33], and *Asteraceae* [34]. A summary revealed that these relationships were not straightforward and not clear. Five studies (*Picea sitchensis, Berberis, Poaceae,* and *Fabaceae*, Tropical versus temperate grasses, 329 tropical versus 527 temperate plants) found positive, seven (*Arachis duranensis, Festuca arundinacea,* North American cultivars of *Zea mays,* 162 British plants, 23 Arctic plants, 22 North American *Zea mays*, and 11 North American *Zea mays*) found negative, and five (*Allium cepa, Dactylis glomerata,* and *Helianthus*) found nonsignificant correlations between genome size and latitude. Additionally, nine were positive, eight were negative, and six were not statistically significant between genome size and altitude [35]. But

the different environmental distribution of the *Lemna* genus (30° to 60° of latitude and below 600 m of altitude) with the other four duckweed genera (0° to 45° of latitude and 600 m to 1200 m of altitude) might explain the large intraspecific genome size variation.

4.3. Intraspecific Variation in Genome Size. Intraspecific genome consistency has been reported in *Allium cepa* [36], *Glycine max* [37], and *Capsicum* and *Gossypium* [38]. We also found a similar result for *Spirodela polyrhiza*, *Landoltia punctata*, *Wolffiella hyalina*, *Wolffiella lingulata*, and *Wolffia australiana*, which do not have statistical intraspecific differences in genome size (Table S1). One explanation is that these species have a mechanism to maintain genome size constancy, for example, by intraspecific stabilizing selection on genome size [39]. On the other hand, we found obvious intraspecific variation in *Lemna minor*, *Lemna aequinoctialis*, *Lemna trisulca*, and *Lemna japonica*. Some artifacts of intraspecific variation in genome size have been noted, such as environmentally induced variations, secondary compounds and fluorescence staining inhibitor, and erroneously determined species [15, 40]. However, our experiments are not complicated by these factors. We developed an easy bar-coding method to correctly identify duckweed species, which allowed us to correct any misnamed duckweed in the collection [17]. As cytosolic components may change in response to changes in the environment, we grew the duckweed plants under identical conditions. We used internal standardization such as *Brachypodium distachyon*, *Arabidopsis thaliana,* and *Physcomitrella patens* that were prepared simultaneously and under the same experimental conditions as the duckweed accessions. Both duckweed and the internal standard have very little secondary compounds, which may affect genome size estimates. Additionally, we performed biological replicate on different days to eliminate instrument bias. In addition, intraspecific differences were independently confirmed by simultaneously measuring two accessions of the same species by FCM (Figures 2(b) and 2(c)).

The intraspecific variation may result from different numbers of repeated sequences, including satellite DNA [41], transposable elements [42], and ribosomal genes [43]. Large-scale polymorphism of heterochromatic repeats exist in the DNA of *Arabidopsis thaliana* and could account for about 50% of the variance among the *Arabidopsis thaliana* accession [44]. In addition, the amount of rDNA accounts for the differences in genome size between closely related lines of *Linum usitatissimum* (flax) [43]. The activity of transposable elements (TE) potentially multiply 20∼100 times (∼0.1–1 Mbp) in a single generation [45]. For example, the BARE-1 TE is positively correlated with genome size within wild barley (*Hordeum spontaneum*) in response to sharp microclimatic divergence [42]. Deletions and insertions (INDELs) are most likely not candidates for genome size differences in duckweed. In *Drosophila melanogaster*, genome loss is only less than 1 bp per generation [46], indicating a small contribution to genome-size variation. However, in the fast growing duckweeds, which only need 2∼5 days for each generation, one could imagine it is more likely that TE have higher rate than other flowering plants to influence genome size within and between species.

5. Conclusion

This is the first extensive analysis of genome sizes in duckweeds and examination of genome size variations across a range of taxonomic levels. We showed that duckweeds, in general, have remarkable smaller genome size compared with other flowering plants. The smallest genome size of *Spirodela polyrhiza*, combined with its sterile and controllable culture, fast growing, and promising application in research, suggest that this species may be good candidates for ongoing whole-genome sequencing projects and a model experimental tool. The 150 Mbp *Spirodela polyrhiza* genome is being sequenced by the DOE-JGI community-sequencing program (CSP), which will address challenges in alternative energy, bioremediation, and global carbon cycling. Also, the availability of a DNA C-values database of duckweeds and a consensus higher-level phylogenetic tree has opened the way for exploring the general processes underlying the evolution of genomes. Obvious intraspecific variation in duckweeds will also provide nice material to study the mechanism of within-species and between-species variation in genome size. However, the main force driving the intraspecific variance and how the genome size affects the phenotype still requires more research.

Acknowledgments

We thank Elias Landolt from Institut für Integrative Biologie (Zürich, Switzerland), Lynn Dickey, Nirmala Rajbhandari, and Peaches Staton from BIOLEX (NC, USA), and Judy Acreman UTCC (Toronto, Canada) to generously provided duckweed accessions. We are grateful to Theresa Hyejeong Choi from EOHSI (Rutgers University) for assistance with flow cytometry. We would like to thank Justina Marcinow, Marina Ermakova and Yiheng Yan for experimental help. We thank Julia Trumbull for maintaining the Landolt Duckweed Collection at Rutgers. This work was supported by an Energy Fellowship and a Waksman Fellowship to W. Wang, and startup funds from the Waksman Institute of Microbiology to T. P. Michael.

References

[1] D. H. Les, D. J. Crawford, E. Landolt, J. D. Gabel, and R. T. Kimball, "Phylogeny and systematics of *Lemnaceae*, the duckweed family," *Systematic Botany*, vol. 27, no. 2, pp. 221–240, 2002.

[2] E. Landolt, *The Family of* Lemnaceae—*A Monographic Study*, vol. 1, Veroffentlichungen des Geobotanischen Institutes ETH, Stiftung Rubel, Zurich, Switzerland, 1986.

[3] W. Journey, P. Skillicorn, and W. Spira, *Duckweed Aquaculture: A New Aquatic Farming System for Developing Countries*, The World Bank, 1993.

[4] W. S. Hillman, "Calibrating duckweeds: light, clocks, metabolism, flowering," *Science*, vol. 193, no. 4252, pp. 453–458, 1976.

[5] A. M. Stomp, "The duckweeds: a valuable plant for biomanufacturing," *Biotechnology Annual Review*, vol. 11, pp. 69–99, 2005.

[6] W. S. Hillman, "Photoperiodism in *Lemna*: reversal of night-interruption depends on color of the main photoperiod," *Science*, vol. 154, no. 3754, pp. 1360–1362, 1966.

[7] K. Chaloupkova and C. C. Smart, "The abscisic acid induction of a novel peroxidase is antagonized by cytokinin in *Spirodela polyrrhiza* L," *Plant Physiology*, vol. 105, no. 2, pp. 497–507, 1994.

[8] R. A. Brain and K. R. Solomon, "A protocol for conducting 7-day daily renewal tests with *Lemna gibba*," *Nature Protocols*, vol. 2, no. 4, pp. 979–987, 2007.

[9] Y. T. Yamamoto, N. Rajbhandari, X. Lin, B. A. Bergmann, Y. Nishimura, and A. M. Stomp, "Genetic transformation of duckweed *Lemna gibba* and *Lemna minor*," *In Vitro Cellular and Developmental Biology—Plant*, vol. 37, no. 3, pp. 349–353, 2001.

[10] N. Ozengin and A. Elmaci, "Performance of duckweed (*Lemna minor* L.) on different types of wastewater treatment," *Journal of Environmental Biology*, vol. 28, no. 2, pp. 307–314, 2007.

[11] J. J. Cheng and A. M. Stomp, "Growing duckweed to recover nutrients from wastewaters and for production of fuel ethanol and animal feed," *Clean—Soil, Air, Water*, vol. 37, no. 1, pp. 17–26, 2009.

[12] R. Lister, B. D. Gregory, and J. R. Ecker, "Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond," *Current Opinion in Plant Biology*, vol. 12, no. 2, pp. 107–118, 2009.

[13] E. Landolt, *The Family of* Lemnaceae—*A Monographic Study*, vol. 2, Veroffentlichungen des Geobotanischen Institutes ETH, Stiftung Rubel, Zurich, Switzerland, 1986.

[14] G. Geber, *Zur Karyosystematik der Lemnaceae*, University of Vienna, Vienna, Austria, 1989.

[15] J. Doležel and J. Bartoš, "Plant DNA flow cytometry and estimation of nuclear genome size," *Annals of Botany*, vol. 95, no. 1, pp. 99–110, 2005.

[16] M. D. Bennett and I. J. Leitch, "Nuclear DNA amounts in angiosperms: progress, problems and prospects," *Annals of Botany*, vol. 95, no. 1, pp. 45–90, 2005.

[17] W. Wang, Y. Wu, Y. Yan, M. Ermakova, R. Kerstetter, and J. Messing, "DNA barcoding of the *Lemnaceae*, a family of aquatic monocots," *BMC Plant Biology*, vol. 10, p. 205, 2010.

[18] J. Doležel, J. Greilhuber, and J. Suda, "Estimation of nuclear DNA content in plants using flow cytometry," *Nature Protocols*, vol. 2, no. 9, pp. 2233–2244, 2007.

[19] D. W. Galbraith, K. R. Harkins, J. M. Maddox, N. M. Ayres, D. P. Sharma, and E. Firoozabady, "Rapid flow cytometric analysis of the cell cycle in intact plant tissues," *Science*, vol. 220, pp. 1049–1051, 1983.

[20] K. Arumuganathan and E. D. Earle, "Nuclear DNA content of some important plant species," *Plant Molecular Biology Reporter*, vol. 9, no. 3, pp. 208–218, 1991.

[21] G. Schween, G. Gorr, A. Hohe, and R. Reski, "Unique tissue-specific cell cycle in *Physcomitrella*," *Plant Biology*, vol. 5, no. 1, pp. 50–58, 2003.

[22] J. Doležel, J. Bartoš, H. Voglmayr, J. Greilhuber, and R. A. Thomas, "Nuclear DNA content and genome size of trout and human," *Cytometry A*, vol. 51, no. 2, pp. 127–129, 2003.

[23] E. Landolt, *Biosystematic Investigations in the Family of Duckweeds (Lemnaceae)*, Veroffentlichungen des Geobotanischen Institutes ETH, Stiftung Rubel, Zurich, Switzerland, 1980.

[24] M. J. Hegarty and S. J. Hiscock, "Genomic clues to the evolutionary success of polyploid plants," *Current Biology*, vol. 18, no. 10, pp. R435–R444, 2008.

[25] C. Seoighe and C. Gehring, "Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome," *Trends in Genetics*, vol. 20, no. 10, pp. 461–464, 2004.

[26] A. R. Leitch and I. J. Leitch, "Genomic plasticity and the diversity of polyploid plants," *Science*, vol. 320, no. 5875, pp. 481–483, 2008.

[27] K. L. Adams and J. F. Wendel, "Polyploidy and genome evolution in plants," *Current Opinion in Plant Biology*, vol. 8, no. 2, pp. 135–141, 2005.

[28] S. P. Otto, "The evolutionary consequences of polyploidy," *Cell*, vol. 131, no. 3, pp. 452–462, 2007.

[29] M. C. J. Bottini, E. J. Greizerstein, M. B. Aulicino, and L. Poggio, "Relationships among genome size, environmental conditions and geographical distribution in natural populations of NW patagonian species of *Berberis* L. (*Berberidaceae*)," *Annals of Botany*, vol. 86, no. 3, pp. 565–573, 2000.

[30] D. Ohri and T. N. Khoshoo, "Genome size in gymnosperms," *Plant Systematics and Evolution*, vol. 153, no. 1-2, pp. 119–132, 1986.

[31] A. L. Rayburn and J. A. Auger, "Genome size variation in *Zea mays* ssp. mays adapted to different altitudes," *Theoretical and Applied Genetics*, vol. 79, no. 4, pp. 470–474, 1990.

[32] M. Ceccarelli, S. Minelli, F. Maggini, and P. G. Cionini, "Genome size variation in *Vicia faba*," *Heredity*, vol. 74, no. 2, pp. 180–187, 1995.

[33] M. A. Lysak, A. Rostkova, J. M. Dixon, G. Rossi, and J. Dolezel, "Limited genome size variation in *Sesleria albicans*," *Annals of Botany*, vol. 86, pp. 399–403, 2000.

[34] J. Chrtek Jr., J. Zahradnícek, K. Krak, and J. Fehrer, "Genome size in *Hieracium* subgenus *Hieracium* (*Asteraceae*) is strongly correlated with major phylogenetic groups," *Annals of Botany*, vol. 104, no. 1, pp. 161–178, 2009.

[35] C. A. Knight, N. A. Molinari, and D. A. Petrov, "The large genome constraint hypothesis: evolution, ecology and phenotype," *Annals of Botany*, vol. 95, no. 1, pp. 177–190, 2005.

[36] M. D. Bennett, S. Johnston, G. L. Hodnett, and H. J. Price, "*Allium cepa* L. cultivars from four continents compared by flow cytometry show nuclear DNA constancy," *Annals of Botany*, vol. 85, no. 3, pp. 351–357, 2000.

[37] J. Greilhuber and R. Obermayer, "Genome size and maturity group in Glycine max (soybean)," *Heredity*, vol. 78, no. 5, pp. 547–551, 1997.

[38] B. Hendrix and J. M. Stewart, "Estimation of the nuclear DNA content of *Gossypium* species," *Annals of Botany*, vol. 95, no. 5, pp. 789–797, 2005.

[39] P. Šmarda, L. Horová, P. Bureš, I. Hralová, and M. Marková, "Stabilizing selection on genome size in a population of *Festuca pallens* under conditions of intensive intraspecific competition," *New Phytologist*, vol. 187, no. 4, pp. 1195–1204, 2010.

[40] J. Greilhuber, "Intraspecific variation in genome size in angiosperms: identifying its existence," *Annals of Botany*, vol. 95, no. 1, pp. 91–98, 2005.

[41] G. Bosco, P. Campbell, J. T. Leiva-Neto, and T. A. Markow, "Analysis of *Drosophila* species genome size and satellite DNA

content reveals significant differences among strains as well as between species," *Genetics*, vol. 177, no. 3, pp. 1277–1290, 2007.

[42] R. Kalendar, J. Tanskanen, S. Immonen, E. Nevo, and A. H. Schulman, "Genome evolution of wild barley (*Hordeum spontaneum*) by BARE-1 retrotransposon dynamics in response to sharp microclimatic divergence," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 12, pp. 6603–6607, 2000.

[43] C. A. Cullis, "Mechanisms and control of rapid genomic changes in flax," *Annals of Botany*, vol. 95, no. 1, pp. 201–206, 2005.

[44] J. Davison, A. Tyagi, and L. Comai, "Large-scale polymorphism of heterochromatic repeats in the DNA of Arabidopsis thaliana," *BMC Plant Biology*, vol. 7, article 44, 2007.

[45] D. A. Petrov, "Evolution of genome size: new approaches to an old problem," *Trends in Genetics*, vol. 17, no. 1, pp. 23–28, 2001.

[46] D. A. Petrov and D. L. Hartl, "High rate of DNA loss in the *Drosophila melanogaster* and *Drosophila virilis* species groups," *Molecular Biology and Evolution*, vol. 15, no. 3, pp. 293–302, 1998.

BMC Plant Biology

RESEARCH ARTICLE Open Access

DNA barcoding of the *Lemnaceae*, a family of aquatic monocots

Wenqin Wang, Yongrui Wu, Yiheng Yan, Marina Ermakova, Randall Kerstetter, Joachim Messing*

Abstract

Background: Members of the aquatic monocot family *Lemnaceae* (commonly called duckweeds) represent the smallest and fastest growing flowering plants. Their highly reduced morphology and infrequent flowering result in a dearth of characters for distinguishing between the nearly 38 species that exhibit these tiny, closely-related and often morphologically similar features within the same family of plants.

Results: We developed a simple and rapid DNA-based molecular identification system for the *Lemnaceae* based on sequence polymorphisms. We compared the barcoding potential of the seven plastid-markers proposed by the CBOL (Consortium for the Barcode of Life) plant-working group to discriminate species within the land plants in 97 accessions representing 31 species from the family of *Lemnaceae*. A *Lemnaceae*-specific set of PCR and sequencing primers were designed for four plastid coding genes (*rpoB, rpoC1, rbcL* and *matK*) and three noncoding spacers (*atpF-atpH, psbK-psbI* and *trnH-psbA*) based on the *Lemna minor* chloroplast genome sequence. We assessed the ease of amplification and sequencing for these markers, examined the extent of the barcoding gap between intra- and inter-specific variation by pairwise distances, evaluated successful identifications based on direct sequence comparison of the "best close match" and the construction of a phylogenetic tree.

Conclusions: Based on its reliable amplification, straightforward sequence alignment, and rates of DNA variation between species and within species, we propose that the *atpF-atpH* noncoding spacer could serve as a universal DNA barcoding marker for species-level identification of duckweeds.

Background

The cost of DNA purification and sequencing has dropped considerably in recent years so that identification of individual species by DNA barcoding has become an independent, subtler method than solely morphological-based classification to distinguish closely related species, which also defines the systematic relationships by analysis of genetic distance. The key element for a robust barcode is a suitable threshold between inter- and intra-specific genetic distances. Sequence variation between species has to be high enough to tell them apart while the distances within species must be low enough for them to cluster together [1]. The mitochondrial coxidase subunit I (*COI*) gene has proven to be a reliable, cost-effective, and easily recovered barcode marker to successfully identify animal species [2-4], but its application in the plant kingdom is impeded by a slow nucleotide substitution rate, which is insufficient for the diagnosis of individual species [5,6]. However, the Consortium for the Barcode of Life (CBOL) plant-working group recently proposed seven leading candidate sequences for use as barcoding markers [7]. Four plastid coding genes (*rpoB, rpoC1, rbcL* and *matK*) and three noncoding spacers (*atpF-atpH*, *psbK-psbI* and *trnH-psbA*) have been selected based on previous investigations among different plant families [8-10]. However, the utility of each of these sequences for individual families of species within the plant kingdom is hardly predictable [11,12].

Although there have been attempts to use the single-locus of *matK* [8], a combination of two loci, *rbcL* and *trnH-psbA* [9], and even multi-loci combinations [13] as barcoding sequences, the use of a unified barcode for the identification of all the land plants would be difficult due to conflicting needs of different researchers. For example, an optimal barcode marker that has been determined empirically to distinguish plants at the family level may prove less useful for making accurate

* Correspondence: messing@waksman.rutgers.edu
Waksman Institute of Microbiology, Rutgers University, 190 Frelinghuysen Road, Piscataway, NJ 08854, USA

species level identifications. Most of the proposed plant barcode markers were designed primarily for identifying distantly related organisms in biodiversity hotspots such as Panama [14] and Kruger National Park in South Africa [8]. So far, little attention and only a few studies have been devoted to developing unified barcodes suitable for making identifications within a family, within a genus, or between closely related sister species. A test of seven other candidate barcoding sequences in the family of *Myristicaceae* was applied to eight species within a genus and yielded two suitable barcodes [15]. Recently, it has been shown that all three markers (*rbcL*, *trnH-psbA* and *matK*) can discriminate 4 sister species of *Acacia* across three continents [16]. The marker *matK* has been reported to distinguish 5 *Dendrobium* species [17]. More complex approaches have been developed at the subfamily level identification of larger groups of related plants [18]. Although an extensive barcode study for 31

Carex species suggested that a single locus or even multiple loci cannot provide a resolution of greater than 60%, it did not include some of the new markers (*atpF-atpH* and *psbK-psbI*) [19]. When *atpF-atpH* and *psbK-psbI* were included for distinguishing *Carex* and *Kobresia*, it could be shown that *matK* identifies 95% as single-locus or 100% of the species when combined with another marker. However, this study used material from a well defined regional perspective, the Canadian Arctic Archipelago, where the number of co-existing closely related species is limited [20]. Our objective was to determine whether one or more of the markers proposed by the CBOL plant-working group would serve as an optimal marker for species-level identification within the family *Lemnaceae*.

The members of the family *Lemnaceae*, commonly called duckweeds, comprise 38 species in five genera [21]. They are all aquatic plants that grow on or below the surface of the water all over the world and they include the smallest flowering plants [22]. They are ideal material for physiological, biochemical, and genomic studies because of their direct contact with medium, rapid growth and relatively small genome sizes [22]. They are valuable means for biomanufacturing through genetic engineering technology and due to the recent progress towards duckweed-based commercial products [23]. They can be easily maintained by vegetative reproduction in aseptic cultivation for decades [23]. The small size of the plant is ideal for maintaining diverse accessions and therefore for evolutionary studies at the DNA level. Some species, such as *Lemna minor*, are used by the Environmental Protection Agency for measuring water quality because their growth rates are sensitive to a wide range of environmental contaminants such as metals, nitrates, and phosphates [24]. Indeed, wastewater treatment with duckweed has been proposed as a "green" way to remediate municipal water supplies [25]. Rapid growth also offers practical applications of duckweeds as a biofuel crop. Some duckweeds form starch-rich over-wintering fronds called turions, which can be easily induced from vegetative fronds by treatment of cold shock, starvation, or with abscisic acid [26,27]. Resulting from their size and density, both vegetative fronds and turions are much more easily harvested than microalgae [28], which make duckweeds an attractive feedstock for bioethanol production that does not compete for agriculturally productive land.

Given these potential uses, the 160-Mb *Spirodela polyrhiza* genome has been selected for whole genome sequencing by the DOE-JGI community-sequencing program (CSP). A reference genome within this family will be invaluable for gene discovery and evolutionary analysis of aquatic monocot species. Furthermore, from a systematic point of view, classification solely based on morphological characteristics has been a significant challenge. The most readily observed anatomical feature of the minute and highly reduced duckweeds are their fronds with or without roots. These few and somewhat variable morphological characters and rarely emerging flowers or fruits make identification of duckweeds extremely difficult even for professional taxonomists [29]. Complementing traditional classification methods with a DNA-based method would be highly applicable for such a family of species. It would permit these species to be classified in a highly reproducible and cost effective manner because DNA-based methods are independent of morphology, integrity, and developmental stage of the organism and can distinguish among species that superficially look alike [30].

Here, we present a simple and accessible protocol to barcode duckweeds and establish a sequence database against which unknown species may be compared and tentative species identifications can be validated. This database also provides a high-resolution phylogenetic resource for this important plant monocot family.

Results

Sampling criteria

The duckweed family consists of 38 species classified into 5 genera [21]. A worldwide collection has been characterized by genome sizes (Wang et al., ms. in prep.). From this collection, 97 ecotypes were sampled for the current work representing all five genera and 31 species (81.6% of the known species; Additional file 1). The ecotypes selected encompass the worldwide geographical distribution of duckweeds originating from different climates and geographical regions, ranging from N60° to S42° latitude and 9 m to 1287 m in altitude (Additional file 1, Figure 1). 85 ecotypes from 19 species were used for statistical calculations and candidate

Figure 1 Google map of the worldwide collection of duckweeds for the current study. The distribution of duckweeds was made by GPS with corresponding latitude and longitude.

barcode evaluations. An additional 12 single-ecotype species were examined to determine the broader applicability of the barcode markers for identification.

Validation of DNA barcoding markers

To simplify identification of different species by DNA barcodes, a target DNA sequence marker has to meet two basic requirements: the first is a high success rate during PCR amplification and DNA sequencing, the second is sufficient DNA sequence polymorphism to permit different species to be distinguished and evolutionary distances between them to be calculated [1]. The CBOL plant-working group proposed 7 leading candidates [7], i.e., 4 coding genes (*rpoB, rpoC1, rbcL* and *matK*) and 3 noncoding spacers (*atpF-atpH, psbK-psbI* and *trnH-psbA*). To evaluate the seven markers, genomic DNA extracted from the 97 ecotypes was subjected to PCR amplification with the primer pairs based on the chloroplast sequence of *Lemna minor*. The PCR primers were also used for sequencing (See Materials and methods). PCR and sequencing were generally successful (≥95%) for all the barcode candidates except *matK* (71%) (Table 1). The maximal and minimal alignment length of PCR product for *rpoB*, *rpoC1*, *rbcL* and *matK* were identical, while that of *atpF-atpH*, *psbK-psbI* and *trnH-psbA* were quite variable, with a range of 579-622 bp, 185-576 bp and 286-504 bp, respectively. It was not unexpected that the coding markers (*rpoB*, *rpoC1*, *rbcL* and *matK*) were conserved in PCR product length, while the noncoding spacers (*atpF-atpH, psbK-psbI* and *trnH-psbA*) displayed more variability due to extensive insertions/deletions (Table 1). These results indicate that the selection of markers by the COBL plant-working group should provide a reasonable level of success for new untested plant families.

Intra- and inter-specific DNA sequence polymorphism

To assess the degree of DNA polymorphism between DNA samples, sequence divergences between and within species were calculated by Kimura 2-parameter (K2P) and uncorrected p-distance, respectively. Both models exhibited the same tendency: higher average interspecific diversity and lower intraspecific distance. For example, the K2P distance within and between species is as follows: *psbK-psbI* (0.1648 and 0.0072), *trnH-psbA* (0.1133 and 0.0058), *matK* (0.0715 and 0.0019), *atpF-atpH* (0.0633 and 0.0008) *rpoB* (0.0388 and 0.0069), *rpoC1* (0.0303 and 0.0006), *rbcL* (0.0216 and 0.0004). The

Table 1 Success ratios of PCR amplification and sequencing for seven candidate barcoding markers

	psbK-psbI	*trnH-psbA*	*matK*	*atpF-atpH*	*rpoB*	*rpoC1*	*rbcL*
Max. length of product*	576	504	725	622	389	450	522
Min. length of product*	185	286	719	579	389	450	522
# tested Samples	97	97	97	97	97	97	97
% Success of PCR and sequencing	100%	95%	71%	99%	98%	100%	100%

* The analyzed product length becomes shorter than corresponding one's due to removal of the end of ambiguous nucleotides

noncoding spacer *psbK-psbI* showed the highest interspecific diversity (66 average substitution sites among 675 bp), while the coding marker *rbcL* is the most conserved one (11 average substitution sites among 522 bp) (Table 2). Wilcoxon signed rank tests further showed that the most variable barcode between species was *psbK-psbI*, followed by *trnH-psbA*, *matK* and *atpF-atpH* (Additional file 2). The lowest intraspecific distance was provided by *atpF-atpH* and *rbcL*, whereas the highest is *trnH-psbA*, *psbK-psbI* and *matK* (Additional file 3). Although none of the seven proposed markers possessed both the highest variation between species and the lowest distance within a species, *atpF-atpH* seemed to show sufficient interspecific but relatively low intraspecific divergence, compared to the other six markers (Table 2, Additional file 2 and 3).

The accuracy of barcoding for species identification depended to a large extent on the barcoding gap between intraspecific and interspecific sequence variations. Effective barcoding became weaker when interspecific and intraspecific distances overlapped. To evaluate whether there was a significant barcoding gap, we calculated the distribution of divergences for the seven markers (Figure 2). Median and Mann-Whitney U tests inferred that the mean of intraspecific divergence was significantly lower than that of interspecific distance in each case ($p < 0.0001$). Even though *psbK-psbI* and *trnH-psbA* exhibited the highest rates of divergence between species, they were also most diverged within species, which could easily result in misidentification (Table 2, Additional file 3 and 4, Figure 2). On the other hand, the adequate variation and the narrow overlapping distance of the *atpF-atpH* marker would ensure accurate ecotype and species identification (Table 2, Additional file 2 and 3, Figure 2).

DNA sequence similarity-based identification

In order to test whether accurate species identification can be made in our samples, we adopted the "best match" function in the program TAXONDNA [31]. The rank order for the correct identification is *atpF-atpH* (92.85%) *psbK-psbI* (84.7%), *trnH-psbA* (82.5%), *matK* (77.77%), *rpoB* (77.5%), *rpoC1* (70.58%), *rbcL* (70.58%) (Table 3). Generally, the three noncoding spacers produced higher rates of successful identifications than those of the four coding markers. Consistent with Figure 3, *atpF-atpH* yielded the best result with 92.85% successful identifications. Among 84 ecotypes (not including species with single sampled ecotypes), 78 samples were successfully discriminated, three were ambiguous and three were incorrectly identified using *atpF-atpH*. When we combined *atpF-atpH* with one of the other five barcoding markers, the percentage of correct identification dropped, except for *psbK-psbI*, which gave an increase of 1.19% (Table 3). The markers *matK* + *atpF-atpH* were not counted because of the small number of sequence comparisons done with *matK*.

Tree-based sequence classification

As an alternative to sequence similarity-based identification, we estimated the proportion of recovered monophyly from multiple conspecific ecotypes per species in the phylogenetic tree for each barcoding marker. Here, we need to stress that the primary purpose of the tree is not so much the evolutionary relationship, but the species identification. The *atpF-atpH* attained the highest score of monophyletic species (73.7%, i.e., 14 correctly identified out of 19 species; Table 4 and Figure 3). The number of successfully identified species with the other six markers was *rpoB* (11), *rpoC1* (11), *rbcL* (11), *trnH-psbA* (10), *psbK-psbI* (8). The *atpF-atpH* marker did not distinguish closely-related pairs of sister species such as *W. gladiata* and *W. oblonga* and *L. minuta* and *L. valdiviana*.

Although the location of most grouped ecotypes in the taxonomic trees did not change in regard to each marker, a close examination consistently revealed two interesting connections. First, despite the fact that very little

Table 2 Measurement of inter- and intra-specific divergences for seven barcoding markers

Region	*psbK-psbI*	*trnH-psbA*	*matK*	*atpF-atpH*	*rpoB*	*rpoC1*	*rbcL*
Aligned length (bp)*	675	520	725	674	389	450	522
Mean interspecific No. of substitution	66	32	48	44	13	13	11
Mean interspecific Kimura 2-parameter distances	0.1648 ± 0.0221	0.1133 ± 0.0120	0.0715 ± 0.0061	0.0633 ± 0.0068	0.0338 ± 0.0051	0.0303 ± 0.0050	0.0216 ± 0.0038
Mean interspecific Kimura 2-parameter distances	0.0072 ± 0.0015	0.0058 ± 0.0014	0.0019 ± 0.0003	0.0008 ± 0.0002	0.0069 ± 0.0008	0.0006 ± 0.0002	0.0004 ± 0.0002
Mean interspecific P-distances	0.1435 ± 0.0156	0.0986 ± 0.0095	0.0671 ± 0.0052	0.0601 ± 0.0059	0.0327 ± 0.0048	0.0295 ± 0.0048	0.0212 ± 0.0037
Mean interspecific P-distances	0.0066 ± 0.0012	0.0057 ± 0.0014	0.0019 ± 0.0003	0.0008 ± 0.0002	0.0062 ± 0.0007	0.0006 ± 0.0002	0.0004 ± 0.0002

* Aligned length becomes longer than corresponding ones due to addition of the gap.

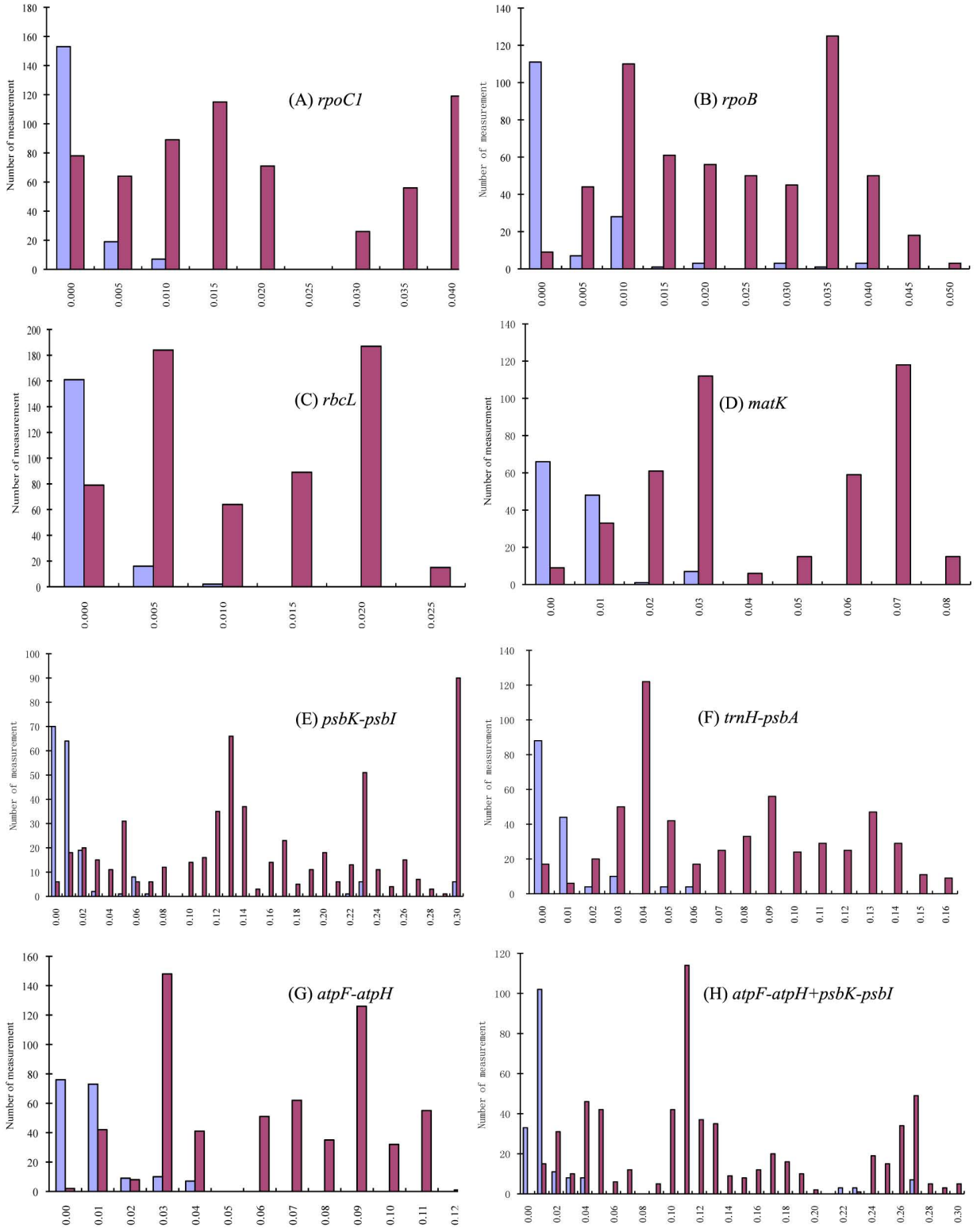


Figure 2 Relative distribution of all intra- and inter-specific divergence for single or combined markers. (A) *rpoC1*. (B) *rpoB*. (C) *rbcL*. (D) *matK*. (E) *psbK-psbI*. (F) *trnH-psbA*. (G) *atpF-atpH*. (H) *atpF-atpH+ psbK-psbI*. × axis is uncorrected p-distance with corresponding increment unit based on variation of each marker. Y axis is the number of occurrences. Barcoding gaps were evaluated with high significance ($p < 0.0001$) by Median and Mann-Whitney U tests for all markers. Blue bars indicate intraspecific distance and red bars are interspecific distance.

Table 3 Identification success based on "best close match" tools

	psbK-psbI	*trnH-psbA*	*matK*	*atp-atpH*	*rpoB*	*rpoCl*	*rbcL*	*psbK-psbI + atp F-atpH*	*trnH-psbA + atp F-atpH*	*matK+ atpF-atpH*	*rpoB +atp F-atpH*	*rpoCl + atp F-atp H*	*rbcL + atpF-atpH*
Correct	72 (84.7%)	66 (82.5%)	49 (77.77%)	78 (92.85%)	62 (77.5%)	60 (70.58%)	60 (70.58%)	79(94.04%)	71(89.87%)	/	77 (91.66%)	77 (91.66%)	77 (91.66%)
Ambiguous	8 (9.41%)	11 (13.75%)	10 (15.87%)	3(3.57%)	12 (15.0%)	21 (24.7%)	21 (24.7%)	0(0.0%)	3(3.79%)	/	2(2.38%)	4(4.76%)	4(4.76%)
Incorrect	5 (5.88%)	2(2.5%)	4(6.34%)	3(3.57%)	6(7.5%)	4(4.7%)	2(2.35%)	5(5.95%)	5(6.32%)	/	5(5.95%)	3(3.57%)	3(3.57%)
No match	0(0.0%)	1(1.25%)	0(0.0%)	0(0.0%)	0(0.0%)	0(0.0%)	2(2.35%)	0(0.0%)	0(0.0%)	/	0(0.0%)	0(0.0%)	0(0.0%)
Threshold	22.12%	4.01%	2.62%	2.96%	2.57%	0.44%	0.38%	22.16%	2.08%	/	2.44%	1.77%	1.67%

"best close match" was analyzed by TAXONDNA program [31] with single region or two-region combinations. The ecotypes was classified into correct, ambiguous, incorrect and no match group. The group number was shown in each well. Number in bracket indicates percentage in all barcoding ecotypes. *matK + atpF-atpH* was not counted due to the small number of sequence comparison done for *matK*. Percentage in the bracket was calculated by dividing each item by all tested sample.

is known about how cross pollination in these tiny flowering plants occurs, *L. japonica* has been suspected to originate from a hybridization event between *L. minor* and *L. turionifera* based on morphological characters [22]. Our data indicates that sequence from each of the seven tested markers of *L. japonica* 7182 was always identical to and clustered with *L. minor* (Figure 3). Since the chloroplast is maternally inherited in many (but not all) plants, our data is consistent with *L. japonica* arising from a cross between *L. minor* and *L. turionifera*.

The second connection was *S. polyrhiza* 9203, which consistently clusters with *S. intermedia* rather than other *S. polyrhiza* in all seven tested markers (Figure 3). We examined 34 ecotypes of *S. polyrhiza* from the collection using the *atpF-atpH* marker and found four additional ecotypes that grouped closely with *S. intermedia* (Additional file 4). This suggested that these accessions might have been misidentified as *S. polyrhiza* due to the overlap in morphological characteristics between these species.

Discussion

Here, we present data validating the most useful DNA barcoding markers for the family of *Lemnaceae* from among those proposed by the CBOL plant-working group. Such a fundamental, whole family-wide analysis lays the groundwork for phylogenetic and genomic studies. Our samples represent a worldwide collection from the same family with many sister species (Figure 1 and 3, Additional file 1). Specimens in previous taxonomic classifications using barcoding markers were mainly from distantly related groups from broadly different families that originated from the local or more defined regions, such as the National Park [8], the Amazon [32], and the Panama region [14]. Because of the diversity of the collection that has accumulated over the years, duckweeds provide a unique system to test the proposed barcoding markers for closely related species. Furthermore, it is difficult to classify members of this family by morphology alone. Therefore, we can not only validate the universal application of barcoding markers, but also apply it to species that may be solely dependent on such an approach for conservation. The advantage of universal barcoding markers is the design of universal primers for barcoding markers from reference sequences, which in this case was *L. minor* [33]. The primers worked very well for all the samples (31 species and 97 ecotypes) with PCR amplification and the sequencing success rates better than 95%, except in the case of *matK*, which yielded a rate as low as 71% (Table 1). In addition, a lower PCR annealing temperature than optimal for *Lemna minor* permits primers to anneal to the target sequences despite sequence polymorphism in related species. It is interesting that most PCR failure existed in the *Wolffioideae* subfamily (Additional file 1). The locus *matK* has been shown to be very variable in numerous phylogenetic studies [34,35]} and other studies have also noted the difficulties of its utilization due to PCR failure and lack of truly universal primer sites [9,10]. Further improvement of primer designs for matK for other targets could increase amplification success, but might fail because of less conserved sites near the most variable sequences of the locus. Although *matK*

Table 4 Number of monophyletic species recovered with the best two phylogenetic methods for six markers

Loci	UPGMA	MP
psbK-psbI	8 (93.3)	8 (87.5)
trnH-psbA	10 (87.5)	10 (85.7)
matK	/	/
atpF-atpH	14 (100)	14 (94.1)
rpoB	11 (83.3)	11 (68.8)
rpoC1	11 (85.7)	11 (68.8)
rbcL	11 (85.7)	12 (68.8)

The number of monophyletic species out19 species was shown in each well. Proportions supported by bootstrap >50% are in brackets.

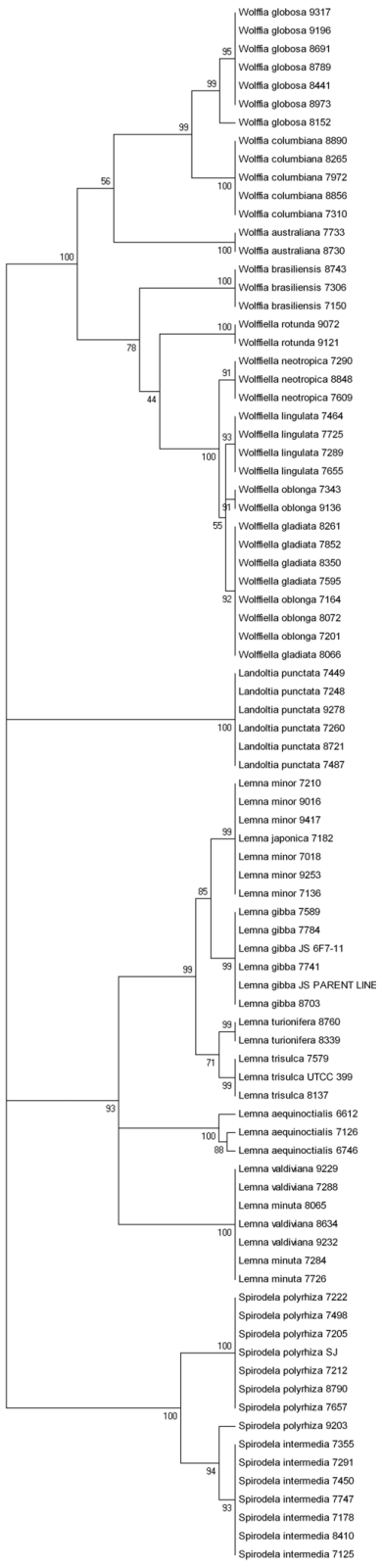


Figure 3 UPGMA tree based *atpF-atpH* sequences. The tree was drawn among 20 species with more than one ecotype except *L. japonica*.

DNA sequences exhibited the highest interspecific variation among the four coding markers (Table 2), the low percentage of successful PCR amplification and sequencing in duckweeds would restrict its extensive use.

It was not surprising that the noncoding spacers showed dramatically higher sequence variability than the coding markers (Table 2). Given the slow evolutionary rate of *rpoB*, *rpoC1* and *rbcL* (especially for *rbcL*, which is strongly recommended for barcoding across all land plants), they work well to distinguish distantly related species either alone or when combined with other more variable regions [6,9]. However, their sequence polymorphisms might not be sufficient to distinguish closely related species. The non-coding spacers of *psbK-psbI* and *trnH-psbA* were the most polymorphic plastid sequences with variable sequence length in duckweeds (Table 1). The size of *trnH-psbA* in *Spirodela* (~504 bp) was 218 bp longer than in the other four genera (~286 bp). The length of the *psbK-psbI* sequence was the most variable, ranging from ~185 bp in *S. polyrhiza* to ~479 bp in *S. intermedia* even though they were sister-species (Table 1 and Figure 3). These significant length variations caused by deletion/insertion, simple sequence repeats and rearrangements were problematic for accurate alignment, but could potentially be adapted for simple diagnostic tests that would not require DNA sequencing. Furthermore, the high sequence polymorphisms of the aligned sequences of *psbK-psbI* and *trnH-psbA* could offer greater distinction between species in a diverse set of genera in certain families [5,8]. Still, one has to use caution for intraspecies comparison where the relatively higher intraspecific distance compromised their power in barcoding duckweed species. One nearly has to cluster samples into two groups, one for ecotypes of the same species and one for species to species comparison (Table 2, 3, and 4, Figure 3). Failure to do so would prevent the detection of true differences between congeneric species and conspecific ecotypes and therefore impede the use of a universal duckweed barcode (Figure 2).

Although previous studies showed that *atpF-atpH* as a barcoding marker was inferior to *psbK-psbI*, *trnH-psbA* and *matK* based on distantly related species [5,8,9], our data suggested that it was the most promising barcoding marker for duckweeds with respect to high PCR amplification, ease of alignment, and sufficient sequence divergence (Table 1, 2, 3, 4 and Figure 2). Therefore, our data differed from the conclusions of evaluating barcoding markers made from unrelated species. Although it was shown that barcoding plants by more than one region tended to be more effective [11-13], combination of *atpF-atpH* with any of the other markers resulted in only slight increases or drops of the rate of successful identification of species compared to itself alone (Table 3), indicating that the

discriminatory power of *atpF-atpH* has already reached an optimum. When the *atpF-atpH* marker was combined with other markers, the reduced resolution lowered the differential value without complementary benefits. A similar finding that a combination of *matK* and *trnH-psbA* did not improve species identification has been reported as well [8].

One of the most significant applications of DNA barcoding is to overcome taxonomic obstacles, where it is difficult to identify unknown or wrongly named species in a family with similar morphology (Figure 3). Furthermore, DNA barcoding could offer us a primary screen for further characterization of cryptic species. Although scientists within the duckweed community were trying to resolve the question of whether *L. japonica* (Lj) originated from hybridization of *L. minor* (Lm) and *L. turionifera* (Lt), preliminary attempts to cross Lm and Lt (50 crosses) to reproduce the hybridization event were not successful [22]. The key problem is that flowering is very rare and the flower is small in size, which makes outcrossing extremely tedious [23]. Here, the sequences from the seven tested chloroplast markers of *L. japonica* 7182 were always identical and clustered with *L. minor* (Figure 3). Therefore, we used the limited nuclear markers (glyceraldehyde-3-phosphate dehydrogenase, histone 3 gene, beta-1,2-xylosyltransferase isoform 1, expression control elements from the *Lemnaceae* family) to uncover the relationship among them by polymorphisms. Unexpectedly, the sequences showed great conservation and there was not sufficient variation to answer this question. However, the identical alleles in *L. japonica* 7182 and *L. minor* support the assumption that *L. japonica* might have come from the cross of *L. minor* and *L. turionifera*.

Generally speaking of members of the duckweed family, the more derived they are, the simpler their morphologies. The reduction in size and simplification in structure make the fronds more mobile and better successfully adapt to variable conditions [22]. *S. intermedia* was characterized by a slight degree of primitivism of more nerves, roots, and ovules compared to *S. polyrhiza*, which suggested that *S. intermedia* was differentiated into *S. polyrhiza* potentially through gradual morphological reduction and isolation. However, gradual differences were sometimes difficult to distinguish from each other due to overlapping characteristics [22]. Our studies for 34 ecotypes of *S. polyrhiza* using *atpF-atpH* markers showed five ecotypes that have been clustered with *S. intermedia* (Additional file 4), which is mainly restricted to South America [22]. Good trace evidence comes from *S. polyrhiza* 9203 (Figure 3). Among five ecotypes, three are derived from South America, while another two are from India. Therefore, a refined classification is necessary to determine whether another four ecotypes except *S. polyrhiza* 9203 should be classified as *S. intermedia* rather than *S. polyrhiza*.

Both phylogenetic data [21] and our barcoding data showed that closely related species *W. gladiata* and *W. oblonga*, *L. minuta* and *L. valdiviana* could not be separated from each other (Figure 3). These sister-species share identical sequences for barcoding markers, which would require a search for additional barcoding markers with greater sequence polymorphism. In fact, a universal DNA barcoding marker has not been reported to distinguish more than 90% of species tested until now [8,32]. Elucidation of recently evolved species sharing identical barcoding sequences still needs further taxonomic or case-by-case morphological, flavonoid, and allozyme analyses. On the other hand, use of next-generation sequencing technologies and corresponding software applications are emerging where low pass coverage of different specimen could provide the necessary resolution.

Conclusions

In this study we have demonstrated that *atpF-atpH* noncoding spacer could serve as a universal DNA barcoding marker for species-level identification of duckweeds. This marker will allow to identify unknown species or to exploit new species of duckweeds by reason of its reliable amplification, straightforward sequence alignment, and rates of DNA variation between species and within species. DNA barcoding developed in this study are a significant contribution to the taxonomical structure in duckweeds compared with insensitive morphological classification.

Methods

Plant materials

The *Lemnaceae* collection originated from the Institut für Integrative Biologie (Zürich, Switzerland), the BIOLEX company (North Carolina, USA), and the University of Toronto Culture Collection of Algae and Cyanobacteria (UTCC, Toronto, Canada) where it was maintained for many years. Detailed information about many of these accessions is included in Dr. Landolt's monographic study [29]. In total, 97 ecotypes representing 31 species (81.6% of the known species) were sampled in this study. Since the intraspecific distance is very important for evaluating a suitable barcoding marker, 2 to 8 representatives per species are included for 19 species, whereas another 12 species are represented by a single ecotype. Moreover, the selected ecotypes represent a worldwide geographical distribution (Figure 1). A summary of all specimens included in this study was listed in Additional file 1.

DNA amplification, sequencing and alignment

All duckweed fronds were grown aseptically in half-strength Schenk and Hildebrandt medium (Sigma, S6765). Total DNA was extracted using CTAB [36]. The chloroplast markers *rpoB, rpoC1, rbcL, matK*, *atpF-atpH, trnH-psbA*, and *psbK-psbI*, which were proposed by the CBOL plant-working group, were amplified with a set of modified primers (Table 5) based on reference sequences from *Lemna minor* [33]. The amplicon sizes were also estimated according to *Lemna minor* (Table 5). PCR reaction conditions also followed guidelines from the CBOL plant-working group. Briefly, 50-100 ng genomic DNA and 5 pmol of each primer are added with the JumpStart™Redtop® ReadyMix™Reaction Mix (P1107, Sigma) Redix in 25 ml of final volume. To improve the universal application of these primers, they were designed to have an annealing temperature (T_a) of 50°C, which is 1 to 6°C lower than the optimal T_a of *Lemna minor* as determined by Beacon Designer software (PREMIER Biosoft International) under reaction conditions of 50 mM monovalent ion and 200 nM nucleic acid concentration (Table 5). The program uses the following formula: optimal T_a = 0.3 × Tm (primer) + 0.7 Tm (product) -14.9 [37]. The PCR products were purified with ExoSap-IT™(USB Corp.) and then sequenced on an ABI3730 automated sequencer using the same primers as in the PCR reactions. Both strands of each PCR product were sequenced and double-checked. The success ratios of PCR amplification and sequencing were counted (Table 1). After the ambiguous nucleotides (~30bp) at the ends of reads were removed, the length of products was measured and multiple DNA sequence alignments were generated using ClustalW in MEGA 4.1 [38].

Genetic distance analysis

Genetic distance was calculated using pairwise alignments of sequences between and within species (Table 2). The average intraspecific distance was calculated with the mean pairwise distance in each species with more than one representative, which eliminated biases due to unbalanced sampling among taxa. We evaluated conspecific and congeneric variability for each pair of marker sequences by Wilcoxon signed rank tests (Additional file 2 and 3) [9]. Median and Mann-Whitney U tests were executed to examine the extent of DNA barcoding gap/overlap between intra- and inter-specific divergences [8].

Evaluation of DNA barcoding markers based on sequence similarity

For assessing success in species assignment or identification among our data set, we adopted the "best match" function in the program TAXONDNA (Table 3) [31]. We calculated pairwise distances as uncorrected pairwise distances and compared two sequences over at least 300 bp except for *psbK-psbI* (230 bp). We suppressed indels when computing distances. The threshold was set at a value below which 95% of all intraspecific pairwise distances were found. Since the best match was based on direct sequence comparison with other conspecific ecotypes, the analysis only counted species with multiple ecotypes per species.

Evaluation of DNA barcoding markers using phylogenetic analysis

The other criterion used to measure success of species identification was based on generating a phylogenetic tree. We built trees with MEGA 4.1 by using the best algorithms methods of UPGMA and MP compared with other tree building techniques for DNA barcoding [8]. UPGMA trees were made from K2P distances. The MP trees were constructed using the close neighbor interchange (CNI) method with search level 1. The initial tree for the CNI search was created by random addition

Table 5 List of primers for the seven proposed DNA barcoding markers

Marker	Primer sequence	Amplicon size (*Lemna minor*)	Ta Optimum (*Lemna minor*)
psbK-psbI	Forward: 5'-TTAGCATTTGTTTGGCAAG-3';	544 bp	51°C
	Reverse: 5'- AAAGTTTGAGAGTAAGCAT -3'		
trnH-psbA	Forward: 5'-GTTATGCACGAACGTAATGCTC-3';	300 bp	55°C
	Reverse: 5'- CGCGCGTGGTGGATTCACAATCC-3'		
matK	Forward: 5'-CGTACTGTACTTTTATGTTTACGAG-3';	862 bp	55°C
	Reverse: 5'- ATCCGGTCCATCTAGAAATATTGGTTC -3'		
atpF-atpH	Forward: 5'-ACTCGCACACACTCCCTTTCC-3';	675 bp	53°C
	Reverse: 5'- GCTTTTATGGAAGCTTTAACAAT -3'		
rpoB	Forward: 5'-ATGCAGCGTCAAGCAGTTCC-3';	406 bp	55°C
	Reverse: 5'- TCGGATGTGAAAAGAAGTATA -3'		
rpoC1	Forward: 5'-GGAAAAGAGGGAAGATTCCG-3';	509 bp	56°C
	Reverse: 5'- CAATTAGCATATCTTGAGTTGG -3'		
rbcL	Forward: 5'-GTAAAATCAAGTCCACCACG-3';	580 bp	56°C
	Reverse: 5'-ATGTCACCACAAACAGAGACTAAAGC -3'		

for 10 replications. Each tree contains the bootstrap values as calculated by the software from 500 replicates. Here, we only calculated the number of successfully clustered species as monophyly among the species with multiple conspecific individuals (Figure 3, Additional file 4, Table 4).

Additional material

Additional file 1: Information of sampled duckweeds and GenBank accession numbers for sequence. A complete list of all species and ecotypes with relevant information including geographical position and marker sequences is provided.

Additional file 2: Wilcoxon signed rank tests of interspecific distance among markers. Values for each marker assessment is provided and ordered.

Additional file 3: Wilcoxon signed rank tests of intraspecific divergence among markers. Values for each marker assessment is provided and ordered.

Additional file 4: UPGMA tree based atpF-atpH sequences for sister species of S. polyrhiza and S. intermedia. Distance analysis was carried out as described under Methods.

Acknowledgements
We thank Elias Landolt from Institut für Integrative Biologie (Zürich, Switzerland), Lynn Dickey, Nirmala Rajbhandari, and Peaches Staton from BIOLEX (North Carolina, USA), and Judy Acreman (UTCC, Toronto, Canada) for their generous provision of duckweed ecotypes. The research described in this manuscript was supported by the Selman A. Waksman Chair in Molecular Genetics.

Authors' contributions
WW designed experiment, analyzed data and wrote the manuscript. YW conducted PCR and sequenced the products. YY and ME kept the duckweeds collection and extracted genomic DNA. RK provided advices for experiments and revised manuscript. JM supervised the work, interpreted data with WW, and revised all versions of the manuscript. All authors read and approved the final manuscript.

Received: 7 June 2010 Accepted: 16 September 2010
Published: 16 September 2010

References

1. Meyer CP, Paulay G: **DNA barcoding: error rates based on comprehensive sampling.** *PLoS Biol* 2005, **3(12)**:e422.
2. Hebert PDN, Cywinska A, Ball SL, deWaard JR: **Biological identifications through DNA barcodes.** *Proceedings of the Royal Society of London Series B: Biological Sciences* 2003, **270(1512)**:313-321.
3. Hebert PDN, Stoeckle MY, Zemlak TS, Francis CM: **Identification of birds through DNA barcodes.** *PLoS Biol* 2004, **2(10)**:e312.
4. Hubert N, Hanner R, Holm E, Mandrak NE, Taylor E, Burridge M, Watkinson D, Dumont P, Curry A, Bentzen P, Zhang J, April J, Bernatchez L: **Identifying Canadian freshwater fishes through DNA barcodes.** *PLoS ONE* 2008, **3(6)**:e2490.
5. Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH: **Use of DNA barcodes to identify flowering plants.** *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102(23)**:8369-8374.
6. Chase MW, Salamin N, Wilkinson M, Dunwell JM, Kesanakurthi RP, Haidar N, Savolainen V: **Land plants and DNA barcodes: short-term and long-term goals.** *Philos Trans R Soc Lond B Biol Sci* 2005, **360(1462)**:1889-1895.
7. Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, Chase MW, Cowan RS, Erickson DL, Fazekas AJ, Graham SW, James KE, Kim K-J, Kress WJ, Schneider H, van AlphenStahl J, Barrett SCH, van den Berg C, Bogarin D, Burgess KS, Cameron KM, Carine M, Chacón J, Clark A, Clarkson JJ, Conrad F, Devey DS, Ford CS, Hedderson TAJ, Hollingsworth ML, *et al*: **A DNA barcode for land plants.** *Proceedings of the National Academy of Sciences* 2009, **106(31)**:12794-12797.
8. Lahaye R, Savolainen , Vincent , Duthoit , Sylvie , Maurin , Olivier , van der Bank, Michelle : **A test of psbK-psbI and atpF-atpH as potential plant DNA barcodes using the flora of the Kruger National Park (South Africa) as a model system.** *Nature Precedings* 2008 [http://hdl.handle.net/10101/npre.2008.1896.1].
9. Kress WJ, Erickson DL: **A two-locus global DNA barcode for land plants: The coding rbcL gene complements the non-coding trnH-psbA spacer region.** *PLoS ONE* 2007, **2(6)**:e508.
10. Chase MW, Cowan RS, Hollingsworth PM, van den Berg C, n S, Petersen G, Seberg O, rgsensen T, Cameron KM, Carine M, Pedersen N, Hedderson TAJ, Conrad F, Salazar GA, Richardson JE, Hollingsworth ML, Barraclough TG, Kelly L, Wilkinson M: **A proposal for a standardised protocol to barcode all land plants.** *Taxon* 2007, **56**:295-299.
11. Pennisi E: **TAXONOMY: Wanted: A barcode for plants.** *Science* 2007, **318(5848)**:190-191.
12. Chase MW, Fay MF: **Barcoding of plants and fungi.** *Science* 2009, **325(5941)**:682-683.
13. Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG, Husband BC, Percy DM, Hajibabaei M, Barrett SCH: **Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well.** *PLoS ONE* 2008, **3(7)**:e2802.
14. Kress WJ, Erickson DL, Jones FA, Swenson NG, Perez R, Sanjur O, Bermingham E: **Plant DNA barcodes and a community phylogeny of a tropical forest dynamics plot in Panama.** *Proc Natl Acad Sci USA* 2009, **106(44)**:18621-18626.
15. Newmaster SG, Fazekas AJ, Steeves RAD, Janovec J: **Testing candidate plant barcode regions in the Myristicaceae.** *Molecular Ecology Resources* 2008, **8(3)**:480-490.
16. Steven GN, Subramanyam R: **Testing plant barcoding in a sister species complex of pantropical Acacia (Mimosoideae, Fabaceae).** *Molecular Ecology Resources* 2009, **9**:172-180.
17. Asahina H, Shinozaki J, Masuda K, Morimitsu Y, Satake M: **Identification of medicinal Dendrobium species by phylogenetic analyses using matK and rbcL sequences.** *Journal of Natural Medicines* 2010, **64(2)**:133-138.
18. Ward J, Gilmore SR, Robertson J, Peakall R: **A grass molecular identification system for forensic botany: A critical evaluation of the strengths and limitations*.** *Journal of Forensic Sciences* 2009, **54(6)**:1254-1260.
19. Starr JR, Naczi RFC, Chouinard BN: **Plant DNA barcodes and species resolution in sedges (Carex, Cyperaceae).** *Molecular Ecology Resources* 2009, **9**:151-163.
20. Clerc-Blain JL, Starr JR, Bull RD, Saarela JM: **A regional approach to plant DNA barcoding provides high species resolution of sedges (Carex and Kobresia, Cyperaceae) in the Canadian Arctic Archipelago.** *Molecular Ecology Resources* 2010, **10(1)**:69-91.
21. Les DH, Crawford DJ, Landolt E, Gabel JD, Kimball RT: **Phylogeny and systematics of Lemnaceae, the duckweed family.** *Systematic Botany* 2002, **27(2)**:221-240.
22. Landolt E: **The family of Lemnaceae - a monographic study Vol 1.** Veroffentlichungen des Geobotanischen Institutes der Eidgenossischen Technischen Hochschule, Stiftung Rubel 1986, **1.**
23. Stomp AM: **The duckweeds: a valuable plant for biomanufacturing.** *Biotechnol Annu Rev* 2005, **11**:69-99.
24. Brain RA, Solomon KR: **A protocol for conducting 7-day daily renewal tests with Lemna gibba.** *Nat Protoc* 2007, **2(4)**:979-987.
25. Ozengin N, Elmaci A: **Performance of duckweed (Lemna minor L.) on different types of wastewater treatment.** *J Environ Biol* 2007, **28(2)**:307-314.
26. Cheryl CS, Anthony JT: **Abscisic-acid-induced turion formation in Spirodela polyrrhiza L. I. Production and development of the turion.** *Plant, Cell and Environment* 1983, **6(6)**:507-514.
27. Appenroth KJ: **Co-action of temperature and phosphate in inducing turion formation in Spirodela polyrhiza (Great duckweed).** *Plant, Cell & Environment* 2002, **25(9)**:1079-1085.
28. Cheng JJ, Stomp AM: **Growing duckweed to recover nutrients from wastewaters and for production of fuel ethanol and animal feed.** *CLEAN - Soil, Air, Water* 2009, **37(1)**:17-26.

29. Landolt E: **Biosystematic investigations in the family of duckweeds (Lemnaceae).** Veroffentlichungen des Geobotanischen Institutes der Eidgenossischen Technischen Hochschule, Stiftung Rubel 1980, **1.**
30. Hebert PDN, Gregory TR: **The promise of DNA barcoding for taxonomy.** *Syst Biol* 2005, **54(5)**:852-859.
31. Meier R, Shiyang K, Vaidya G, Ng PKL: **DNA barcoding and taxonomy in Diptera: A tale of high intraspecific variability and low identification success.** *Syst Biol* 2006, **55(5)**:715-728.
32. Gonzalez MA, Baraloto C, Engel J, Mori SA, Petronelli P, Riera B, Roger A, Thebaud C, Chave J: **Identification of Amazonian trees with DNA barcodes.** *PLoS One* 2009, **4(10)**:e7483.
33. Mardanov A, Ravin N, Kuznetsov B, Samigullin T, Antonov A, Kolganova T, Skyabin K: **Complete sequence of the duckweed (Lemna minor) chloroplast genome: structural organization and phylogenetic relationships to other angiosperms.** *Journal of Molecular Evolution* 2008, **66(6)**:555-564.
34. Shaw J, Lickey EB, Beck JT, Farmer SB, Liu W, Miller J, Siripun KC, Winder CT, Schilling EE, Small RL: **The tortoise and the hare II: relative utility of 21 noncoding chloroplast DNA sequences for phylogenetic analysis.** *Am J Bot* 2005, **92(1)**:142-166.
35. Hilu KW, Borsch T, Muller K, Soltis DE, Soltis PS, Savolainen V, Chase MW, Powell MP, Alice LA, Evans R, Sauquet H, Neinhuis C, Slotta TAB, Rohwer JG, Campbell CS, Chatrou LW: **Angiosperm phylogeny based on matK sequence information.** *Am J Bot* 2003, **90(12)**:1758-1776.
36. Murray MG, Thompson WF: **Rapid isolation of high molecular weight plant DNA.** *Nucl Acids Res* 1980, **8(19)**:4321-4326.
37. Rychlik W, Spencer WJ, Rhoads RE: **Optimization of the annealing temperature for DNA amplification in vitro.** *Nucleic Acids Res* 1990, **18(21)**:6409-6412.
38. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: xMolecular evolutionary genetics analysis (MEGA) software version 4.0.** *Mol Biol Evol* 2007, **24(8)**:1596-1599.

doi:10.1186/1471-2229-10-205
Cite this article as: Wang *et al.*: DNA barcoding of the *Lemnaceae*, a family of aquatic monocots. *BMC Plant Biology* 2010 **10**:205.

High-Throughput Sequencing of Three *Lemnoideae* (Duckweeds) Chloroplast Genomes from Total DNA

Wenqin Wang, Joachim Messing*

Waksman Institute of Microbiology, Rutgers, The State University of New Jersey, Piscataway, New Jersey, United States of America

Abstract

Background: Chloroplast genomes provide a wealth of information for evolutionary and population genetic studies. Chloroplasts play a particularly important role in the adaption for aquatic plants because they float on water and their major surface is exposed continuously to sunlight. The subfamily of *Lemnoideae* represents such a collection of aquatic species that because of photosynthesis represents one of the fastest growing plant species on earth.

Methods: We sequenced the chloroplast genomes from three different genera of *Lemnoideae*, *Spirodela polyrhiza*, *Wolffiella lingulata* and *Wolffia australiana* by high-throughput DNA sequencing of genomic DNA using the SOLiD platform. Unfractionated total DNA contains high copies of plastid DNA so that sequences from the nucleus and mitochondria can easily be filtered computationally. Remaining sequence reads were assembled into contiguous sequences (contigs) using SOLiD software tools. Contigs were mapped to a reference genome of *Lemna minor* and gaps, selected by PCR, were sequenced on the ABI3730xl platform.

Conclusions: This combinatorial approach yielded whole genomic contiguous sequences in a cost-effective manner. Over 1,000-time coverage of chloroplast from total DNA were reached by the SOLiD platform in a single spot on a quadrant slide without purification. Comparative analysis indicated that the chloroplast genome was conserved in gene number and organization with respect to the reference genome of *L. minor*. However, higher nucleotide substitution, abundant deletions and insertions occurred in non-coding regions of these genomes, indicating a greater genomic dynamics than expected from the comparison of other related species in the *Pooideae*. Noticeably, there was no transition bias over transversion in *Lemnoideae*. The data should have immediate applications in evolutionary biology and plant taxonomy with increased resolution and statistical power.

Citation: Wang W, Messing J (2011) High-Throughput Sequencing of Three *Lemnoideae* (Duckweeds) Chloroplast Genomes from Total DNA. PLoS ONE 6(9): e24670. doi:10.1371/journal.pone.0024670

Editor: Jonathan H. Badger, J. Craig Venter Institute, United States of America

Received July 5, 2011; **Accepted** August 15, 2011; **Published** September 9, 2011

Funding: No current external funding sources were received for this study.

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

* E-mail: messing@waksman.rutgers.edu

Introduction

Plants are defined by primary plastids, encompassing algae, Streptophytes, and land plants [1]. Each plant cell has three genomes, separated in three subcellular compartments, the nucleus, the chloroplasts, and the mitochondria. Chloroplasts are key organelles of green plants for photosynthesis. They are also responsible for storage of starch, and synthesis of chlorophyll, nucleic acids, and 50% of soluble protein in leaves. Chloroplasts are highly conserved in terms of their structure, genome size (from 120 to 217 Kb) and its gene content (~130 genes) [2]. Typically chloroplast genomes in plants contain two identical inverted repeats (IRa and IRb), separated by unique sequences, the large single copy (LSC) and the small single copy (SSC) regions [3]. Chloroplasts contain multiple copies of a circular, double-stranded DNA molecule. For instance, leaf cells of tobacco and pea typically have ~100 chloroplasts and up to 10,000 DNA copies [4]. Total genomic DNA could have as much as 5,000 times the copies of chloroplast DNA relative to nuclear gene copies as tested in monocots and dicots [5]. In addition to its important biological roles, chloroplast genome sequences are widely used in evolutionary studies, comparative genomics [6], and biotechnology [7].

Lemnoideae (duckweeds) are a subfamily of the *Araceae* of aquatic flowering monocot plants [8]. However, their minute size and simple morphologically characteristics made them extremely difficult for systematic analysis and species identification. Integration of morphological, flavonoid, allozyme, and DNA markers have yielded a single and well-resolved maximum parsimonious tree, but the resolution for closely related species is problematic with very low value of bootstrap support [9]. The same is true for DNA barcoding of the *Lemnoideae* subfamily. Actually, the *atpF-atpH* marker appeared to be the most powerful barcode to distinguish individual species of *Lemnoideae* with 14 out of 19 species, still short of complete coverage [10]. Indeed, a prevalent feature of chloroplast genomes is their high degree of sequence conservation. Choices of greater numbers of divergent sequences should increase resolution both for the exploration of plant relationships and DNA plant barcoding. Because the chloroplast genome in contrast to the nuclear genome is haploid and is uniparentally inherited, acting as a single locus, it has the potential to become the elusive universal single-locus for plant species identification and systematic analysis.

Duckweeds also have great potential industrial applications. Their biomass doubles every 1 or 2 days. They contain a starch

content of 45.8% (dry weight) growing in wastewater [11]. They can keep accumulating starch as high as 65% when switching from frond to the turion phase [12]. Therefore, duckweeds have been proposed as an alternative starch source for fuel production. Taking into account the recent improvements in transplastomic techniques, which provides an environmentally benign method of plant genetic engineering and accumulates extraordinarily high levels of foreign proteins [7], duckweed chloroplast transformation would greatly accelerate the exploration of its biofuel potential.

Traditionally, chloroplast genomes have been sequenced by primer walking based on closely related known genomes [13] or by shotgun sequencing [6]. However, with the advent of next generation sequencing platforms a new cost-effective option to capture multiple genomes on a larger scale has arisen [14]. Still, the separation of plastid DNA from nuclei and mitochondria can be tedious and would require the use of multiple long PCR reactions to obtain overlapping fragments (5 to 10 Kb) of the entire chloroplast genome, which could produce long gaps if some PCR reactions would fail [14,15]. Another way is to use a modified chloroplast isolation protocol and further amply them by multiple-primed rolling circle methods [16]. Either way, it would need substantial efforts to obtain enriched chloroplast DNA that could contain significant amounts of contaminating non-target DNA.

A recent study reported that chloroplast genome sequences were recovered from total DNA including nuclei, chloroplasts, and mitochondria by using an Illumina-based sequencing platform. Still, many gaps could not be bridged because of highly divergent regions [17]. However, here we could demonstrate that it is possible to assemble complete chloroplast genome sequences from total leaf DNA with the SOLiD sequencing platform at a high level of accuracy, following the same principles that have been applied to the first genome assembled entirely by shotgun DNA sequencing [18]. To obtain regions from the chloroplast genome that diverged from a reference genome, *de novo* assembly was employed using paired reads based on the concept of universal synthetic primers [19]. Before assembly, SOLiD reads from mitochondrial and nuclear DNA, were filtered electronically. Furthermore, we could use the chloroplast genome of the closely related species *L. minor* as a reference that has been sequenced with traditional overlapping long reads [13]. Genome assembly, the comparative and phylogenetic analyses of these genomes are presented here.

Methods

DNA isolation and SOLiD DNA sequencing

Duckweeds sequenced in this study (Table 1) were grown from a cluster of 3–5 fronds produced by a single mother frond. Total DNA was extracted from whole plant tissue by the CTAB method [20]. Sequencing runs were done on a SOLiD^{TM} 3 Analyzer (Applied Biosystems, Foster City, CA) at the Waksman Genomics Core Facility of Rutgers University. Mate-paired libraries with approximately 1.5 Kb inserts were constructed from 20 μg of genomic DNA following the manufacturer's instructions (SOLiD sample preparation protocol for Mate-Paired library sequencing), and deposited in one spot of a quadrant slide. Fifty nucleotide-long reads were obtained from each of the F3 and R3 tags, with more than 100 million reads obtained for each of the genomes.

Sequence data analysis pipeline

To assemble the chloroplast genomes using SOLiD reads and close the remaining gaps with long reads from capillary electrophoresis (CE) sequencers, we used the following steps (Fig. 1). Because all chloroplast genomes contain two identical inverted repeats (IRs), we first assembled genomes without IRb's and with LSC, SSC, IRa, but added them later on for the full-length molecules.

1) Data filtering: SOLiD mate-paired short reads were preprocessed by Mean Filter of a Perl script [21]; i.e., reads were truncated to 40 bp and average quality of reads were set to exceed the threshold QV score of 20. Because coverage is very high, only successful mate-pair reads went into the next step. 2) Selection of chloroplast-related reads: The filtered mate-pair colorspace reads from each of the three samples were aligned to the chloroplast genome of *L. minor* [13] (GenBank accession number: DQ400350) using the BWA short-read alignment component with default parameters [22]. At least one end of the paired-end reads was anchored to the chloroplast genome of *L. minor* before interrogating the second end to map to a linked sequence or to a gap. 3) 1st run of genome assembly: *De novo* assembly was performed with identified chloroplast-related reads using the SOLiD^{TM} System *de novo* Accessory Tools 2.0 (http://solidsoftwaretools.com/gf/project/denovo/) in conjunction with the Velvet assembly engine [23]. These tools are designed to simplify and optimize parameters for ease of use and best performance. They sample an optimal subset of reads and automatically estimate optimal parameters for each step. Velvet parameters generated from the tools were deposited in Table 2 with hash length 19 and coverage cut-off 11. The assembly assistant module in the tool kit took the input from Velvet and produced scaffolds with 120 mate-pair confirmations to make confident scaffolding at the conclusion of this pipeline. 4) 2nd run of genome assembly: After the first run, all scaffolds were concatenated into pseudomolecules. In order to maximize chloroplast-related reads, the artificial molecule functioned as a new reference and step 2 and 3 were then reiterated. 5) Correction

Table 1. Species used for comparative genomic analysis.

Species	Source	Nuclear Genome Size[b] (Mbp)	Chloroplast Genome Size (bp)	Inverted Repeats Size (bp)	Genbank Number
Spirodela polyrhiza 7498	North Carolina, Durham Co., Durham, 'USA	160	168788	31755	JN160603
Lemna minor (reference) [a]	Russia	356-604	165955	31223	DQ400350
Wolffiella lingulata 7289	Amazonas, Manaus, Rio Negro, 'Brazil	655	169337	31683	JN160604
Wolffia australiana 7733	Mount Lofty Range, Torrens Gorge, 'South Australia	357	168704	31930	JN160605

[a]Reference chloroplast genome [13];
[b]Nuclear genome sizes [29].
doi:10.1371/journal.pone.0024670.t001

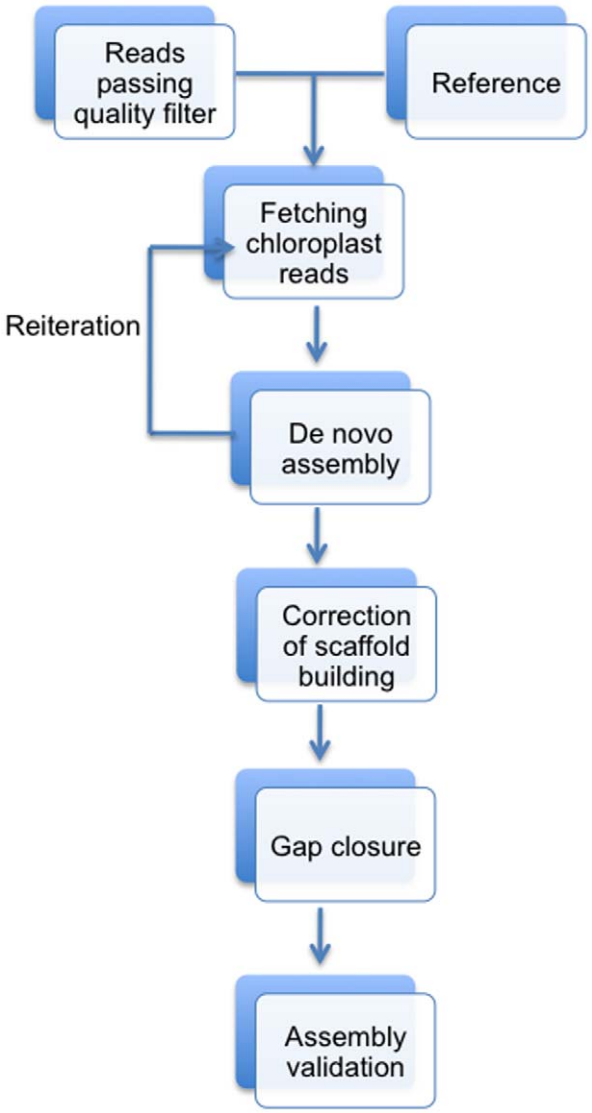


Figure 1. Pipeline of chloroplast genome assembly. Details are described under Methods.
doi:10.1371/journal.pone.0024670.g001

of scaffold building: The biggest scaffolds of each genome were aligned with the most closely related reference genome of *L. minor* using BLAST2 (http://blast.ncbi.nlm.nih.gov/). Indeed in a few instances, non-contiguous genomic regions were found in juxtaposed positions at gap positions. At these gaps scaffolds were broken and contigs reordered in collinearity with the reference genome. Smaller contigs were manually ordered based on the reference genome. All scaffolds were then concatenated into a single full-length molecule, where each gap in the sequence was marked with one N. 6) Gap closure: Gaps were small enough so that flanking primer pairs could be chosen (Table S1) to isolate missing sequences by PCR and apply CE sequencing methods (ABI 3730XL) for closure. PCR amplification and conditions have been described recently [10]; 7) Assembly validation: Because PCR amplification of gaps required correct ordering of contigs into scaffolds, the long CE reads provided validation of overlapping sequences and the correct ordering of short read assemblies. Accumulative overlaps and discrepancies between alignments of sequences from both methods were summarized using DNASTAR software (http://www.dnastar.com/), which would reveal sequencing errors of the SOLiD platform. Because of mate-pair data junctions between IRb and LSU or SSU could be confirmed with CE sequencing of PCR products. 8) GenBank deposition: The fully sequenced genomes of the three species were annotated by DOGMA [24], checked manually, and have been deposited into GenBank as a whole genome shotgun project (Table 1).

To assess the contribution of the filtering step with the reference genome to the performance of Velvet as an assembly tool, we also performed an assembly with total DNA reads including the nuclear and mitochondria DNA. Under these conditions, we could not use the default set-up parameters for the pipeline, which requires uniform coverage by a single genome. Otherwise, the precomputed parameters would extract sub-set reads that represent a mixture of three genomes with different coverage. To avoid this, we determined the optimized parameters after omitting data filtering as step 1 by empirically testing parameters for step 2, 3, and 4 and then manually accessing the SOLiD^{TM} System *de novo* Accessory Tools 2.0 as shown in Table 2. All other assembly steps were the same as described with selected reads.

Whole genome alignments, comparison, and phylogenetic analysis

Lemnoideae chloroplasts, *S. polyrhiza* 7498 (S.pol), *L. minor* (L.min), *W. lingulata* 7289 (W.lin), *W. australiana* 7733 (W.aus) were aligned by a program of global multiple alignments of finished sequences (Multi-LAGAN) [25] and annotation for the reference genome of *L. minor* [13] was used to construct sequence conservation plots in the program mVISTA [26].

The 81 protein coding nucleotide sequences from duckweeds were retrieved after annotation by DOGMA, concatenated as one full-length molecule and pair-wisely aligned with each other by Multi-LAGAN. MEGA 5 was used to detect transitions, transversions, and INDELs (insertion/deletion) for all genomes except the IRb regions and protein coding sequences. A similar analysis of 71 common genes was done for chloroplast genomes of species in the subfamily of the *Pooideae*, i.e., wheat (AB042240), barley (EF115541) and *Brachypodium* (EU325680). They were chosen because wheat and barley belong to the same tribe of *Triticeae*, whereas *Brachypodium* belongs to the different tribe of *Brachypodieae* within the same subfamily. This is taxonomically equivalent to the division within the subfamily of the *Lemnoideae*. The *Spirodela* and *Lemna* species belong to the same tribe, but *Wolffiella* and *Wolffia* to a different one [8,9].

To examine whether the genome-wide phylogenetic analyses were consistent with those of morphological, flavonoid, and allozyme markers, as well as selected DNA sequences [9], we employed Maximum Parsimony to reconstruct the *Lemnoideae* phylogeny with whole chloroplast sequences by using MEGA 5 [27]. *Phoenix dactylifera* is in the same class of *Liliopsida* as *Lemnoideae* and functions as an outgroup here [28]. However, one of the two inverted repeat regions (IRb) was excluded from phylogenetic analyses.

Results

De novo assembly of short sequence reads yields high quality contigs

The chloroplast genomes of *S. polyrhiza*, *W. lingulata* and *W. australiana* in this study were selected on the basis of phylogenetic diversity of the subfamily *Lemnoideae* and their extensive variation of nuclear genome sizes (Table 1) [29]. The three genomes were sequenced using mate-paired libraries with the SOLiDTM 3

Table 2. *De novo* assembly statistics for the three sequenced species.

Species	Read processing	Sc	N50 Sc (bp)	Co	N50 Co (bp)	Co length (bp)	Hash length	Expected coverage	Coverage cut-off[a]	min_pair_ count[b]	Total reads (X10^6)	Aligned reads (%)	Average genome coverage	Average nuclear coverage
Spirodela polyrhiza 7498	with selection	3	92558	60	4246	136597	19	150	11	120	153	12.9	5474	42
	without selection	6	36267	73	4062	132134	19	150	30	120				
Wolffiella lingulata 7289	with selection	1	136457	53	4708	139523	19	150	11	120	155	2.5	1070	12
	without selection	8	25221	82	2333	133615	19	150	6	120				
Wolffia australiana 7733	with selection	2	134892	39	8677	137183	19	150	11	120	111	6.2	1912	15
	without selection	3	98687	60	3743	132446	19	150	11	120				

[a]Coverage cut-off: minimum coverage required to form a contig.
[b]min_pair_count: number of mate-pair confirmations required for confident scaffolding. Sc = scaffold; Co = contig.
doi:10.1371/journal.pone.0024670.t002

System. The previously sequenced *L. minor* chloroplast genome was used as computational filter to separate chloroplast reads from nuclear and mitochondria reads. Considering the identical feature of the two inverted repeats, we first assembled 136 Kb of the chloroplast genome from the LSC, IRa, SSC regions. All three genomes were each processed into one single large scaffold of 92 Kb (S.pol), 136Kb (W.lin), and 134 Kb (W.aus), respectively. Assembly of SOLiD reads resulted between 39 to 60 contigs and 1 to 3 scaffolds per genome (Table 2). With the second largest scaffold of 40 Kb for S.pol, the length of all the added contigs already reached a size expected for a chloroplast genome excluding the IRb region. However, alignment of these assemblies with the reference genome suggested between one to three misassembled scaffolds that needed to be corrected. Most contigs were interrupted by mononucleotide repeats and low complexity sequences.

Clearly, read length is a critical factor for assembly programs, but how critical is the filtering step for separating the mixture of nuclear, mitochondria, and chloroplast genomic sequences for the assembly tool used here. We therefore modified the parameters and the steps in the data processing protocol empirically to produce sequence assemblies without prior selection of chloroplast sequences. *De novo* assembly from total reads generated 60–82 contigs with 2333–4062 bp of N50 contig length, whereas assembly from selected chloroplast reads gave us a significant improvement with 18% to 35% lower contig numbers and longer N50 values of contig lengths (Table 2). If the computational read selection were omitted, 13–29 additional PCR reactions would have been required to close the gaps from total reads assembly and validate order of contigs and scaffolds as described below.

Using the ends of contigs separated by Ns, primers were designed for PCR amplification. Because of the alignment with the reference genome, the correct ordering of contigs could be confirmed by the fact that PCR amplification occurred. Furthermore, when PCR products were sequenced by the CE ABI 3730XL platform, overlapping sequences could be used to close gaps and validate the order of contigs. Accumulative overlaps for the three genomes totaled 48 Kb. When short read assemblies were compared with CE long read sequences, the cumulative differences amounted to just 0.041%, reflecting a high consensus between the two sequencing methods. We also could test the short read assembler by mapping *de novo* assemblies back to the complete genome. Although only 2.5~12.9% of the reads were successfully aligned, keeping in mind the DNA mixture from plant tissue, this was sufficient to give a mean coverage between 1,070 to 5,474 times (Table 2 and Fig. 2). The IRa and IRb regions had lower coverage due to random placement of repetitive read pairs when mapping. For nuclear genome sequences, we found 12 to 42-fold coverage by ignoring mitochondrial DNA reads (Table 2). Based on these assessments, there were approximately 100 chloroplast genome copies for every nuclear genome copy.

Sequence comparison and phylogeny among *Lemnoideae* chloroplast genomes

The chloroplast genomes of duckweeds appeared to be within a short range of 165,955 bp to 169,353 bp in length (Table 1). All of them include a pair of inverted repeats of around 31 Kb separated by SSC and LSC. Large single copy (LSC) and Small Single Copy (SSC) regions were close to 90 Kb and 10 Kb long, respectively. *S. polyrhiza*, *W. lingulata* and *W. australiana* contain the same gene number and order as the reference genome *L. minor* (Fig. 3).

The conservation of the overall structure of the chloroplast genomes allowed us to align the sequences of four duckweed species at the genome-wide level. Comparison of the sequences

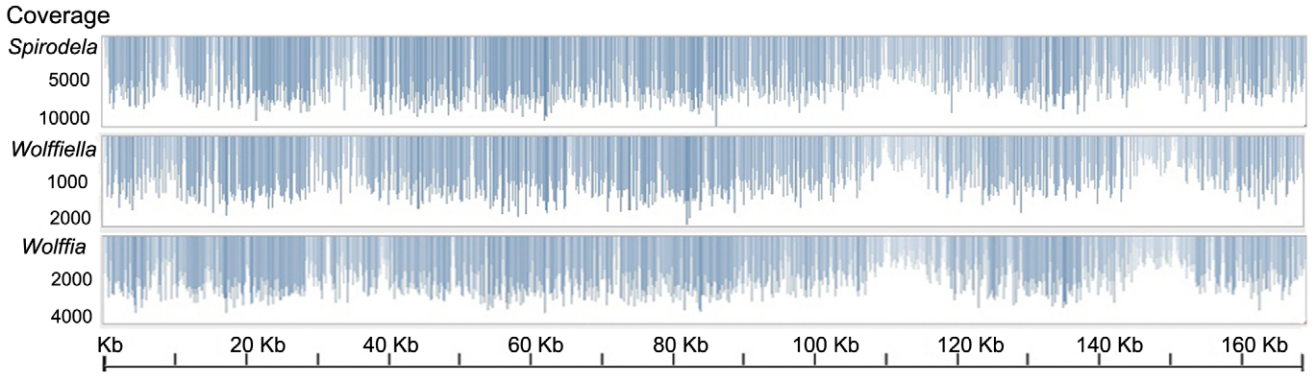


Figure 2. Coverage of *Lemnoideae* chloroplast genome by SOLiD system reads. Depth of coverage was plotted along the genome coordinates. Blue peaks show the coverage.
doi:10.1371/journal.pone.0024670.g002

revealed multiple hotspots of high sequence length polymorphism (Fig. 3). The IRs showed lower sequence divergence than the single-copy regions. The majority of highly divergent regions were in non-coding regions as illustrated in an mVISTA alignment plot. The region between *rpoB* and *psbD* from position 28 Kb to 36 Kb is one of the most polymorphic regions. For example, *W. australiana* has a 425-bp deletion in the 29 Kb *rpoB*-tRNA-Cys region. *S. polyrhiza* has a 505-bp deletion compared with 100-bp deletions in *W. lingulata*, whereas a 353-bp insertion occurred at 31 Kb of the intergenic *petN-psbM* region of *W. australiana*. Both *W. lingulata* and *W. australiana* have a 460-bp deletion in the 32 Kb *psbM*-tRNA-Asp region. Moreover, some INDELs existed in introns, such as a 123-bp insertion in *atpF* of *Spirodela* at 13 Kb, and 114-bp deletion in *ndhA* for *W. lingulata* and 105-bp for *W. australiana* at the 132 Kb region (Fig. 3).

Maximum parsimony produced a single fully resolved tree with strong node support (Fig. 4). Our phylogenetic results showed *Wolffiella* and *Wolffia* were more closely related than the others. Furthermore, our analysis strongly supported that *Spirodela* was at the basal position of the taxon, followed by *Lemna* and *Wolffiella*, whereas *Wolffia* was the most derived (Fig. 4).

Evolution of *Lemnoideae* and *Pooideae*, with chloroplast genomes in different orders

To further evaluate the pace of evolutionarily divergence, we compared chloroplast genomes from different monocot orders by quantifying nucleotide substitution rates and INDELs ratios. The subfamily of *Pooideae* within the *Poaceae* belongs to the order of the *Poales*, whereas the *Lemnoideae* belong to the order of the *Alismatales*. When such a comparison is made, duckweeds have a higher rate of substitution than species of the *Pooideae* at the whole genome level and in protein-coding regions. Moreover, INDELs were very prominent in duckweed genomes with ratios of 0.061 to 0.095, whereas they were much higher than the values between 0.006 and 0.012 in conservative coding regions. When we compared duckweeds with species of the *Pooideae*, duckweeds had twice as many INDELs in their chloroplast genomes than the *Pooideae's* species based on the same level of intra-tribe or inter-tribe comparisons (Table 3). Based on INDELs length in genome and coding regions (Table 3), we could conclude that most INDELs were located in non-coding regions. Interestingly, we found that transversions were higher than transitions in the subfamily of *Lemnoideae* with R-values from 0.6 to 0.7 of the total genome. The same result was discovered in protein coding regions except between *S. polyrhiza* and *L. minor* (R = 1.1). However, these values were completely the opposite in the species of the subfamily of *Pooideae* with R-values from 1.2 to 1.7, where transitions were more numerous than transversions (Table 3).

Discussion

Next generation sequencing platforms have mainly been used for re-sequencing, SNP analysis, and expression profiling because it has been difficult to develop *de novo* assembly tools for short sequence reads [30]. Whereas re-sequencing or sequencing of related genomes can be very productive for SNP detection and for map-based cloning of mutant alleles, short-read assemblies often fail to detect large INDELs and variable regions in new genomes because technically there is no reference for them. *De novo* assemblies of short reads could cover all insertions, deletions, and rearrangements that would otherwise be incorrectly assembled based on alignments with a reference genome [14]. The pipeline of the SOLiD™ System *de novo* Accessory Tools 2.0, however, has been well adapted to assemble high-coverage SOLiD reads of microbial genomes [31]. Because chloroplasts are even smaller than bacterial genomes, more in the order of large viruses, they represent an exception where such method can be applied. Moreover, we could use paired reads from the same DNA fragment to anchor one end to a contig and the other to a gap that could overlap with other unanchored ends. For this purpose, we used a module Assembly Assistant for SOLiDTM to maximally fill gaps in scaffolds by sufficiently utilizing benefits of these paired ends (http://solidsoftwaretools.com/gf/project/denovo/). Indeed, we got good assemblies by using high quality reads and minimizing non-target DNA from read mixtures. However, interference for contig building arose mainly from long mononucleotide repeats and low complexity sequence. Final mapping of SOLiD reads back to the complete chloroplast genome yielded only 2.5~12.9% alignment due to 1,000 times smaller genome size than nuclear genome. After comparison of the assembly from computationally selected chloroplast reads with that from total reads, we could show that there is a significant advantage of masking non-chloroplast reads if a related genome sequence is available. Furthermore, without masking, the minimum coverage required to form a contig (coverage cut-off) for Velvet needs to be empirically determined to favor the higher coverage of chloroplast reads over the much lower coverage of nuclei and mitochondria genome sequences to enter the assembly program. Exploration of different computational filters, however, could be used to mask chloroplast sequences instead to favor the assembly of either nuclear or mitochondrial genomic DNA in parallel from the same

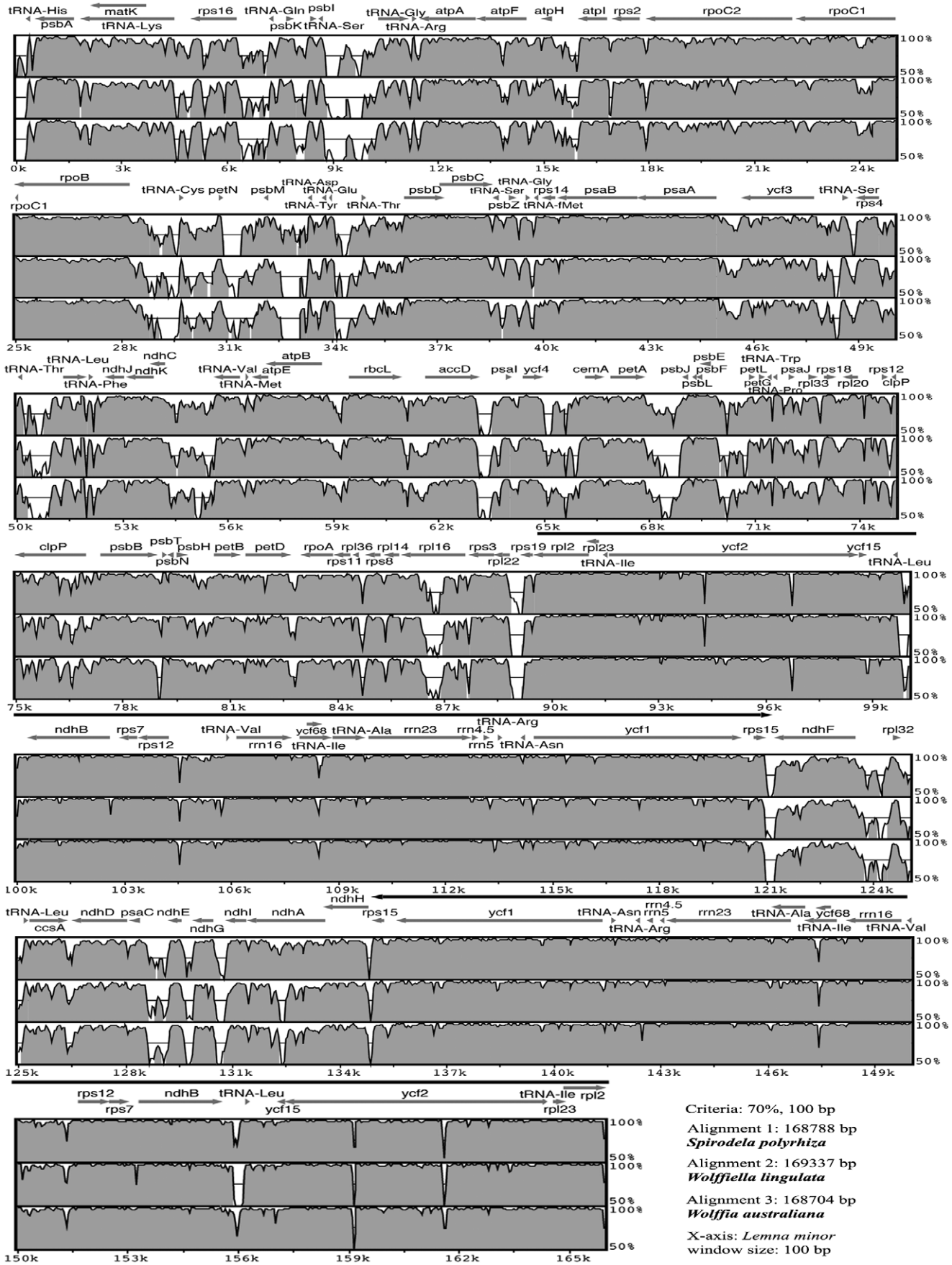

tRNA-His matK rps16 tRNA-Gln psbI tRNA-Gly atpA atpF atpH atpI rps2 rpoC2 rpoC1
psbA tRNA-Lys psbK tRNA-Ser tRNA-Arg
0k 3k 6k 9k 12k 15k 18k 21k 24k
rpoB tRNA-Cys petN psbM tRNA-Asp tRNA-Glu psbC tRNA-Gly psbD tRNA-Ser rps14 psaB psaA ycf3 tRNA-Ser
rpoC1 tRNA-Tyr tRNA-Thr psbZ tRNA-fMet rps4
25k 28k 31k 34k 37k 40k 43k 46k 49k
tRNA-Leu ndhC atpB psbE tRNA-Trp
tRNA-Thr ndhJ ndhK tRNA-Val atpE rbcL accD psaI ycf4 cemA petA psbJ psbF petL psaJ rps18 rps12
tRNA-Phe tRNA-Met psbL petG rpl33 rpl20 clpP
tRNA-Pro
50k 53k 56k 59k 62k 65k 68k 71k 74k
clpP psbB psbT psbH petB petD rpoA rpl36 rpl14 rpl16 rps3 rps19 rpl2 rpl23 ycf2 ycf15
psbN rps11 rps8 rpl22 tRNA-Ile tRNA-Leu
75k 78k 81k 84k 87k 90k 93k 96k 99k
ndhB rps7 tRNA-Val ycf68 tRNA-Ala rrn23 rrn4.5 tRNA-Arg ycf1 rps15 ndhF rpl32
rps12 rrn16 tRNA-Ile rrn5 tRNA-Asn
100k 103k 106k 109k 112k 115k 118k 121k 124k
tRNA-Leu ndhD psaC ndhE ndhI ndhA ndhH rps15 ycf1 tRNA-Asn rrn4.5 rrn5 rrn23 tRNA-Ala ycf68 rrn16
ccsA ndhG tRNA-Arg tRNA-Ile tRNA-Val
125k 128k 131k 134k 137k 140k 143k 146k 149k
rps12 ndhB tRNA-Leu ycf2 tRNA-Ile rpl2
rps7 ycf15 rpl23
150k 153k 156k 159k 162k 165k
100%
50%
Criteria: 70%, 100 bp
Alignment 1: 168788 bp
Spirodela polyrhiza
Alignment 2: 169337 bp
Wolffiella lingulata
Alignment 3: 168704 bp
Wolffia australiana
X-axis: Lemna minor
window size: 100 bp

Figure 3. Alignment of *Lemnoideae* chloroplast genomes. The sequence of *L. minor* chloroplast genome was compared to those of *S. polyrhiza* (top), *W. lingulata* (middle), *W. australiana* (bottom). Sequences were aligned in mVISTA and the annotation shown above the alignment corresponds to the *L. minor* genome. Grey arrows above the alignment indicate genes and their orientation. Thick black lines show the position of the IRs. The grey peaks determine the percent identity between two sequences of *L. minor* as the reference and our sequenced genomes.
doi:10.1371/journal.pone.0024670.g003

data set, provided a deep enough genome coverage. Assuming that read length will improve for next generation sequencing platforms as they did for conventional methods in the transition from gel to capillary separation techniques, the major advances in shotgun DNA sequencing are now throughput and computational capacity [32].

It is generally assumed that there is a universal transition bias over transversion, probably as a consequence of the fundamental biochemical basis of mutations [33]. This rule appears to hold quite well in many vertebrate species [34] and it also works very well in the *Pooideae* subfamily as we have calculated here. Surprisingly, this is not the case for the *Lemnoideae* subfamily, where a transition bias is absent. Although there is an exemption of transition bias in coding regions of *Spirodela* and *Lemna*, which could be explained by a selection of nonsynonymous substitutions. If all types of substitutions were to be equal, a 1:2 ratio of transition/transversion would be expected because of two possibilities of transitions (AG+CT) and four of transversions (AT+AC+GT+GC). Excluding nucleotide mutations in coding regions from whole genomes of duckweed chloroplasts, the

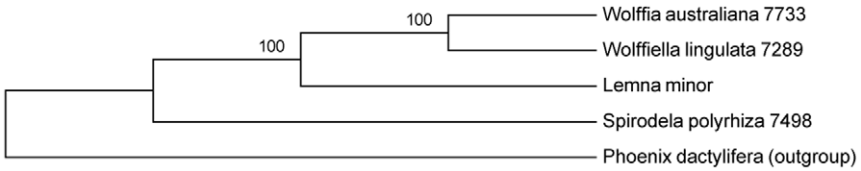


Figure 4. Complete chloroplast genome phylogeny of *Lemnoideae*. The phylogram was drawn by Maximum Parsimony with 1000 replicates of bootstrap test. The tree was rooted by *Phoenix dactylifera* as an outgroup. Support from bootstrap value was shown at the nodes. The GenBank accessions used for the analyses are JN160603 (*S. polyrhiza*), DQ400350 (*L. minor*), JN160604 (*W. lingulata*), JN160605 (*W. australiana*) and GU811709 (*P. dactylifera*). The whole genome sequences were aligned by Multi-LAGAN and MEGA 5 was used to draw the tree.
doi:10.1371/journal.pone.0024670.g004

Table 3. Pairwise sequence divergence of the whole genome and protein coding regions in the subfamily *Lemnoideae* compared with those of the subfamily *Pooideae* (wheat, barley and *Brachypodium*).

Comparative Type	Alignment Region	Pair Alignment	Alignment Length	Substitution Rate[a]	R = si[b]/sv[c]	INDELs Length	INDELs Ratio[d]
intra-tribe	whole genome	S.pol+L.min	141014	0.05	0.7	10262	0.073
intra-tribe	whole genome	W.lin+W.aus	141506	0.04	0.6	8635	0.061
inter-tribe	whole genome	S.pol+W.lin	143722	0.07	0.6	12757	0.089
inter-tribe	whole genome	S.pol+W.aus	142828	0.07	0.6	11849	0.083
inter-tribe	whole genome	L.min+W.lin	142965	0.07	0.6	13543	0.095
inter-tribe	whole genome	L.min+W.aus	141968	0.07	0.6	12429	0.088
intra-tribe	whole genome	wheat+barley	115940	0.02	1.2	4365	0.038
inter-tribe	whole genome	wheat+B.dis	117055	0.04	1.2	6615	0.057
inter-tribe	whole genome	barley+B.dis	116768	0.04	1.3	6196	0.053
intra-tribe	81 Protein genes	S.pol+L.min	69247	0.03	1.1	420	0.006
intra-tribe	81 Protein genes	W.lin+W.aus	69503	0.03	0.8	633	0.009
inter-tribe	81 Protein genes	S.pol+W.lin	69539	0.04	0.9	819	0.012
inter-tribe	81 Protein genes	S.pol+W.aus	69459	0.04	0.9	682	0.010
inter-tribe	81 Protein genes	L.min+W.lin	69521	0.04	0.9	831	0.012
inter-tribe	81 Protein genes	L.min+W.aus	69468	0.04	0.9	748	0.011
intra-tribe	71 Protein genes	wheat+barley	58607	0.01	1.5	290	0.005
inter-tribe	71 Protein genes	wheat+B.dis	58658	0.03	1.7	1045	0.018
inter-tribe	71 Protein genes	barley+B.dis	58647	0.03	1.7	1034	0.018

[a]Substitution Rates = substitution/alignment length;
[b]si (Transitional Pairs) = AG+CT;
[c]sv (Transversional Pairs) = TA+TG+CA+CG;
[d]INDELs Ratio = INDELs length/alignment length. AG means A is mutated to G and others follow the same rules. S.pol = *S. polyrhiza*, L.min = *L. minor*, W.lin = *W. lingulata*, W.aus = *W. australiana,* B.dis = *B. distachon*
doi:10.1371/journal.pone.0024670.t003

number of R-values for non-coding region would be very close to 0.5. In such a case, there would be no significant difference between transition and transversion rates. However, in a study of grasshopper pseudogenes a transition/transversion bias was not universal and both substitution rates reached a 1:1 ratio [35]. Interestingly, transversions could also occur more frequently than transitions in chloroplasts of green algae [36].

Despite the overall high conservation of genome content across different duckweed species, our results demonstrate that substitution rates, insertion and deletion events are more frequent in duckweed chloroplast genomes than in species of the *Pooideae*, especially in non-coding regions (Table 3, Fig. 3). Recent studies also support the observation that *Lemnoideae* have a higher rate of chloroplast sequence evolution relative to *Pistia* and related *Araceae* [37].

Nucleotide substitutions and INDEL mutations are generated during DNA replication or are due to DNA damage [38,39]. Although the enzymes responsible for the maintenance of chloroplast replication and DNA repair are highly faithful, under certain conditions chloroplasts may have to tolerate some level of oxidative damage that occurs spontaneously due to an abundance of reactive oxygen species from the water-splitting activity of the photosystem [36]. Because duckweeds float on water surface, are fully exposed to sunlight, and produce biomass at such a fast rate, their plastid genomes probably transmit and accumulate mutations more frequently than other plants. Once the genome of *Spirodela* has been sequenced, it will be interesting to analyze its nuclear genes that are involved in DNA replication and repair of the plastid genome and how they have evolved compared to terrestrial slow growing plants.

So far, all phylogeny constructions of *Lemnoideae* have used selected genes or partial regions as markers. However, with sequenced chloroplast genomes of four species in this subfamily and the powerful program to align them, it is possible for the first time to perform whole chloroplast genome phylogenetic analysis. The topology of nodes, all with 100% bootstrap values, conforms to the accepted phylogeny based on extensive analysis from morphology and DNA sequence markers. However, there were two nodes that were problematic with only 42% and 53% bootstrap values in *Wolffia* [9]. Therefore, our results contradict the hypothesis that *Wolffia* arose from a merger of *Wolffiella* and *Lemna*, which was based on the *trnL-trnF* marker only [37]. Clearly, the addition of more informative sites from whole genome sequences will improve resolution and confidence in phylogenetic analyses.

In summary, our data gave evidence that next-generation platforms have the capacity to sequence the chloroplast genome at over 1,000 times coverage in just an individual spot on a quadrant slide without plastid purification (Table 2). In order to gain an improved understanding of genome evolution in members of the duckweed subfamily, we generated chloroplast genomes for three species from different genera using *L. minor* as a reference. Our analysis further suggests that (i) gene content is very conserved in duckweeds; (ii) fast nucleotide substitution and abundant INDELs played a key role in the evolution of chloroplast genomes of duckweeds; (iii) duckweed chloroplast genome sequences are very promising to become an elusive single-locus plant barcode for systematic analysis. This information will be critical for the development of a chloroplast transformation system in industrial applications of duckweeds.

Supporting Information

Table S1 Amplification primers for genome gap closure and validation.
(XLS)

Acknowledgments

We thank David Sidote and Mark Diamond for their help with the genome sequencing by SOLiD platform, as well as David Sidote's useful suggestions for bioinformatics analysis.

Author Contributions

Conceived and designed the experiments: WW JM. Performed the experiments: WW. Analyzed the data: WW JM. Contributed reagents/materials/analysis tools: WW. Wrote the paper: WW JM.

References

1. Delwiche CF, Timme RE (2011) Plants. Curr Biol 21: R417–422.
2. Palmer JD (1985) Comparative Organization of Chloroplast Genomes. Annu Rev Genet 19: 325–354.
3. Kolodner R, Tewari KK (1979) Inverted repeats in chloroplast DNA from higher plants. Proc Natl Acad Sci USA 76: 41–45.
4. Shaver J, Oldenburg D, Bendich A (2006) Changes in chloroplast DNA during development in tobacco, Medicago truncatula, pea, and maize. Planta 224: 72–82-82.
5. Lutz K, Wang W, Zdepski A, Michael T (2011) Isolation and analysis of high quality nuclear DNA with reduced organellar DNA for plant genome sequencing and resequencing. BMC Biotechnol 11: 54.
6. Bortiri E, Coleman-Derr D, Lazo G, Anderson O, Gu Y (2008) The complete chloroplast genome sequence of Brachypodium distachyon: sequence comparison and phylogenetic analysis of eight grass plastomes. BMC Research Notes 1: 61.
7. Bock R (2007) Plastid biotechnology: prospects for herbicide and insect resistance, metabolic engineering and molecular farming. Curr Opin Biotechnol 18: 100–106.
8. Cabrera LI, Salazar GA, Chase MW, Mayo SJ, Bogner J, et al. (2008) Phylogenetic relationships of aroids and duckweeds (Araceae) inferred from coding and noncoding plastid DNA. Am J Bot 95: 1153–1165.
9. Les DH, Crawford DJ, Landolt E, Gabel JD, Kimball RT (2002) Phylogeny and Systematics of Lemnaceae, the Duckweed Family. Syst Botany 27: 221–240.
10. Wang W, Wu Y, Yan Y, Ermakova M, Kerstetter R, et al. (2010) DNA barcoding of the Lemnaceae, a family of aquatic monocots. BMC Plant Biology 10: 205.
11. Cheng JJ, Stomp AM (2009) Growing Duckweed to Recover Nutrients from Wastewaters and for Production of Fuel Ethanol and Animal Feed. CLEAN – Soil, Air, Water 37: 17–26.
12. Landolt E (1986) The family of Lemnaceae - a monographic study, Vol 1: Veroffentlichungen des Geobotanischen Institutes ETH, Stiftung Rubel, Zurich.
13. Mardanov A, Ravin N, Kuznetsov B, Samigullin T, Antonov A, et al. (2008) Complete Sequence of the Duckweed (Lemna minor) Chloroplast Genome: Structural Organization and Phylogenetic Relationships to Other Angiosperms. J Mol Evol 66: 555–564-564.
14. Cronn R, Liston A, Parks M, Gernandt DS, Shen R, et al. (2008) Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. Nucleic Acids Res 36: e122–e122.
15. Diekmann K, Hodkinson TR, Fricke E, Barth S (2008) An Optimized Chloroplast DNA Extraction Protocol for Grasses (Poaceae) Proves Suitable for Whole Plastid Genome Sequencing and SNP Detection. PLoS ONE 3: e2813.
16. Atherton R, McComish B, Shepherd L, Berry L, Albert N, et al. (2010) Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. Plant Methods 6: 22.
17. Nock CJ, Waters DLE, Edwards MA, Bowen SG, Rice N, et al. (2010) Chloroplast genome sequences from total DNA for plant identification. Plant Biotech J 9: 328–333.
18. Gardner RC, Howarth AJ, Hahn P, Brown-Luedi M, Shepherd RJ, et al. (1981) The complete nucleotide sequence of an infectious clone of cauliflower mosaic virus by M13mp7 shotgun sequencing. Nucleic acids research 9: 2871–2888.
19. Vieira J, Messing J (1982) The pUC plasmids, an M13mp7-derived system for insertion mutagenesis and sequencing with synthetic universal primers. Gene 19: 259–268.
20. Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res 8: 4321–4325.
21. Sasson A, Michael TP (2010) Filtering error from SOLiD Output. Bioinformatics 26: 849–850.

22. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.
23. Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821–829.
24. Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20: 3252–3255.
25. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, et al. (2003) LAGAN and Multi-LAGAN: Efficient Tools for Large-Scale Multiple Alignment of Genomic DNA. Genome Res 13: 721–731.
26. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: computational tools for comparative genomics. Nucleic Acids Res 32: W273–W279.
27. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Mol Biol Evol.
28. Yang M, Zhang X, Liu G, Yin Y, Chen K, et al. (2010) The Complete Chloroplast Genome Sequence of Date Palm (Phoenix dactylifera L.). PLoS ONE 5: e12762.
29. Wang W, Kerstetter R, Michael T (2011) Evolution of genome size in duckweeds (Lemnaceae). Journal of Botany.
30. Paszkiewicz K, Studholme DJ (2010) De novo assembly of short sequence reads. Brief Bioinform 11: 457–472.
31. den Bakker H, Cummings C, Ferreira V, Vatta P, Orsi R, et al. (2010) Comparative genomics of the bacterial genus Listeria: Genome evolution is characterized by limited gene acquisition and limited gene loss. BMC Genomics 11: 688.
32. Larson R, Messing J (1982) Apple II software for M13 shotgun DNA sequencing. Nucleic acids research 10: 39–49.
33. Wakeley J (1996) The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. Trends Ecol Evol 11: 158–162.
34. Nachman MW, Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. Genetics 156: 297–304.
35. Keller I, Bensasson D, Nichols RA (2007) Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes. PLoS Genet 3: e22.
36. Guhamajumdar M, Sears BB (2005) Chloroplast DNA base substitutions: an experimental assessment. Mol Genet Genomics 273: 177–183.
37. Rothwell GW, Van Atta MR, Ballard HE, Stockey RA (2004) Molecular phylogenetic relationships among Lemnaceae and Araceae using the chloroplast trnL-trnF intergenic spacer. Mol Phylogenet Evol 30: 378–385.
38. Friedberg EC (2003) DNA damage and repair. Nature 421: 436–440.
39. Magee AM, Aspinall S, Rice DW, Cusack BP, Semon M, et al. (2010) Localized hypermutation and associated gene losses in legume chloroplast genomes. Genome Res 20: 1700–1710.

OPEN ACCESS Freely available online

PLOS ONE

The Mitochondrial Genome of an Aquatic Plant, *Spirodela polyrhiza*

Wenqin Wang, Yongrui Wu, Joachim Messing*

Waksman Institute of Microbiology, Rutgers, The State University of New Jersey, Piscataway, New Jersey, United States of America

Abstract

Background: *Spirodela polyrhiza* is a species of the order Alismatales, which represent the basal lineage of monocots with more ancestral features than the Poales. Its complete sequence of the mitochondrial (mt) genome could provide clues for the understanding of the evolution of mt genomes in plant.

Methods: *Spirodela polyrhiza* mt genome was sequenced from total genomic DNA without physical separation of chloroplast and nuclear DNA using the SOLiD platform. Using a genome copy number sensitive assembly algorithm, the mt genome was successfully assembled. Gap closure and accuracy was determined with PCR products sequenced with the dideoxy method.

Conclusions: This is the most compact monocot mitochondrial genome with 228,493 bp. A total of 57 genes encode 35 known proteins, 3 ribosomal RNAs, and 19 tRNAs that recognize 15 amino acids. There are about 600 RNA editing sites predicted and three lineage specific protein-coding-gene losses. The mitochondrial genes, pseudogenes, and other hypothetical genes (ORFs) cover 71,783 bp (31.0%) of the genome. Imported plastid DNA accounts for an additional 9,295 bp (4.1%) of the mitochondrial DNA. Absence of transposable element sequences suggests that very few nuclear sequences have migrated into *Spirodela* mtDNA. Phylogenetic analysis of conserved protein-coding genes suggests that *Spirodela* shares the common ancestor with other monocots, but there is no obvious synteny between *Spirodela* and rice mtDNAs. After eliminating genes, introns, ORFs, and plastid-derived DNA, nearly four-fifths of the *Spirodela* mitochondrial genome is of unknown origin and function. Although it contains a similar chloroplast DNA content and range of RNA editing as other monocots, it is void of nuclear insertions, active gene loss, and comprises large regions of sequences of unknown origin in non-coding regions. Moreover, the lack of synteny with known mitochondrial genomic sequences shed new light on the early evolution of monocot mitochondrial genomes.

Citation: Wang W, Wu Y, Messing J (2012) The Mitochondrial Genome of an Aquatic Plant, *Spirodela polyrhiza*. PLoS ONE 7(10): e46747. doi:10.1371/journal.pone.0046747

Editor: Senjie Lin, University of Connecticut, United States of America

Received July 12, 2012; **Accepted** September 4, 2012; **Published** October 4, 2012

Funding: The authors have no support or funding to report.

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

* E-mail: messing@waksman.rutgers.edu

Introduction

Usually, a plant cell contains three genomes: plastid, mitochondrial, and nuclear. In a typical *Arabidopsis* leaf cell, there are about 100 copies of mitochondrial DNA (mtDNA), about 1,000 copies of chloroplast DNA (cpDNA), and two copies of nuclear DNA (ncDNA) [1].

The mitochondrial genome plays fundamental roles in development and metabolism as the major ATP production center via oxidative phosphorylation [2]. The mitochondrial genetic system in flowering plants exhibit multiple characteristics that distinguish them from other eukaryotes: large genome size with dispersed genes, an incomplete set of tRNAs, trans-splicing, and frequent uptake of plastid DNA or of foreign DNA fragments by horizontal and intracellular gene transfer [2,3,4,5,6]. Plant mtDNAs are a major resource for evolutionary studies, because coding regions evolve slowly, in contrast to the flexible non-coding DNA. Therefore, the structural evolution and plasticity of plant mtDNAs make them powerful model for exploring the forces that affect their divergence and recombination.

With the emergence of second-generation sequencing technologies, the number of completed plant mitochondrial genomes deposited in the GenBank database (http://www.ncbi.nlm.nih.gov/genomes/GenomesGroup.cgi?taxid = 33090&opt = organelle. Accessed 2012 Sep 11) has increased until August of 2012 to 69. Most are from Chlorophyta (17 of green algae) and seed plants (26 of eudicotyledons). So far, among 11 sequenced monocot mt genomes, 10 are from the Poales, which have been extensively studied and only one, *Phoenix*, a palm, from the order of Arecale has been sequenced [7]. Obviously, complete mt sequence data will be needed not only from closely but also distant related taxa to give us a broader perspective of mt genome organization and evolution.

Spirodela polyrhiza, with great potential for industrial and environmental applications, is a small, fast growing aquatic plant in the Araceae family of the Alismatales order [8,9]. There are 14 families, 166 genera, and about 4,500 species in this order. The early diverging phylogenetic position of Alismatales offers a broader view at features of monocot mt genomes. Plant mitochondria could also open a strategy for transgenes with high

expression level and biological containment because of their maternal inheritance [10]. Here, we demonstrate the *de novo* assembly of a complete mt genome sequence from total leaf DNA using the SOLiD sequencing platform and a genome copy number-sensitive algorithm that can filter chloroplast and nuclear sequences. Indeed, comparative analysis of this genome provides us with unique features and new insights of this class of plants that differ from other monocots.

Materials and Methods

DNA Isolation and SOLiD DNA Sequencing

The methods for DNA extraction and DNA sequencing by the SOLiD platform followed a protocol as previously published [11]. Briefly, total genomic DNA was extracted from the clonally grown whole plant tissue of *Spirodela polyrhiza*. A mate-paired library was made with 1.5 Kb insertions and read length was 50 bp. Since nucleic, mitochondrial and chloroplast sequence all exist in reads from total DNA preparation, copy number between three genomes was significantly different [12,13], so that it was feasible to *de novo* assembly both chloroplast and mitochondria genomes using the same dataset but with different coverage cut-off numbers as described previously [11].

Genome Assembly, Finishing and Validation

The coverage cut-off was fully utilized to only allow the target organellar genome to be assembled due to obvious differentiation of copy number for three genomes in total reads [12]. Furthermore, low-level contaminating sequences from foreign DNA (mainly nuclear DNA) were discarded by this approach. Quality control and other details were described recently [11]. Before we assembled the mitochondria genome using mate-paired reads, we masked chloroplast reads to reduce effects due to plastid sequence predominance. The detailed pipeline was shown below (Fig. 1).

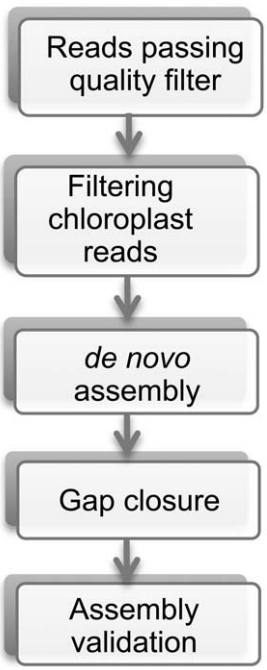


Figure 1. Pipeline of mitochondrial genome assembly. Details were described in Methods.
doi:10.1371/journal.pone.0046747.g001

1) Filtering chloroplast reads: we mapped total high quality reads to existing chloroplast genome (GenBank # JN160603) by BWA short-read alignment component with default parameters [14]. Only unmapped reads were used in the next step. 2) *de novo* assembly: the assembly was executed using the SOLiD^{TM} System *de novo* Accessory Tools 2.0 (http://solidsoftwaretools.com/gf/project/denovo/) in conjunction with the Velvet assembly engine [15]. 3) Gap closure: since chloroplast reads were pre-removed before mitochondrial assembly, theoretically, any location with chloroplast insertion in mtDNA would create a gap. Using flanking primers bridging 57 gaps, the missing sequences were amplified and sequenced with the ABI 3730×l system, yielding a complete contiguous mtDNA sequence (Table S1). To validate the circularity of the *Spirodela* mtDNA, PCR products were sequenced with pairs of primers bridging gaps and overlapping with the assembled linear scaffold. 4) Most gaps were small enough for single CE (capillary electrophoresis) sequence reads and overlapping sequences served as a measure for the accuracy of the SOLiD assembly and error rate. Therefore, PCR amplification and CE sequence provided validation of the order of contigs and also revealed sequencing discrepancies between these two platforms.

Genome Annotation and Sequence Analysis

The main pipeline for mitochondrial genome annotation was adapted from other sources [5]. Databases for protein-coding genes, rRNA and tRNA genes were compiled from all previously sequenced seed plant mitochondrial genomes. BLASTX and tRNAscan-SE were the mainly used programs [5]. The boundaries for each gene were manually curated. The sequin file including sequence and annotation was submitted to NCBI GenBank as JQ804980. The graphical gene map was processed by OrganellarGenomeDRAW program [16]. The codon usages for all protein coding genes in *Spirodela* and *Oryza* were calculated by using the Sequence Manipulation Suite [17].

Cp-derived tRNAs were identified by aligning all tRNA in annotated cpDNA to mtDNA with 80% of identity, an e-value of 1e-10 and a 50% coverage threshold. All remaining sequences were further scanned by EMBOSS getorf for open reading frames (ORFs) with more than 300 bp [18].

Putative RNA editing sites in protein-coding genes were identified by the PREP-mt Web-based program based on the evolutionary principle that editing increases protein conservation among species (http://prep.unl.edu/. Accessed 2012 Sep 11) [19]. The optimized cut-off value 0.6 was set in order to achieve the maximal accurate prediction. RNA editing sites from four genes were validated by RT-PCR with gene-specific primers (Table S2).

Sequences transferred to mtDNA were found by BLASTN search of mtDNA against the *Spirodela* chloroplast genome with 80% of identity, e-value of 1e-10 and 50 bp of length threshold. Repeat sequence analysis was predicted by using REPuter web-based interface, including forward, palindromic, reverse and complemented repeats with a cut-off value of 50 bp [20]. The mitochondrial genome was screened by repeatmasker under cross_match search engine (http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker. Accessed 2012 Sep 11) for interspersed repeats and low complexity DNA sequences [21].

Phylogenetic Analysis

We aligned 19 homologous protein-coding gene sequences (*nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5*, *nad6*, *nad7*, *nad9*, *cob*, *cox1*, *cox2*, *cox3*, *atp1*, *atp4*, *atp6*, *atp8*, *atp9* and *rps3*) from the *Spirodela* mitochondrial genome and other seven plant organisms (Table S8, *Cycad*, NC_010303; *Phoenix*, NC_016740; *Spirodela*, JQ804980; *Oryza*, NC_011033; *Zea*, NC_007982; *Boea*, NC_016741; *Nicotiana*,

NC_006581; *Arabidopsis*, NC_001284) and constructed a phylogenetic tree. Annotations were revalidated and sequences were concatenated into a single continuous sequence from 18,537 to 19,041 bp to initiate alignment by MEGA5 [22]. The phylogeny of the mitochondrial genome was estimated by maximum likelihood (ML) with 1,000 Bootstrap of replicates. *Cycas* was used as the outgroup.

Comparison of Global Genome Structure

The conserved regions for protein-coding and rRNA genes were identified between *Spirodela* and *Oryza* sequences by BLASTN. The synteny together with the annotation file were uploaded to a web-based genome synteny viewer GSV [23]. The relative ordering of a set of homologous genes was illustrated in Fig. 2.

Results and Discussion

The *de novo* Assembly of SOLiD Reads

The optimal parameter of the SOLiD^{TM} System *de novo* Accessory Tools 2.0 for the assembly of the *Spirodela* mitochondrial genome has a hash length of 25 and coverage cut-off of 45. Under these conditions, assembly of SOLiD reads from total leaf DNA resulted in 15 scaffolds and 88 contigs, of which three scaffolds were mitochondrial (173,697, 47,896, 1824 bp) (Table 1). As expected, the other scaffolds were mainly copies of ribosomal RNA genes and retroelements of the nuclear genome because their copy number was comparable to the copies of mitochondrial genomes per leaf cell. To validate the assemblies, gaps were amplified with PCR for dideoxy sequencing with the CE ABI 3730×l system. With this information the order of the three scaffolds were resolved. Furthermore, after the SOLiD short read assembly was aligned with the CE long read sequences, only 0.036% discrepancy was found within 19 Kb sequence of overlaps, demonstrating high consistency between the two platforms. When we mapped the total reads back to the complete mtDNA, a total of 467-fold coverage was calculated. Considering the 5,474-fold chloroplast coverage, we found 41-fold coverage of nuclear genome sequences (Table 1). This level of coverage from assembled sequences was consistent with the expected representation of the three genomes in total leaf DNA, yielding chloroplast, mitochondria, and nuclei with the approximate ratio of 100:10:1.

Here, we applied a layered approach of sequencing organelle genomes without fractionation from total leaf DNA. Thanks to an assembly algorithm of sequence reads that is sensitive to the differential copy number of organelle and nuclear genomes, we did not physically need to fractionate plastid, mitochondrial, and nuclear DNA for deep sequencing. Therefore, we first assembled the complete *Spirodela* chloroplast genome from ABI SOLiD and gap-closure 3730xl reads, which permitted us to mask all plastid DNA reads before assembling mitochondrial DNA, which is in access of nuclear DNA but not as abundant as plastid DNA [11,24]. Furthermore, we can take advantage of the ratios of these genomes to limit the value of coverage cut-off with identical dataset of SOLiD reads, which is taken in consideration for the assembly algorithm to distinguish between plastid, mitochondrial, and nuclear genome sequence reads [12,13]. Assemblies were validated like in the case of chloroplast DNA by PCR and gap sequencing of long reads with the traditional ABI 3730×l sequencing system. Following this protocol, we obtained a complete mitochondrial genome from an aquatic plant in a very cost-efficient way, which can serve as a reference for future mt genomics.

Features of the *Spirodela* Mitochondrial Genome

The mitochondrial genome was assembled into a 228,493 bp master circle (Fig. 3), which makes it the smallest genome of all sequenced monocots, much smaller than the 715,001 bp of *Phoenix dactylifera* [7], 490,520 bp of *Oryza sativa* [25], or 569,630 bp of *Zea mays* mitochondria [26]. Because *Spirodela* diverged at a very early stage in the monocot lineage, it suggests that either the common ancestor of monocots had a relatively compact genome, with a series of independent expansions by accumulation of chloroplast and nuclear sequences or proliferation of pairs of repeats, leading to the large genomes in rice and maize [5,25,26], or a number of size contractions happened in *Spirodela* from the large genome of their ancestor. The GC content in the mtDNA was 45.7%, slightly higher than 43.8% of *Oryza* and 43.9% of *Zea* [25,26]. The coding sequences covered 31% of the mitochondrial genome compared with 57.4% of the chloroplast genome [11] (Table 2). There were 57 functional genes and 4 pseudogenes in total, encoding 35 proteins, 19 tRNAs and 3 rRNAs (Table S3). Therefore, it gave rise to a density of 4.0 Kb per gene. Noticeably, eight genes (*ccmFc*, *cox2*, *nad1*, *nad2*, *nad4*, *nad5*, *nad7*, *rps3*) had 15 *cis*-spliced group II introns, whereas *nad1*, *nad2* and *nad5* were disrupted by 6 *trans*-splicing sites (Table 2 and S1). Previous studies suggested that *trans*-splicing had evolved before the emergence of hornworts [27]. In general, the numbers and locations of introns in the *Spirodela* mtDNA were rather well conserved in other sequenced monocot genomes.

Protein Genes and Transcript Editing

The content of key protein coding genes in *Spirodela* mtDNA is highly conserved with other angiosperms [26,28,29,30]. There were nine subunits of the oxidative phosphorylation complex I (*nad1*, 2, 3, 4, 4L, 5, 6, 7 and 9); one subunit of complex II (*sdh4*); one subunit of complex III (*cob*); three subunits of complex IV (*cox1*, *cox2* and *cox3*); five subunits of complex V (*atp1*, 4, 6, 8 and 9); and four subunits of a complex involved in cytochrome c biogenesis (*ccmB*, *ccmC*, *ccmFn* and *ccmFc*). Other genes encoding maturase (*matR*) and transport membrane protein (*mttB*) were also present in *Spirodela* mtDNA. As in maize [26], the matR gene in *Spirodela* also resided in the intron 4 of *nad1*, which is trans-spliced after transcription. In *Spirodela*, there were ten functional ribosomal genes and two pseudogenes of *rps14* and *rps19* with early stop codons, whereas rice had a functional *rps19* and a non-functional *rps14* [25] and both were missing in maize (Table S3) [26]. All annotated genes and coordinates were listed in Table S3 and shown in a graphical map (Fig. 3).

Post-transcriptional editing occurs in nearly all plant mitochondria, which results in altered amino acid sequences of the translated protein by converting specific Cs into Us in their transcripts. We used the program of the predictive RNA editor of plant mitochondrial genomes (PREP-mt) to predict the location of RNA editing sites, which are based on well-known principles that plant organelles maintain the conservation of protein sequences across many species by editing mRNA [19]. By setting the cut-off value to 0.6 within the 35 protein-coding genes of *Spirodela* mtDNA 600 sites were predicted as C-to-U RNA editing sites (Table S4). To validate the accuracy of this prediction, we compared RNA transcripts from *atp9*, *nad9*, *cox3* and *rps12* by RT-PCR with the corresponding genomic sequences yielding a confirmation for 90.8% of the predicted sites. Considering a level of about 10% artificial predictions, we estimate about 540 RNA editing sites, a number that lies between the 441 of protein-coding genes of *Oryza* [25] and 1,084 of *Cycas* [29].

It is generally accepted that RNA editing is essential for functional protein expression as it is required to modify amino

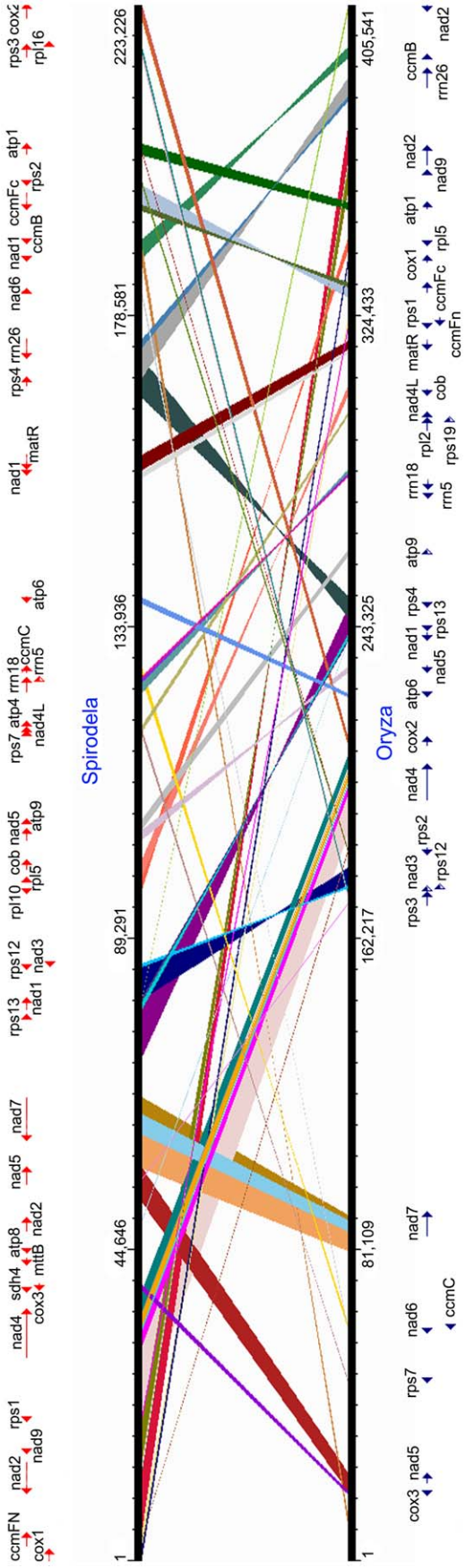


Figure 2. Comparison of synteny in conserved gene loci of *Spirodela* and *Oryza* mitochondrial genomes. The annotated protein-coding genes were indicated for *Spirodela* and *Oryza*. Major conserved regions were bridged by lines. The visualized genome synteny was performed by GSV: a web-based genome synteny viewer [23].
doi:10.1371/journal.pone.0046747.g002

acids to maintain appropriate structure and function [31], or to generate new start or stop codons [32]. Indeed, the abundance of RNA editing sites in *Spirodela* mtDNA might have increased genome complexity and pace of divergence. We summarized the number of potentially modified codons of *Spirodela* mtDNA in Table S5. Three edited codons (TCA (S) = >TTA (L); TCT (S) = >TTT (F); CCA (P) = >CTA (L)) were found most frequently, whereas three editing events from two codons (CAA (Q) = >TAA (X); CAG (Q) = >TAG (X)) resulted in stop codons (Table S5). Even though three new stop codons are located close at the carboxyterminal end of proteins (ccmC, rps1 and rpl16), it is not clear whether these small truncations affect their functions or not, which would require experimental evidence.

The rRNA, tRNA Genes and Codon Usage

Spirodela mtDNA contains 3 ribosomal RNA genes (*rrn5*, *rrn18*, *rrn26*) and one pseudogene of *rrn26*. The 19 putatively expressed tRNA genes are specific for 15 amino acids (Table S3). Four of them (*trnN*-GTT, *trnH*-GTG, *trnM*-CAT and *trnS*-GGA) are probably chloroplast-derived because of high sequence similarity. They are also predicted as chloroplast origin in maize, rice, sugar beet and *Arabidopsis* except *trnS*-GGA in maize [26]. Therefore, they were not recently acquired from chloroplast, but more likely an event of horizontal transfer in a common ancestor. One *trnH*-GTG is considered to be a non-functional pseudogene. Functional tRNA genes for the amino acids Ala, Arg, Leu, Thr and Val are absent. Because all 20 amino acids are required for protein synthesis, and all 64 codons are used in the *Spirodela* mt genome based on a codon-usage scan (Fig. 4 and Table S6) [17], the missing tRNAs are presumably encoded by the nuclear genome

Table 1. *de novo* assembly statistics for the *Spirodela* mitochondrial genome.

Statistical list	Number
Number of scaffolds	15
N50 scaffolds (bp)	173,697
Number of contigs	88
N50 contigs (bp)	6,528
Sum contig length (bp)	240,987
Hash length	25
Expected coverage	90
Coverage cut-off[a]	45
Total reads (X10^6)	153
Aligned reads (%)	1.4
Average chloroplast coverage[b]	5,474
Average mitochondrial coverage	467
Average nuclear coverage	41

[a]Coverage cut-off: minimum coverage required to form a contig.
[b]Average chloroplast coverage was cited from *Spirodela* chloroplast genome assembly [11].
doi:10.1371/journal.pone.0046747.t001

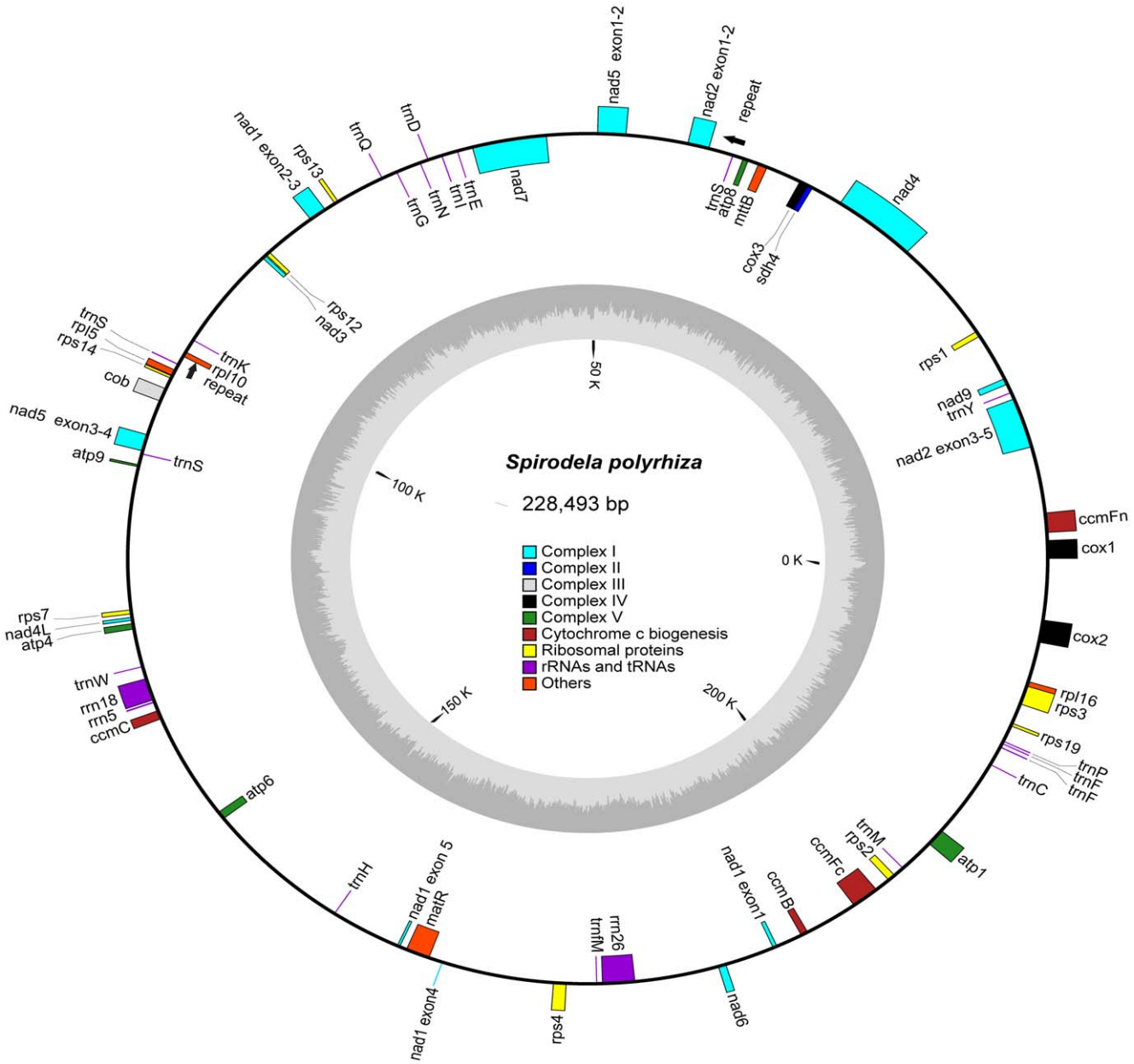


Figure 3. The gene map of *Spirodela polyrhiza* mitochondrial genome. Genes indicated as closed boxes on the outside of the circle are transcribed clockwise, whereas those on the inside were transcribed counter-clockwise. Pseudogenes were indicated with the prefix "Ψ". The biggest repeat pair was also marked by arrows. The genome coordinate and GC content are shown in the inner circle.
doi:10.1371/journal.pone.0046747.g003

and imported from the cytosol into the mitochondria [33–34]. We also found that the two codons for TAT-Tyr and TTT-Phe are highly preferred in *Spirodela* and *Oryza* and overall other codon usage is rather similar between the two species (Table S6).

ORFs and Intergenic Sequences

Only ORFs encoded by a hypothetical gene with more than 300 bp in length between start and stop codons and no match with a known mt coding sequence were counted. Based on this cut-off, we found 39 mitochondrial ORFs, most of which were not cp migrations and specific to *Spirodela* (Table S3). We named ORFs using their amino acid numbers. When the same length of ORFs happened, a lower case letter (a, b, c, etc) was added. Given the large amount of intergenic DNA in *Spirodela* mtDNA, it is not surprising to find an abundance of additional ORFs in its genome. Rarely, ORFs showed conservation to any other plants so that putative ORFs were considered to be spurious prediction [35]. However, orf100a had an ortholog of a NADH-ubiquinone oxidoreductase chain in *Nicotiana tabacum* (GenBank: YP_717128) and orf257 had sequence similarity to DNA polymerase (GenBank: YP_003875487) found in plant mt plasmids [4]. Some studies found that unidentified ORFs had transcripts in rapeseed [36] or to be actively transcribed in sugar beet [37], but further studies are needed to determine whether they encode functional proteins.

A striking feature of *Spirodela* mtDNA was that 81% of the intergenic regions were species-specific and showed no sequence

Table 2. Summary of general features for *Spirodela* mitochondrial genome.

Feature	Value
Genome size (bp)	228,493
GC content (%)	45.7
Coding sequences (%)[a]	31.4
Protein coding gene #	35
ORFs #	39
cis−/*trans-intron* #	15/6
tRNA gene #	19
rRNA gene #	3
Chloroplast-derived (%)	4.1
Gene density (bp)	4009

[a]coding sequences include identified mitochondrial genes, pseudogenes, ORFs and *cis*-spliced introns.
doi:10.1371/journal.pone.0046747.t002

similarity to any other known sequence. It seemed that anonymous sequences in intergenic DNA were quite common. For instance, unidentifiable sequences comprised 70% of *Beta vulgaris* mtDNA [38]. Although they split about 50 million years ago, 76% of rice mtDNA sequences appeared to be highly divergent from maize in intergenic regions [26]. The repetitive DNAs [39], mt plastidal migrations [40] and viral DNA insertions [41] could contribute to the expansion of intergenic regions, but still comprised a rather small fraction in most seed plant mt genomes. On the other hand, it was quite common that multipartite mt genomes could be generated through large repeat pairs with high frequency [35]. Indeed, 29 potential candidates of repeat pairs with more than 50 bp were found in *Spirodela* mtDNA by using REPuter [20] (Table S7). However, we could not detect repeat-specific contigs from the assembly that could be explained of isomeric and subgenomic molecules derived from a master circle after recombination. Probably, the high rate of non-coding sequence turnover in *Spirodela* mtDNA was mainly generated through the process of micro-homologous recombination or non-homologous end joining, later on of active rearrangement and continuous reshuffling. Still, the high proportion of enigmatic non-coding regions in mtDNA is quite extensive. To understand where all these enigmatic sequences might come from and why they appeared to be so common would require additional sequences from closely related species.

Phylogenetic Analysis and Gene Loss in Angiosperm Mitochondrial Genomes

After re-examining mitochondrial genome annotations from seven species, a selection of 19 conserved genes (*nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5*, *nad6*, *nad7*, *nad9*, *cob*, *cox1*, *cox2*, *cox3*, *atp1*, *atp4*, *atp6*, *atp8*, *atp9* and *rps3*) was concatenated to permit alignment analysis of 19,824 sites in eight genomes, listed in Table S8 (dicot: *Arabidopsis*, *Nicotiana* and *Boea*; monocot: *Spirodela*, *Phoenix*, *Oryza* and *Zea*; outgroup: *Cycas*). The gene tree topology from multiple loci (Fig. 5) was largely congruent with the known phylogenetic relationships inferred from analysis of rbcL. There were two subclades of monocots and dicots within the angiosperm [42]. Previous studies of fossil records [43], morphology and molecular analysis [44] also supported that Alismatales (*Spirodela*) was a basal monocot followed by Arecales (*Phoenix*), whereas the Poales (rice and maize) resided in the most developed positions.

The loss of protein coding and tRNA genes in seven genomes relative to the outgroup was examined based on the phylogenetic tree. Generally, most losses were limited in their phylogenetic depth to a single family and must have occurred recently (Fig. 5). Three ribosomal protein genes *rps10*, *rps11* and *rpl2* were missing in *Spirodela* mtDNA. Frequent gene losses of ribosomal protein genes also occurred in other species. At a closer look, *rps2* seemed to have been lost early in the evolution of dicots, whereas *rps2* was

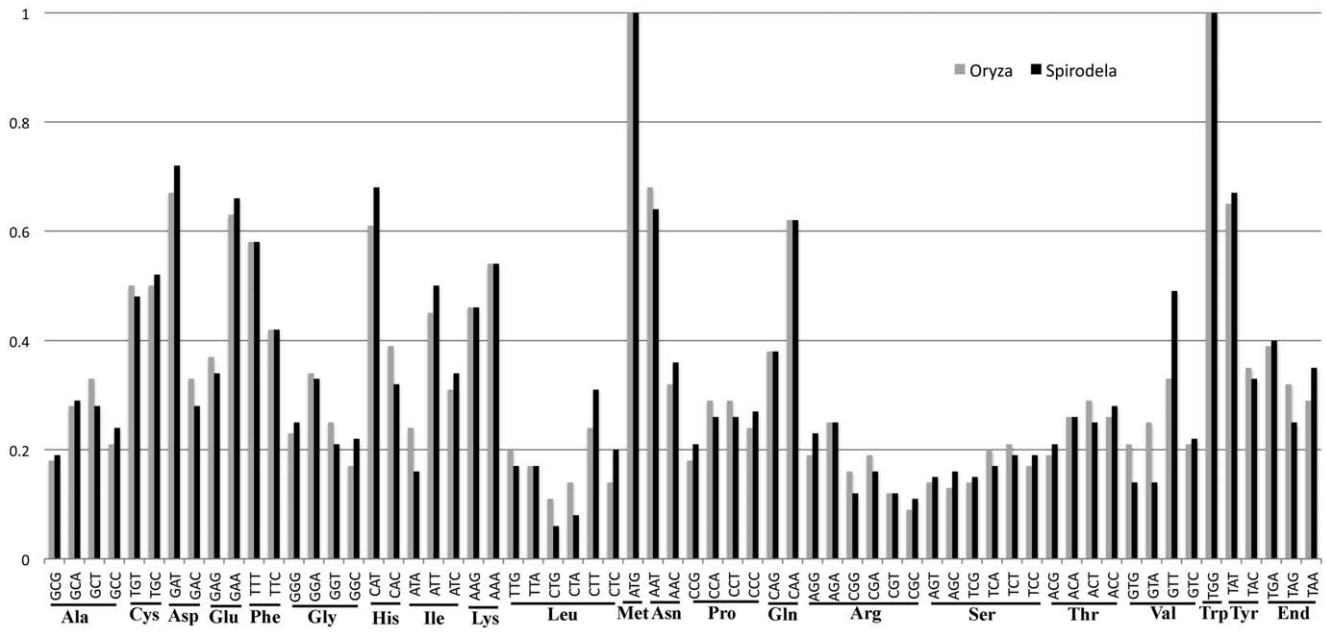


Figure 4. The fraction of each codon usage among the same amino acid in *Spirodela* compared to that in *Oryza*. Black bar was *Spirodela* and grey was *Oryza*. The fraction of each codon usage was shown on Y-axis.
doi:10.1371/journal.pone.0046747.g004

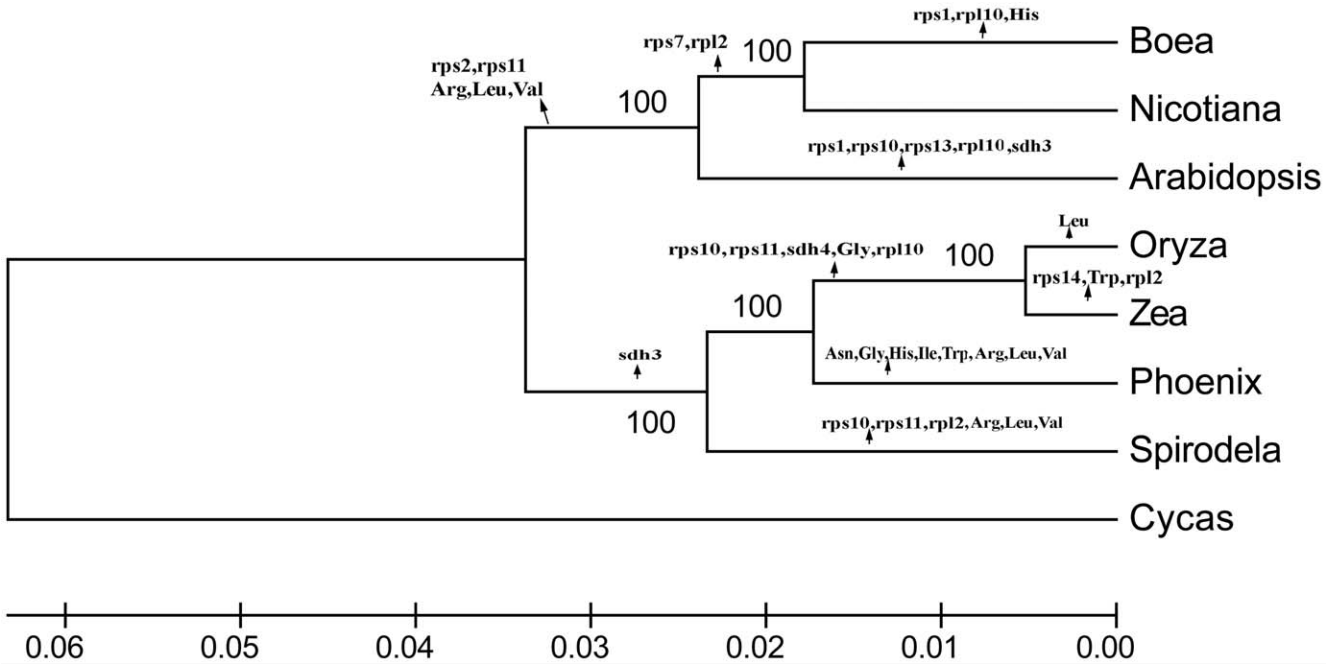


Figure 5. Phylogenetic tree based on 19 conserved genes in mitochondrial genomes. The ML calculation was run by MEGA5 with 1,000 bootstrap replicates. All the gene losses were mapped on the tree branches. *Cycas* was included in the analysis as an outgroup. The signs of Amino Acid (Ala, Arg, Leu, Thr, His, Trp, Ile, Gly, Leu and Val) mean corresponding functional tRNA genes were absent in their mtDNAs.
doi:10.1371/journal.pone.0046747.g005

present in *Cycas*, *Marchantia*, and other monocots [45]. The *rps11* gene was missing in dicots (*Arabidopsis*, *Nicotiana* and *Boea*) and also in some monocots (*Spirodela, Oryza* and *Zea*). The corresponding mt *rps2* and *rps11* genes have been transferred to the nucleus in *Arabidopsis*, soybean, and tomato, suggesting that gene loss followed functional transfer to the nucleus [6,45]. The unparallel loss of *rps11* and *rpl2* in *Spirodela* compared with other monocots suggested that the loss of many genes might have occurred independently in various lineages during speciation of angiosperms. The *sdh3* gene was absent and the *sdh4* gene was present in both *Spirodela* and *Phoenix*, whereas neither was retained in rice and maize (Fig. 5 and Table S8). A previous study showed that *sdh4* losses were concentrated in the monocots and no losses were detected in basal angiosperms by Southern blot survey of 280 angiosperm genera, which further showed most of the losses were limited in phylogenetic depth to a single family [46].

Our data lend support to previous studies that most gene losses occurred with mt ribosomal protein genes and rarely with respiratory genes, which was well documented with a Southern blot survey of gene distribution in 281 diverse angiosperms [6]. When a gene was missing from mtDNA of a given species, it was generally assumed that the original copy had been transferred to the nucleus. Therefore, our results strongly suggested that intracellular gene transfer of ribosomal protein and tRNA genes from mitochondria to the nuclear genome was a frequent process, which in return allowed the nucleus to control the organelle by encoding organelle-destined proteins [33,34]. Still, functional copies of these putative transferred genes will have to be confirmed after the whole nuclear genome sequence will be available. The finding of many intermediate stages of the *cox2* gene transfer in legumes had shown that physical movement of mtDNA to the nuclear genome was an ongoing process [47].

Chloroplast DNA Insertions

The *Spirodela* mtDNA contained multiple cp-originated insertions, ranging in size from 69 to 1,048 bp. These sequences added up to 9,295 bp of the total amount of transferred cpDNA (Table S9), accounting for 4.07% of the mtDNA. A total of 4,436 bp was derived from the inverted repeats of the chloroplast genome, whereas 4,859 bp was transferred from single copy regions of cpDNA. The similarity level of each insertion to the chloroplast genome varied between 75% and 100%. Moreover, the migrated plastid fragments had 732 substitutions, 28 insertions, and 49 deletions within 9,295 bp. They also contained fragments of plastid genes, such as *psbA*, *petB*, *psbC* and *ycf1* (Table S9). All of the protein-coding genes of plastid origin in *Spirodela* mtDNA were likely to be non-functional as a result of truncations and mutations, whereas four tRNAs of plastidal origin appeared to be intact. Indeed, chloroplast-derived sequences were very common in plant mt genomes, such as 6% in rice [25], 4% in maize [26] and 1% in *Arabidopsis* [28]. Surprisingly, 42.4% of the chloroplast genome of *Vitis* has been incorporated into its mt genome [48]. And a large segment of 113 Kb from chloroplast sequences was captured by the *Cucurbita* mt genome [5].

Integrated Nuclear DNA

It is believed that transposable elements in mitochondria are nuclear-derived and are therefore common in mt intergenic regions [38,49]. For instance, 4% of *Arabidopsis* mtDNA was probably derived from transposons of nuclear origins [28]. Four fragments of transposable elements were found in maize mtDNA [26] and nineteen were identified in rice [25]. However, we could not find any transposons in the *Spirodela* mt genome when we searched against the Repbase repetitive element database [50]. This suggests that either very few nuclear sequences have migrated into *Spirodela* mtDNA or *Spirodela* mitochondria select against transposable elements.

Comparison of Genome Synteny

A significant degree of synteny was found within mitochondrial genomes of liverworts, mosses, and chlorophytes at the base of land plants, including a set of gene clusters (more than two genes together), such as the ribosomal protein cluster, *ccm* gene cluster, and two regions containing the *nad* and *cox* genes [51]. It was clear that the sequences of protein-coding genes were highly conserved, but the relative order of genes was greatly rearranged between *Spirodela* and rice (Fig. 2). Many ribosomal proteins were independently lost in both *Spirodela* and rice (Fig. 5); therefore, synteny between the remaining genes became harder to detect. The ancestral *cob-nad1-cox3-cox2-nad6-atp6-rps7-rps12-nad2-nad4-nad5* gene order of basal land plants has been lost due to various recombination and rearrangement events in angiosperm mtDNA evolution. [4,41,52].

In summary, our data provides further evidence that SOLiD platforms can assemble both chloroplast and mitochondrial genomes with regular coverage without any organellar purification (Table 1) [11]. Our analysis of the mt genome of *Spirodela*, having the smallest size among sequenced monocots, elucidates the evolutionary change among monocot mt genomes. Although the critical genes for the electron transport chain in *Spirodela* mtDNA are well conserved, different types of ribosomal protein genes are missing in comparison to other monocots. The number of RNA editing in protein coding genes is within a typical range as other plants. Still, no known transposable elements can be found in its genome, suggesting a rather rare migration from the nucleus to the mitochondria. Sequence-based phylogenetic analysis clearly supports the hypothesis that *Spirodela* is at the very basal lineage of monocots. Comparative analyses of mitochondrial genes between *Spirodela* and rice have shown that the relative order of genes is greatly rearranged over a very short evolutionary time. In this regard, additional complete mitochondrial sequences from closely related species will be needed to fortify the distinct evolution of plant mitochondrial genomes.

Supporting Information

Table S1 Primer pairs for gap closure and Sanger sequencing.
(XLS)

Table S2 Primer pairs for RNA editing validation by RT-PCR.
(XLS)

Table S3 Gene content for Spirodela mitochondrial genome. Gene content includes protein-coding genes, tRNA, rRNA and putative ORFs. "Ψ" means pseudo gene and "cp-" means chloroplast-derived gene.
(XLS)

Table S4 Predicted RNA editing numbers in each protein-coding gene for Spirodela mtDNA. The cutoff value for each predicted site was the percentage of matches in alignment to the corresponding amino acid across species.
(XLS)

Table S5 Type and number of codon modification in predicted RNA editing sites of Spirodela mtDNA.
(XLS)

Table S6 Comparison of codon usage between Spirodela and Oryza. [a]Results for 35 protein coding genes in Spirodela with 30,790 bp. [b]Results for all CDS from Genbank in Oryza with 44,875 bp.
(XLS)

Table S7 Predicted repeat pairs in Spirodela mtDNA by using REPuter. "F" and "P" means forward and palindromic matches.
(XLS)

Table S8 Protein-coding and tRNA gene list in the 8 representative plant mtDNAs. "+" means present and "–" means absent.
(XLS)

Table S9 The regions in Spirodela mtDNA originated from cpDNA with corresponding coordinates and identity.
(XLS)

Acknowledgments

We thank David Sidote and Mark Diamond for SOLiD library construction and sequencing. We thank Gregory Thyssen and Qinghua Wang for their invaluable comments on the manuscript. We also thank Brian Schubert from Waksman Genomics Laboratory for program installation and server management.

Author Contributions

Conceived and designed the experiments: WW YW JM. Performed the experiments: WW YW. Analyzed the data: WW YW JM. Contributed reagents/materials/analysis tools: WW. Wrote the paper: WW YW JM. Supervised the work: JM.

References

1. Logan DC (2006) The mitochondrial compartment. J Exp Bot 57: 1225–1243.
2. Mackenzie S, McIntosh L (1999) Higher plant mitochondria. Plant Cell 11: 571–586.
3. Keeling PJ, Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. Nat Rev Genet 9: 605–618.
4. Sloan DB, Alverson AJ, Storchova H, Palmer JD, Taylor DR (2010) Extensive loss of translational genes in the structurally dynamic mitochondrial genome of the angiosperm *Silene latifolia*. BMC Evol Biol 10: 274.
5. Alverson AJ, Wei X, Rice DW, Stern DB, Barry K, et al. (2010) Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). Mol Biol Evol 27: 1436–1448.
6. Palmer JD, Adams KL, Cho Y, Parkinson CL, Qiu YL, et al. (2000) Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. Proc Natl Acad Sci U S A 97: 6960–6966.
7. Fang Y, Wu H, Yang M (2012) A Complete Sequence and Transcriptomic Analysis of Date Palm (*Phoenix dactylifera* L.) Mitochondrial Genome. NCBI GenBank NC_0167401.
8. Cabrera LI, Salazar GA, Chase MW, Mayo SJ, Bogner J, et al. (2008) Phylogenetic relationships of aroids and duckweeds (Araceae) inferred from coding and noncoding plastid DNA. Am J Bot 95: 1153–1165.
9. Stomp AM (2005) The duckweeds: A valuable plant for biomanufacturing. Biotechnology Annual Review 11: 69–99.
10. Ljaz S (2010) Plant mitochondrial genome: "A sweet and safe home" for transgene. African Journal of Biotechnology 9: 4.
11. Wang W, Messing J (2011) High-throughput sequencing of three *Lemnoideae* (duckweeds) chloroplast genomes from total DNA. PLoS One 6: e24670.
12. Zhang T, Zhang X, Hu S, Yu J (2011) An efficient procedure for plant organellar genome assembly, based on whole genome data from the 454 GS FLX sequencing platform. Plant Methods 7: 38.
13. Lutz K, Wang W, Zdepski A, Michael T (2011) Isolation and analysis of high quality nuclear DNA with reduced organellar DNA for plant genome sequencing and resequencing. BMC Biotechnology 11: 54.
14. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.
15. Zerbino DR, Birney E (2008) Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18: 821–829.

16. Lohse M, Drechsel O, Bock R (2007) OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr Genet 52: 267–274.
17. Stothard P (2000) The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. Biotechniques 28: 1102, 1104.
18. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16: 276–277.
19. Mower JP (2009) The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. Nucleic Acids Res 37: W253–259.
20. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, et al. (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29: 4633–4642.
21. Bergman CM, Quesneville H (2007) Discovering and detecting transposable elements in genome sequences. Brief Bioinform 8: 382–392.
22. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24: 1596–1599.
23. Revanna KV, Chiu CC, Bierschank E, Dong Q (2011) GSV: a web-based genome synteny viewer for customized data. BMC Bioinformatics 12: 316.
24. Les DH, Crawford DJ, Landolt E, Gabel JD, Kimball RT (2002) Phylogeny and Systematics of *Lemnaceae*, the Duckweed Family. Systematic Botany 27: 221–240.
25. Notsu Y, Masood S, Nishikawa T, Kubo N, Akiduki G, et al. (2002) The complete sequence of the rice (*Oryza sativa* L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. Mol Genet Genomics 268: 434–445.
26. Clifton SW, Minx P, Fauron CM, Gibson M, Allen JO, et al. (2004) Sequence and comparative analysis of the maize NB mitochondrial genome. Plant Physiol 136: 3486–3503.
27. Malek O, Knoop V (1998) Trans-splicing group II introns in plant mitochondria: the complete set of cis-arranged homologs in ferns, fern allies, and a hornwort. RNA 4: 1599–1609.
28. Unseld M, Marienfeld JR, Brandt P, Brennicke A (1997) The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. Nat Genet 15: 57–61.
29. Chaw SM, Shih AC, Wang D, Wu YW, Liu SM, et al. (2008) The mitochondrial genome of the gymnosperm *Cycas taitungensis* contains a novel family of short interspersed elements, Bpu sequences, and abundant RNA editing sites. Mol Biol Evol 25: 603–615.
30. Zhang T, Fang Y, Wang X, Deng X, Zhang X, et al. (2012) The complete chloroplast and mitochondrial genome sequences of *Boea hygrometrica*: insights into the evolution of plant organellar genomes. PLoS One 7: e30531.
31. Giege P, Brennicke A (1999) RNA editing in *Arabidopsis* mitochondria effects 441 C to U changes in ORFs. Proc Natl Acad Sci 96: 15324–15329.
32. Takenaka M, Verbitskiy D, van der Merwe JA, Zehrmann A, Brennicke A (2008) The process of RNA editing in plant mitochondria. Mitochondrion 8: 35–46.
33. Woodson JD, Chory J (2008) Coordination of gene expression between organellar and nuclear genomes. Nat Rev Genet 9: 383–395.
34. Schneider A (2011) Mitochondrial tRNA import and its consequences for mitochondrial translation. Annu Rev Biochem 80: 1033–1053.
35. Mower JP, Sloan DB, Alverson AJ (2012) Plant Mitochondrial Genome Diversity: The Genomics Revolution In: Wendel JF, Greilhuber J, Dolezel J, Leitch IJ, editors. Plant Genome Diversity: Springer Vienna. 123–144.
36. Handa H (2003) The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (*Brassica napus* L.): comparative analysis of the mitochondrial genomes of rapeseed and *Arabidopsis thaliana*. Nucleic Acids Res 31: 5907–5916.
37. Satoh M, Kubo T, Nishizawa S, Estiati A, Itchoda N, et al. (2004) The cytoplasmic male-sterile type and normal type mitochondrial genomes of sugar beet share the same complement of genes of known function but differ in the content of expressed ORFs. Mol Genet Genomics 272: 247–256.
38. Satoh M, Kubo T, Mikami T (2006) The Owen mitochondrial genome in sugar beet (*Beta vulgaris* L.): possible mechanisms of extensive rearrangements and the origin of the mitotype-unique regions. Theor Appl Genet 113: 477–484.
39. Lilly JW, Havey MJ (2001) Small, repetitive DNAs contribute significantly to the expanded mitochondrial genome of cucumber. Genetics 159: 317–328.
40. McDermott P, Connolly V, Kavanagh TA (2008) The mitochondrial genome of a cytoplasmic male sterile line of perennial ryegrass (*Lolium perenne* L.) contains an integrated linear plasmid-like element. Theor Appl Genet 117: 459–470.
41. Alverson AJ, Zhuo S, Rice DW, Sloan DB, Palmer JD (2011) The mitochondrial genome of the legume *Vigna radiata* and the analysis of recombination across short mitochondrial repeats. PLoS One 6: e16404.
42. Janssen T, Bremer K (2004) The age of major monocot groups inferred from 800+ *rbcL* sequences. Botanical Journal of the Linnean Society: 4.
43. Stockey RA (2006) The fossil record of basal monocots. 22: 16.
44. Borsch T, Hilu KW, Quandt D, Wilde V, Neinhuis C, et al. (2003) Noncoding plastid *trnT-trnF* sequences reveal a well resolved phylogeny of basal angiosperms. J Evol Biol 16: 558–576.
45. Perrotta G, Grienenberger JM, Gualberto JM (2002) Plant mitochondrial *rps2* genes code for proteins with a C-terminal extension that is processed. Plant Mol Biol 50: 523–533.
46. Adams KL, Rosenblueth M, Qiu YL, Palmer JD (2001) Multiple losses and transfers to the nucleus of two mitochondrial succinate dehydrogenase genes during angiosperm evolution. Genetics 158: 1289–1300.
47. Adams KL, Song K, Roessler PG, Nugent JM, Doyle JL, et al. (1999) Intracellular gene transfer in action: Dual transcription and multiple silencings of nuclear and mitochondrial *cox2* genes in legumes. Proc Natl Acad Sci 96: 13863–13868.
48. Goremykin VV, Salamini F, Velasco R, Viola R (2009) Mitochondrial DNA of *Vitis vinifera* and the issue of rampant horizontal gene transfer. Mol Biol Evol 26: 99–110.
49. Knoop V, Unseld M, Marienfeld J, Brandt P, Sunkel S, et al. (1996) copia-, gypsy- and LINE-like retrotransposon fragments in the mitochondrial genome of *Arabidopsis thaliana*. Genetics 142: 579–585.
50. Smit A, Hubley R, Green P (1996–2010) RepeatMasker Open-3.0. Available: http://www.repeatmasker.org. Accessed 2012 Sep 11.
51. Terasawa K, Odahara M, Kabeya Y, Kikugawa T, Sekine Y, et al. (2007) The mitochondrial genome of the moss *Physcomitrella patens* sheds new light on mitochondrial evolution in land plants. Mol Biol Evol 24: 699–709.
52. Chang S, Yang T, Du T, Huang Y, Chen J, et al. (2011) Mitochondrial genome sequencing helps show the evolutionary mechanism of mitochondrial genome formation in *Brassica*. BMC Genomics 12: 497.

BMC Plant Biology

RESEARCH ARTICLE Open Access

Analysis of ADP-glucose pyrophosphorylase expression during turion formation induced by abscisic acid in *Spirodela polyrhiza* (greater duckweed)

Wenqin Wang[1,2] and Joachim Messing[1*]

Abstract

Background: Aquatic plants differ in their development from terrestrial plants in their morphology and physiology, but little is known about the molecular basis of the major phases of their life cycle. Interestingly, in place of seeds of terrestrial plants their dormant phase is represented by turions, which circumvents sexual reproduction. However, like seeds turions provide energy storage for starting the next growing season.

Results: To begin a characterization of the transition from the growth to the dormant phase we used abscisic acid (ABA), a plant hormone, to induce controlled turion formation in *Spirodela polyrhiza* and investigated their differentiation from fronds, representing their growth phase, into turions with respect to morphological, ultra-structural characteristics, and starch content. Turions were rich in anthocyanin pigmentation and had a density that submerged them to the bottom of liquid medium. Transmission electron microscopy (TEM) of turions showed in comparison to fronds shrunken vacuoles, smaller intercellular space, and abundant starch granules surrounded by thylakoid membranes. Turions accumulated more than 60% starch in dry mass after two weeks of ABA treatment. To further understand the mechanism of the developmental switch from fronds to turions, we cloned and sequenced the genes of three large-subunit ADP-glucose pyrophosphorylases (*APLs*). All three putative protein and exon sequences were conserved, but the corresponding genomic sequences were extremely variable mainly due to the invasion of miniature inverted-repeat transposable elements (MITEs) into introns. A molecular three-dimensional model of the SpAPLs was consistent with their regulatory mechanism in the interaction with the substrate (ATP) and allosteric activator (3-PGA) to permit conformational changes of its structure. Gene expression analysis revealed that each gene was associated with distinct temporal expression during turion formation. *APL2* and *APL3* were highly expressed in earlier stages of turion development, while *APL1* expression was reduced throughout turion development.

Conclusions: These results suggest that the differential expression of *APLs* could be used to enhance energy flow from photosynthesis to storage of carbon in aquatic plants, making duckweeds a useful alternative biofuel feedstock.

Keywords: Duckweed, *Spirodela*, Starch, Turion, ADP-glucose pyrophosphorylase

Background

Duckweed is an aquatic plant seen on water surfaces in many locations in the world. Because it consists mainly of a leaf-like body that performs photosynthesis, it is probably the most efficient multicellular biological solar energy converter that we have. Its structure during this stage of the life cycle is referred to as fronds. Greater duckweed or *Spirodela polyrhiza* is extremely simple with only one frond (merging leaf and stem) and some roots into a compact structure. Fronds grow vegetatively and can increase biomass rapidly, lowering carbon dioxide in the air and reducing nitrogen and phosphor in the water [1]. Many species of duckweeds can double their biomass every 2 or 3 days [2,3]. In addition, the tiny and

* Correspondence: messing@waksman.rutgers.edu
[1]Waksman Institute of Microbiology, Rutgers University, 190 Frelinghuysen Road, Piscataway, NJ 08854, USA
Full list of author information is available at the end of the article

free-floating duckweeds need very little amount of lignin to support their growth [4]. On the contrary, they might save the extra energy to synthesize more protein and carbohydrate. *Spirodela polyrhiza* has low amount of lignin [4], which contains 29.1% of protein [2] and up to 70% carbohydrate in dry weight [5]. The relatively easy harvesting process compared to algae is to skim of the floating fronds by net or collect them at the outlet of water by a grid [5].

There are conditions like temperature shifts due to seasons that can cause a morphological change to a different structure, called turions. Many species of the subfamily *Lemnoideae* can produce this kind of dormant fronds, which are characterized by more starch, smaller vacuoles and air space [6,7]. This developmental change is also accompanied by a shift in metabolism. The energy harvested during photosynthesis is shifted to starch biosynthesis, resulting in the accumulation of starch in turions. Because the volume of intercellular air space shrinks and starch increases the density of the tissue, it can sink to the bottom of waters where the organism can survive even if the top of the water freezes. Turions can change back to fronds vegetatively using the starch as an energy source, demonstrating a highly evolved adaptation to the environment. Because fronds have little lignin, which would interfer with the digestion of the carbohydrate fraction of biomass, and turions have high starch content, duckweed might also be suitable as an alternative source of bioenergy. Whereas cellulose is a crystalline, compact and structural compound resistant to biological attack and enzymatic degradation, starch is readily digested. Even though many advances over the past years have been made in the commercialization of cellulosic biomass [8], the cost of producing equal amounts of ethanol from cellulosic biomass is still much higher than production directly from starch [9]. Therefore, growing attention is being devoted to use duckweeds as a source of carbon compounds and convert duckweed biomass into bio-ethanol [10]. Fronds growing in swine wastewater contain 45.8% (dry weight) of starch. Moreover, 50.9% of the original dry biomass can be enzymatically hydrolyzed into a reducing sugar, which contributes to 25.8% fermented ethanol of dry biomass [10].

Recent studies have focused on the influence of various environmental conditions for turion formation or germination [11-15], the sensitivity threshold of ABA for turion formation [13,16] and the different structure (air space, vacuole, starch and cell wall) of fronds and turions [7]. On the other hand, information of starch content, granule size, and derivation of starch granules involvement with turion formation, which is critical to explore the potential biofuel of duckweed, is less well understood.

The pathway of starch synthesis is very complex, but ADP-glucose pyrophosphorylase (AGPase) plays a pivotal role in regulating starch levels and in determining patterns of starch deposition in plants. This enzyme comprises two identical large subunits (APLs) and two same small subunits (APSs) in angiosperms, each of which is encoded by distinct genes. Even though the roles of each AGPase subunit in the enzyme are not clear, it is generally proposed that APLs modify the response to allosteric regulators, whereas APSs act as the catalytic part [17]. Recent studies suggest that AGPase are usually in plastidial forms except for a cytosolic one in cereal endosperms [18,19]. Here, we compared the distinctive attributes between fronds and turions in *S. polyrhiza* and investigated starch production during development upon induction with abscisic acid (ABA), a plant hormone. To gain further insight into the function of the large subunit of AGPase (*APLs*) in starch synthesis as well, we cloned the *Spirodela* genes, analyzed them, and quantified their expression, which will allow in the future targeting expression of transgenes.

Results

Turion induction with ABA

Spirodela polyrhiza was grown under controlled light conditions as described under Methods. Fronds were harvested and examined under a dissecting microscope. Dividing fronds, representing single leaf-like bodies, were connected, thin, and elliptical (~8 mm in length and ~6 mm in width). The top of fronds was bright green, whereas the bottom extended a few roots that were submerged into water (Figure 1a). Continued growth in the presence of ABA gave rise to turion formation with different morphological features (Figure 1b). After 5 days of ABA application a significant shift to starch accumulation took place in samples collected from both wet and dry tissues. Starch accumulation during turion development exhibited a characteristic pattern. There was a progressive increase of starch from 5 days to 10 days after ABA application and after 14 days, the starch content became almost stable. The final starch content in turions for wet tissues was 24.4%, which corresponds to 60.1% in dry mass (Figure 2a). Turions were also harvested and examined under a dissecting microscope. They appeared thicker and smaller in nearly round shape (~2 mm in length and ~3 mm in width). Turions were dark green, spotted with many anthocyanin pigments, and retained only rudimentary roots that are not visible by naked eye (Figure 1b).

Frond samples were then examined by electron microcopy. The frond cell had normal discal chloroplasts with a few small starch grains (Figure 3a and 3c). Most frond cells contained a single larger vacuole and bigger intercellular air space, while turion cells have multiple smaller vacuoles and bigger air space between cells (Figure 3b). The turion cell accumulated many starch granules, which

Figure 1 Morphological comparison of frond and turion formed after 14 days of ABA treatment. **a**) dorsal and ventral frond; **b**) dorsal and ventral turion. Bars = 1 mm.

almost occupied 1/4-2/3 of cell volume (Figure 3b and 3d). The kidney-shaped starch granule was surrounded with stacks of thylakoid membranes in chloroplasts (Figure 3e and 3f). The increased starch granules at the expense of the vacuolar expansion also contributed to the distortion of chloroplasts (Figure 3e) and a shift in tissue density that caused turions to sink to the bottom of liquid medium (left panel of Figure 4a). Placed on filter paper, they looked like "green seeds" compared to fronds (Figure 4b).

Cloning and sequencing of members of the *Spirodela* APL gene family

The level of starch accumulation in turions (Figure 2) and the convenience of collecting them from the flask bottom (Figure 4) are key features for biofuel applications as described above. To examine the metabolic regulation of these features, this study seeks to identify key enzymes, whose manipulation at the molecular level could optimize the timing and level of starch production. Common knowledge would then suggest investigating the differential expression of key enzymes in starch biosynthesis. Therefore, we decided to clone the large subunit of the ADP-glucose pyrophosphorylase gene family (*APLs*) from *Spirodela polyrhiza*. Because this gene is very conserved among angiosperms, we used the known *Arabidopsis* protein sequences to design degenerate primers to amplify *APL* coding sequences as described under Methods. Cloned DNA fragments were then sequenced and overlapping fragments were used to reconstruct the entire three cDNA-copies from *Spirodela*. We named them *SpAPL1*, *SpAPL2* and *SpAPL3* with Genbank accession numbers of JN180634, JN180635, JN180636. Based on the cDNA sequences primers were then designed to clone the corresponding gene sequences from total genomic DNA as described under Methods. The cloned genes of *SpAPL1*, *SpAPL2* and *SpAPL3* were then also sequenced and deposited

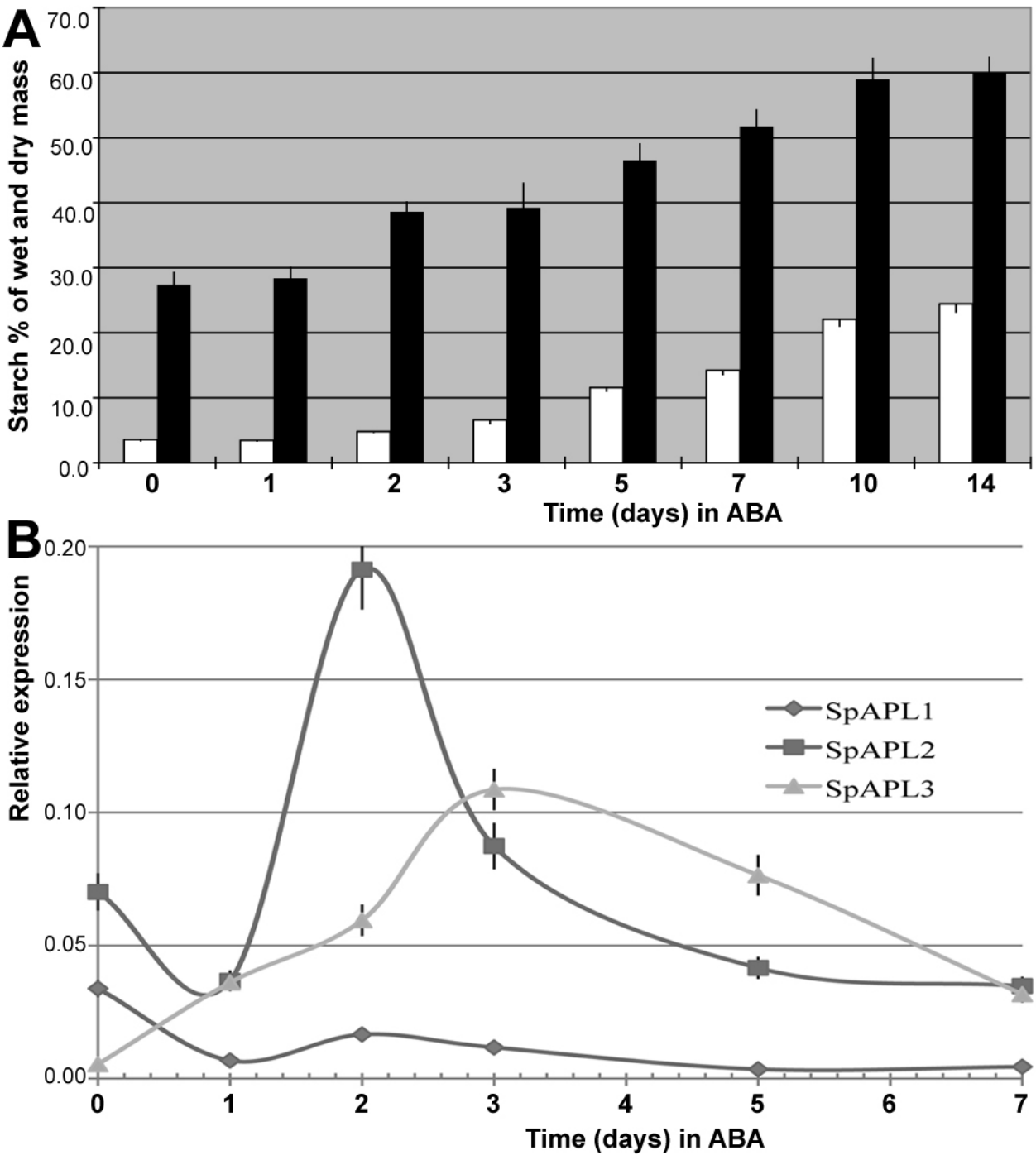


Figure 2 Starch accumulation and expression of the APL genes during turion development. **a**) White bars stands for wet tissue and black bars for dry tissue. Y-axis shows starch content (mg) for every 100 mg wet tissue or dry tissue. **b**) qPCR was used to quantify expression of APLs based on RNA from 0, 1, 2, 3, 5, 7 days of ABA treatment. Standard error was shown by vertical bar. Sample collections for starch analysis and *APLs* expression quantification were listed in Table 1.

into GenBank with accessions JN180631, JN180632, JN180633, respectively. After aligning cDNAs with their corresponding genomic sequences, all introns could be identified. Accordingly, all *SpAPLs* consisted of 15 exons and 14 introns (Figure 5). Whereas the coding sequences of the *SpAPL1*, *SpAPL2* and *SpAPL3* genes were slightly different in length with 1,554, 1,611, 1,620 bp or 517, 536, and 539 amino acids, respectively, the corresponding

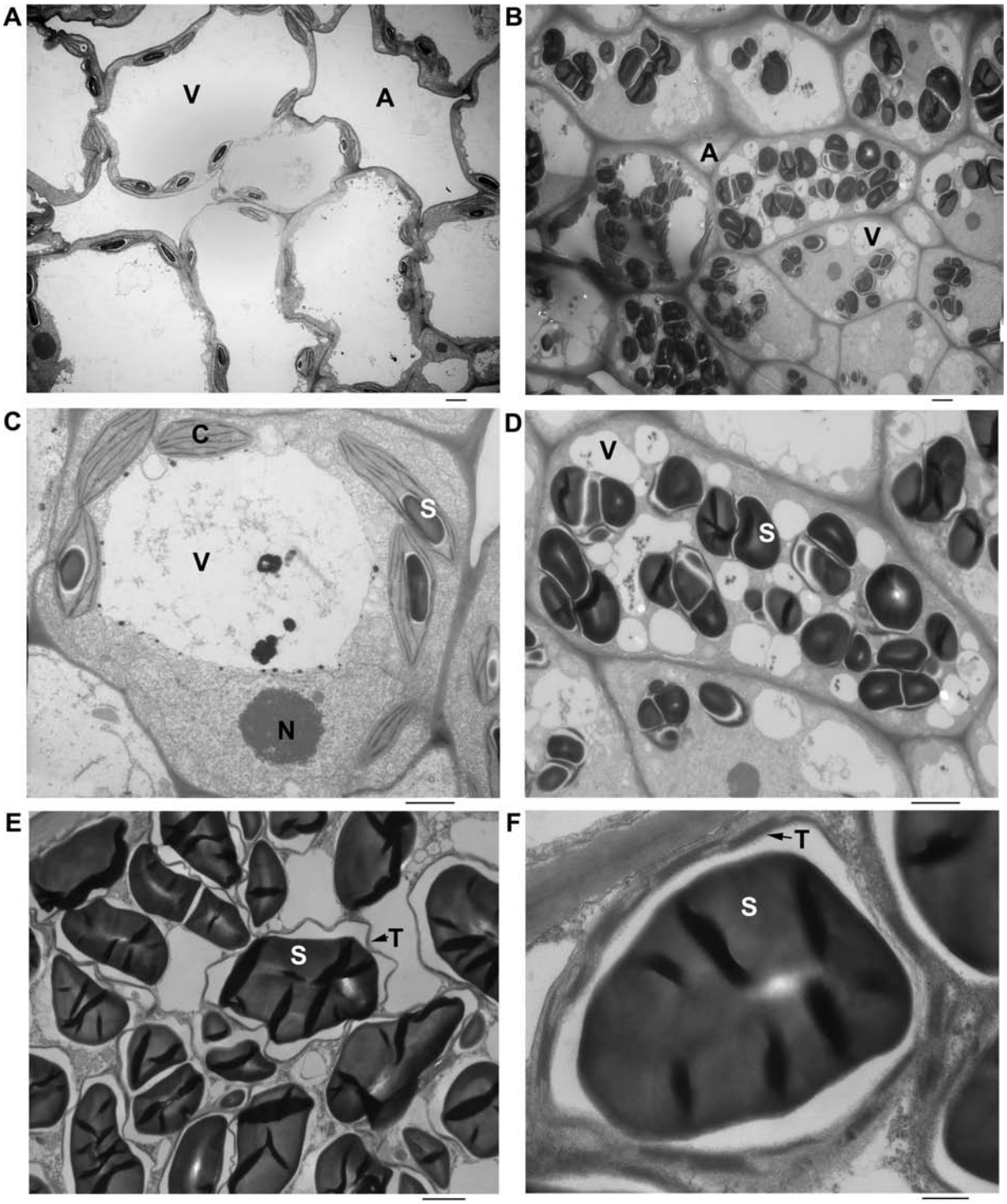


Figure 3 Microscopic study. **a**) Transmission electron microscopic (TEM) picture of frond cells with lower magnification, Bars = 2 μm; **b**) TEM picture of turion cells with lower magnification, Bar = 2 μm; **c**) TEM picture of a frond cell with higher magnification, Bar = 2 μm; **d**) TEM picture of a turion cell with higher magnification, Bar = 2 μm; **e**) TEM picture of a section of a turion cell with higher magnification, Bar = 2 μm; **f**) TEM picture of a section of a turion cell with the highest magnification, Bar = 500 nm. Abbreviations are chloroplast (C), starch granule (S), vacuole (V), intercellular air space (A), thylakoid membrane (T), nucleus (N).

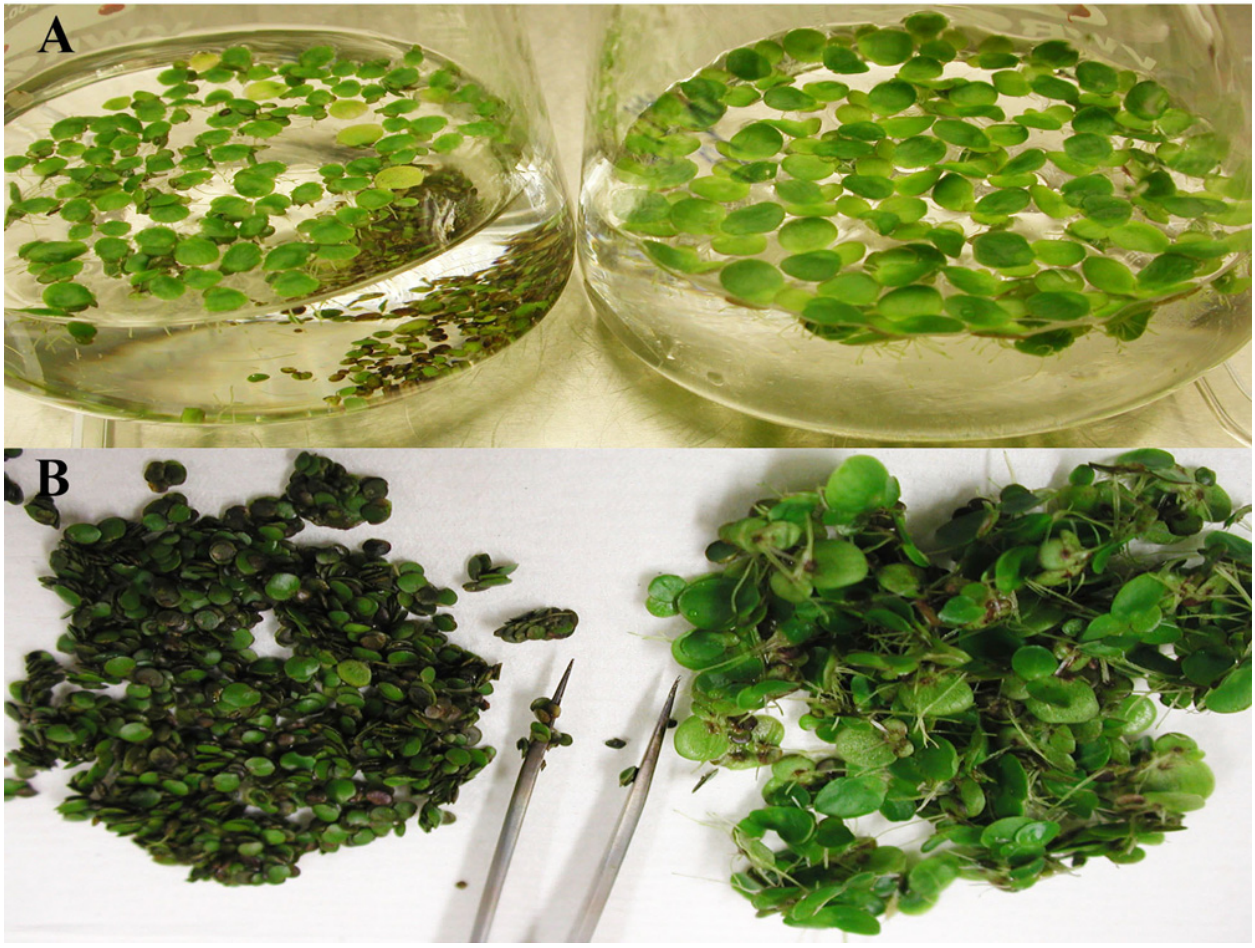


Figure 4 Turion formation induced by ABA. **a**) Turions (left panel) on the bottom and fronds swimming with roots down (right panel) in flasks; **b**) turions (left) and fronds (right) placed on filter paper. Bars = 1 mm.

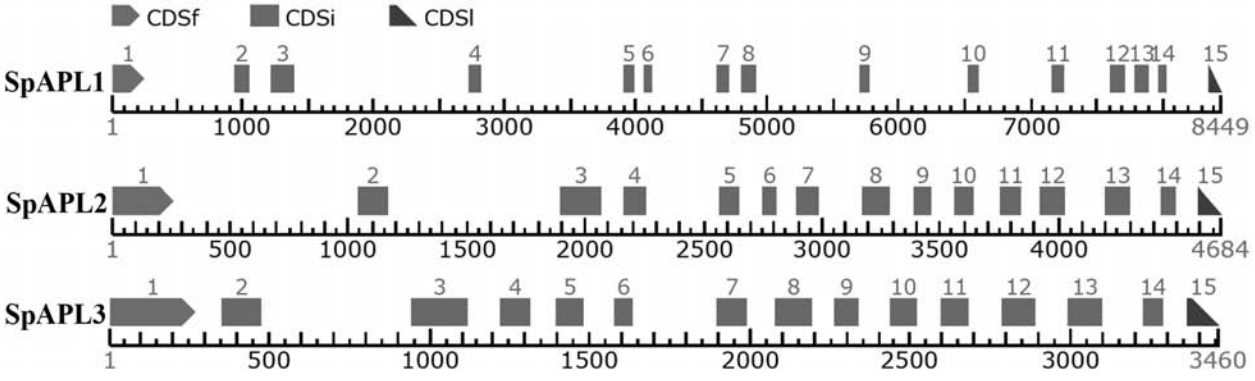


Figure 5 Structural organization of the large subunit of AGPase (*SpAPL*) genes. The coding exons were depicted as gray boxes. Introns were depicted as bar scaled by x-axis (bp). CDSf–First (Starting with Start codon), CDSi–internal (internal exon), CDSl–last coding segment (ending with stop codon).

Table 1 Gene features of *APL* family

Gene Name	Gene Length (bp)	ORF Length (bp)	Putative Protein Length (aa)	Intron Length (bp)	MITE Length (bp)	Ratio = MITE/ Intron
SpAPL1	8449	1554	517	6895	2507	0.36
SpAPL2	4684	1611	536	3073	659	0.21
SpAPL3	3460	1620	539	1840	126	0.07

genomic regions differed significantly with 8,449, 4,684 and 3,460 bp (Table 1), reflecting intron expansions.

Structure and phylogeny of members of the Spirodela APL gene family

The basis for the variation in protein sizes became clear when their primary structures were compared with other known APLs. Sequence alignments of the deduced amino acid sequences of SpAPL1, SpAPL2, and SpAPL3 proteins showed high homology except for their N-terminal regions (Additional file 1: Figure S1). APLs are usually targeted to the plastid through a signal peptide at their amino-terminus. SpAPL2 and SpAPL3 had conserved plastid-targeting signals with cleavage sites at positions 78 and 64 based on TargetP http://www.cbs.dtu.dk/services/TargetP/. The shorter protein of SpAPL1 had a very weak targeting signal and internal deletions similar to the rice APLs. Although there were differences in the amino-terminal regions, the coding sequences from exon 3 to 15 were of the same size and very conserved.

The corresponding introns, however, have diverged significantly in length and composition. Interestingly, comparison to the TIGR Plant Repeat Database [20] indicated that expansion of introns could be largely due to miniature inverted-repeat transposable elements (MITEs). When the MUST system was applied that was used to predict MITEs rather than depending on sequence homology alone, the sequence data suggested then that MITEs had invaded the introns of *SpAPL1*, *SpAPL2*, and *SpAPL3*, comprising now 36%, 21%, and 7% of total intron sequences, respectively (Table 1).

Using the amino acid sequence alignment of *APL*s from *S. polyrhiza*, rice, and maize, we constructed a maximum likelihood phylogeny of the *APL* family. This phylogenetic tree separated the *APL*s into three main clades: *SpAPL1* clustered together with the plastidial forms of *OsAPL1* and *ZmAPL1* in branch APL-I. *SpAPL2* shares the branch APL-II with the plastidial forms of *OsAPL4* and *ZmAPL4*; *SpAPL3* shares a common ancestor with both plastidial (*OsAPL3* and *ZmAGP1*) and cytosolic forms (*OsAPL2* and *ZmSH2*) in rice and maize [21] (Figure 6).

A structural model of the APLs

To confirm the inference of their function, three-dimensional structures of SpAPLs were built by using the experimental protein structure (PDB 1yp3) from potato as a suitable template. Amino acid sequence alignment of the regulatory site of APLs from potato and *S. polyrhiza* showed five key conserved residues (P44, P52, P66, K414 and K452) (Figure 7a) in all three SpAPLs. Molecular modeling analysis of APLs suggested a critical role of APLs for allosteric regulation in this region with binding sites for ATP and 3-PGA (Figure 7b). P44 was important for accommodating ATP phosphate groups, as it was located between a conserved GGXGXRL loop region and the strongly conserved "PAV" region, which involved catalysis and allosteric regulation [22]. P52 was predicted to be located in flexible loops close to the lysine residues (K414 and K452), while P66 lied in a helix. Site-directed mutagenesis of the P52 and P66 in potato showed dramatic changes in affecting enzyme regulatory properties, while P44 mutants resulted in a nearly catalytically inactive enzyme [23]. K414 and K452 were shown to be involved in the increase of the affinity for the activator 3-PGA [22,24]. Model structures of APL1, APL2 and APL3 were identical in these features. Therefore, only APL1 was shown in Figure 7 as an example.

Expression patterns of APL genes in developing turions

With three different gene copies present in *Spirodela polyrhiza*, the question arises how each enzyme is expressed temporally during turion formation. We therefore isolated total mRNA from leaf-like tissue 0, 1, 2, 3, 5, and 7 days after the addition of ABA. To measure expression of each APL gene copy, we applied qPCR to mRNA samples using specific primer pairs to distinguish between transcripts from each gene. Expression of *SpAPL2* and *SpAPL3* dramatically increased two-and 10-fold, respectively, as turion development was initiated (1-3 days). Furthermore, there seemed to be a difference in the expression of *SpAPL2* and *SpAPL3*. *SpAPL2* was critically in the first phase of induction, whereas *SpAPL3* seemed to replace *SpAPL2* in a second burst of activity. There was no obvious increased expression of *SpAPL1* after ABA induction. Indeed, *SpAPL1* appeared to be more active in initial fronds compared to *SpAPL3* (0 days of ABA application). When turions went into mature phase (after 5 days), the expression of all *SpAPLs* was leveling off (Figure 2b).

Discussion

We began dissecting the process of turion formation in duckweeds. Usually turion development occurs in late

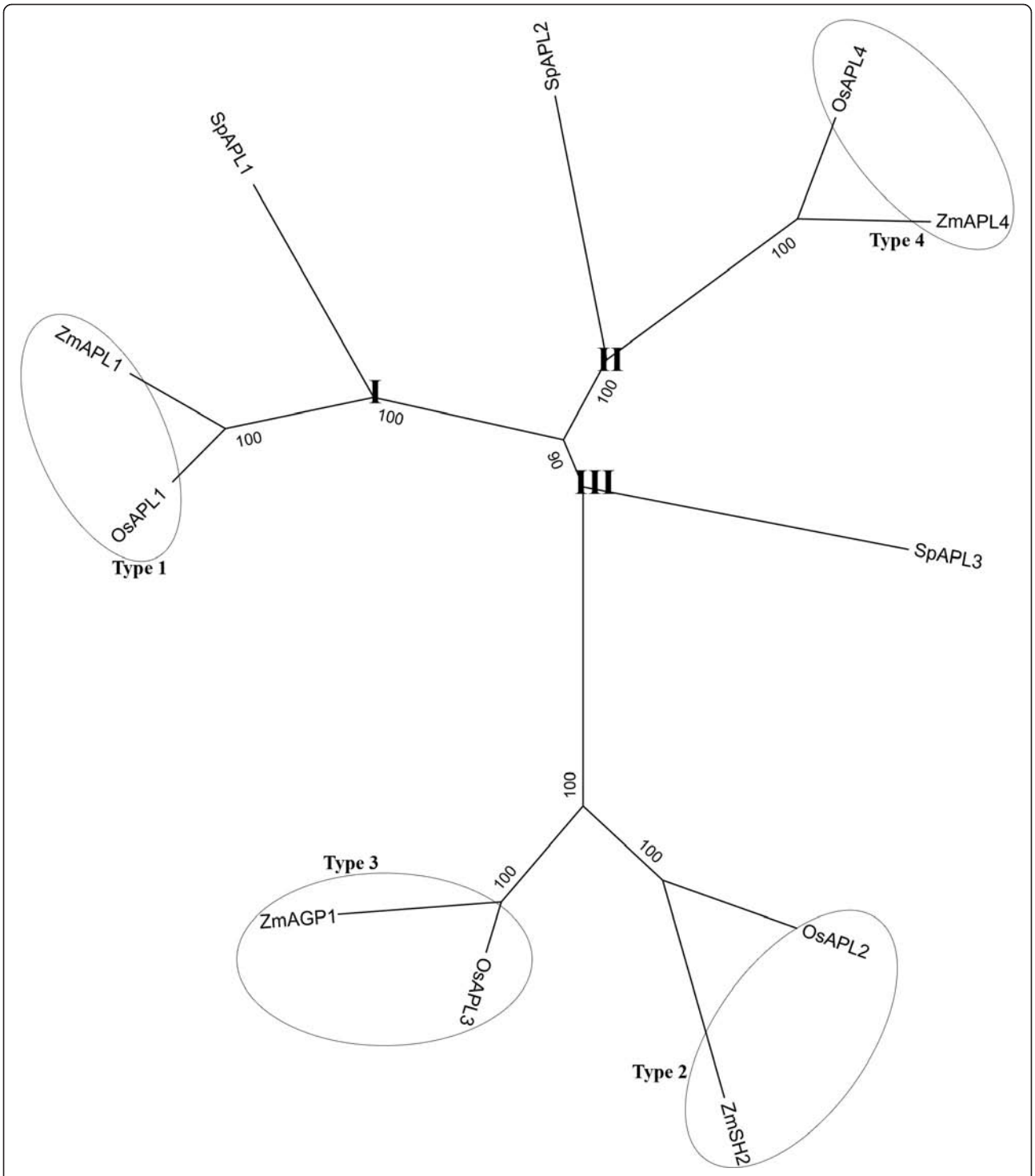


Figure 6 Phylogenetic tree based on the amino acid sequence alignment of large subunits of AGPase (*APLs*) from *S. polyrhiza* (Sp), *Oryza sativa* (Os), and *Zea mays* (Zm). The protein names are those published previously or predicted from CDS: OsAPL1, NP_001051184; ZmAPL1, NP_001106017; SpAPL1, JN180634; OsAPL2, NP_001043654; ZmSH2, NP_001121104; OsAPL3, NP_001056424; ZmAGP1, NP_001105717; SpAPL3, JN180636; OsAPL4, NP_001059276; ZmAPL4, NP_001106058; SpAPL2, JN180635. Three clades were designated APL-I, APL-II, APL-III. The classification of APLs in grasses has previously been published [21].

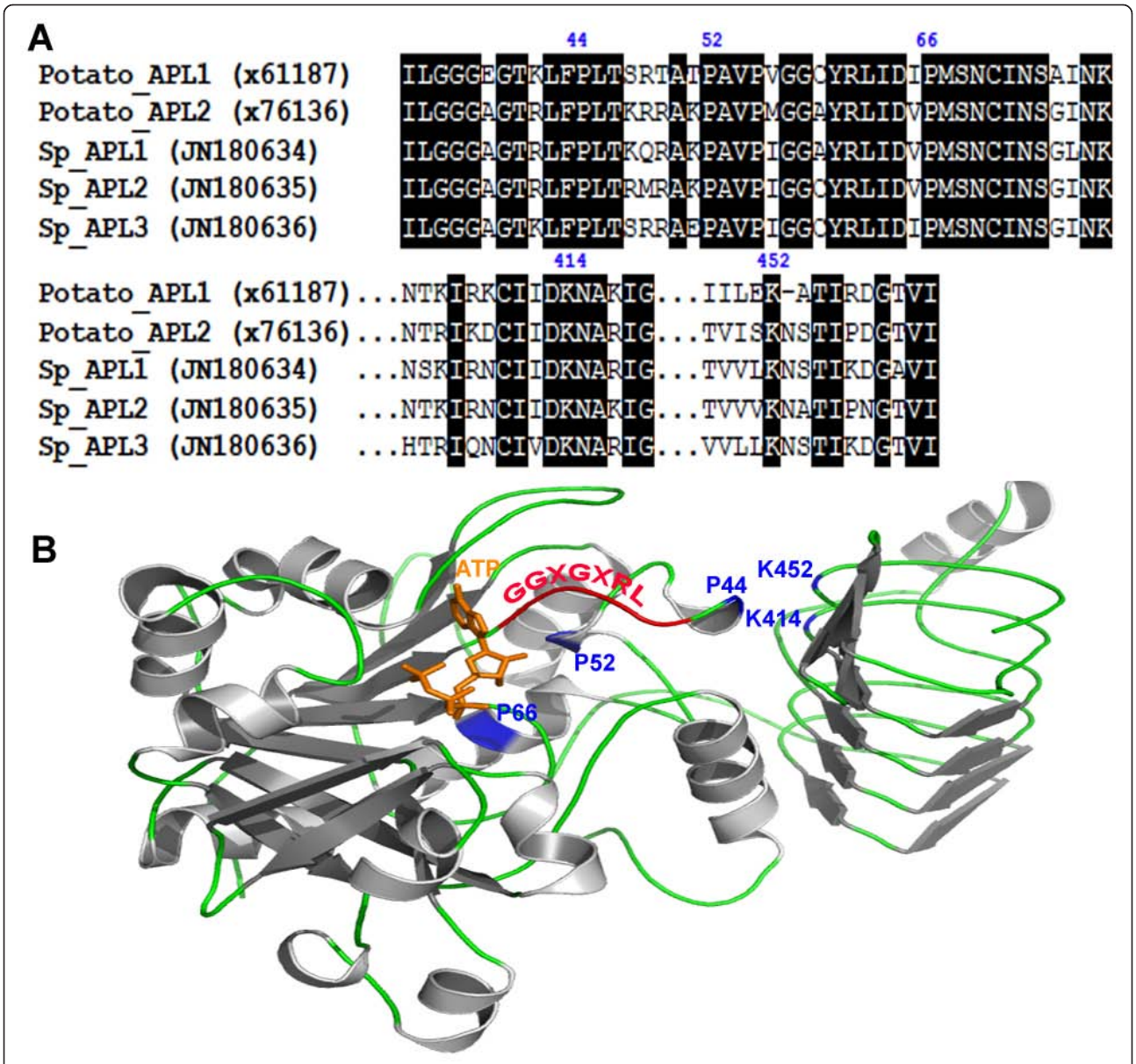


Figure 7 Amino acid sequence alignment for the regulatory sites of APLs between potato and S. polyrhiza and modeled structure of S. polyrhiza APL1. **a**) GenBank accession numbers were listed in parentheses. Important proline (P44, P52 and P66) and lysine (K414 and K452) residues critical for allosteric regulation were numbered corresponding to potato AGPase large subunit (x61187). Identical residues were shaded in black. **b**) The N-terminal region containing the putative ATP binding site and the regions containing the putative 3-PGA binding sites of APL1 were modeled by comparison with the known structure of the potato AGPase small subunit (PDB 1yp3) with 54.93% identity. The modeled position of ATP in orange was shown. The α-helix and β-sheet were colored in gray and the loop was in green. Important proline (P) and lysine (K) residues in APL1 were indicated by blue color. The conserved GGXGXRL loop region was in red.

summer or early autumn because of starvation and lower temperatures [25]. *Spirodela* turions can also be induced under controlled laboratory conditions by increasing the concentration of ABA in the growth medium [11,13], decreasing temperature [15], or depriving phosphorus in the medium [12]. Here, we have taken advantage of ABA as an inducer and could reproduce the morphological changes that occur during turion formation. Turions are germinated into new fronds in the presence of light and nitrogen in the following spring using starch storage as an energy source [26,27]. Therefore, the drastic starch accumulation during turion formation marks a turning point in the switch process from low-starch fronds to high-starch turions.

The reported contents of starch varied from 14% to 43% depending on the species, developmental states

(fronds, resting fronds, or turions) [28] and tested methods [5,29]. Starch content could even go up to 75% of the dry weight in resting fronds of *Spirodela oligorrhiza* (renamed into *Landoltia punctata*) growing in phosphor-deficient cultures [12], a level that is comparable to cereal seeds of corn, sorghum and wheat [30]. Even though regular fronds have as low as 16% starch in dry mass, turions of *S. polyrhiza* can reach up to 62% starch [25]. Our use of exogenous ABA produces the same developmental switch, as the different morphological features are easily distinguishable. The switch is rapid, providing advantages for biochemical and physiological analysis [13]. We obtained 60.1% starch from dry mass after 2 weeks of ABA induction (Figure 2a), which is comparable to the Henssen's study. The size of mature starch grains from turions was around 4 μm in diameter as estimated by TEM (Figure 3e and 3f), whereas starch grains from wheat, corn and rice reach a size of 30 μm, 25 μm and 20 μm, respectively [31]. In a different study, the size of starch granules illuminated by red light for different times have also been measured using SEM scans arriving at similar values [32]. Interestingly, it has been suggested that smaller starch granules are more easily hydrolyzed into sugars than larger ones, regardless of botanical source [33]. After 72 h of continuous irradiation, the sizes of starch granules in turions are significantly reduced to about 1.5 μm [32]. Although duckweeds might have adapted to rapidly switch back to a growth phase faster than seed plants, this property also might provide a more efficient way for producing bio-ethanol than from maize.

Amyloplasts in non-photosynthetic tissue, such as seeds, roots, and stems, which lack chlorophyll and internal membranes, are the main organelles responsible for the synthesis and storage of starch granules in most plants. However, turions remain green or dark-green throughout their development (Figure 1b and 4b). The plastids in turions, where starch synthesis takes place, still retain abundant stacks of thylakoids (Figure 3e and 3f). It therefore suggests that chloroplasts with a simple structure as in duckweeds can function both as source and sink. The starch-storing plastids of turions are directly derived from chloroplasts, and retain chloroplast-like characteristics throughout their development. This adaptation greatly saves energy by directly depositing sucrose generated from photosynthesis into starch storage without the need for transport through a vascular system and the use of a glucose phosphate transporter [21]. A similar system exists also in a non-aquatic plants such as pea embryos, where starch-storing plastids also directly originate from chloroplast [34,35]. Moreover, using TEM light-induced degradation of starch granules in turions of *Spirodela polyrhiza* also exhibited a transition from amyloplasts to chloroplasts [32]. Both studies would demonstrate that differentiation from chloroplast to amyloplast could be reversed based on physiological changes. Indeed, the cell structure of turions appears to be well organized for its function. Its lack of intercellular air space and presence of smaller vacuoles allow them to survive in deep water, where the temperature is more moderate than on the surface. The numerous starch grains provide a bank of energy when turions germinate in the following spring. This life cycle is also consistent with starch content in fronds and turions.

Because starch biosynthesis is an important feature for the developmental switch from fronds to turions, it also provides us with the first entry point to dissect the developmental regulation of turion formation. Therefore, we reasoned that the first step in this line of investigation consists of the identification and characterization of key regulatory genes known in starch biosynthesis, which are the ADP-glucose pyrophosphorylases. We successfully cloned three copies of *APLs* of *Spirodela polyrhiza*. *APLs* are expressed in different organs of grass species, type 1 in leaves, type 2 and type 3 in seeds, and type 4 in both seeds and leaves [21,36]. Based on phylogeny and spatial expression of *SpAPLs* (Figure 6 and 2b), they have their homologs in grass species. *SpAPL2* and *SpAPL3* are active in turions, while *SpAPL1* is expressed at a higher level in fronds. The transcript level of *SpAPL2* and *SpAPL3* are active at an early phase of turion formation, while all transcript level of *SpAPLs* decline towards the end phase. It could account for the inhibition of total RNA synthesis after 3 days in ABA, which leads to the shutdown of all primary processes and onset of the dormant state [37]. Analysis of networks of gene expression during *Arabidopsis* seed filling has also shown that expression of carbohydrates occurred early in seed development [38]. Noticeably, the transcription of *SpAPL1* and *SpAPL2* is suppressed right after one day of ABA addition, which is quite consistent with previous findings that ABA could inhibit DNA, protein, and RNA synthesis during turion development [37]. But this inhibitory effect of ABA during turion development is selective for that the synthesis of certain turion specific proteins increases [37]. Indeed, the pattern of expression was consistent with a rate-limiting role for this protein in starch biosynthesis. Furthermore, the regulation of gene copies underwent divergence and probably sub-functionalization to permit metabolic differentiation.

In plants, ADP-glucose pyrophosphorylases consist of large and small subunits that share many amino acids due to the proposed origination from a common ancestral gene [39]. For example, APLs and APSs, which make up the heterotetrametic potato enzyme, share significant sequence homology (53% identity and 73% similarity) [40]. Here we selected the large subunit for our analysis because we made the assumption that both are coordinately expressed and that the large subunit should suffice

as a marker of the developmental switch between frond and turion stage of the life cycle. Furthermore, the current sequencing of the entire genome will provide an opportunity to locate the gene copies of the small subunit as well. The model structure of the large subunit confirms that N-and C-terminal regions of the SpAPLs are essential for the allosteric regulatory properties of the heterotetrameric enzyme AGPase (Figure 7b) [23]. Even though APLs are considered as a catalytic-disabled subunit, the ability of binding effectors (3-PGA) and substrates (ATP) is likely to undergo a conformational transition similar to the APSs during its catalytic cycle [41].

Phylogenetic analysis showed that *SpAPL1* and *SpAPL2* descended from common ancestors of the plastidial form Type 1 and Type 4 of the grasses, respectively, while *SpAPL3* shares the same branch with the ancestor of cytosolic Type 2 and plastidial Type 3 of grasses (Figure 6) [17]. Studies suggest that cytosolic Type 2 in grass evolved from a duplication of an ancestral gene encoding a Type 3 plastidial *APL* by loss-of-function of the transit peptide cleavage site [21]. A similar process might have taken place in *Spirodela*, where *SpAPL1* does not exhibit a clear transit peptide. Interestingly, the opposite seems to be true for *SpAPL3*, which clusters with cytosolic Type 2 *APL*s, but encodes a transit peptide. Based on this, we classify it as a plastidial Type 3 *APL* of the grasses. The phylogenetic relationship will become clearer when we know whether these copies are clustered or dispersed in the *Spirodela* genome. Interestingly, there is differential invasion of MITES in the introns of these genes with the most pronounced invasion in the *SpAPL1* gene (Table 1). This is reminiscent of the grasses, where one of the smallest genomes, rice, had a relative high percentage of MITEs (13.3% of all repeat elements compared to 0.4% in maize), but low retrotransposon content (59.5% compare to 92.7% in maize). *Spirodela polyrhiza* was namely chosen for sequencing because of its small genome size. Given the genome size variation among *Lemnoideae*, perhaps a similar relationship of genome size and MITEs exists among *Lemnoideae* as has been found in grass species [42].

Conclusion

In summary, turions of *S. polyrhiza* contain high starch content, small size of starch granules, and low lignin proportion, which provides a solid foundation for developing them as an alternative biofuel source. For further investigation of the role of *SpAPL2* and *SpAPL3* genes in starch synthesis, studies using transgenic plants will be needed.

Methods

Plant material and growth conditions

For our studies we chose *S. polyrhiza* (Sp) 7498 because this will serve as a reference genome for the *Lemnoideae*. One cluster of 3-5 fronds was aseptically transplanted into half-strength Schenk and Hildebrandt basal salt mixture (Sigma, S6765) with 1% sucrose liquid medium at pH 5.8. The cultures were kept in chamber maintained at 100 $\mu mol.m^{-2}.s^{-1}$ and 23°C through a 16 h-light, 8 h-dark photoperiod. After a couple of days' growth, 1 μM abscisic acid (ABA, Sigma, A1049) was added.

Microscopic analysis of frond and turion

Vegetative fronds without ABA treatment and turions with 14 days ABA treatment were fixed, embedded, and dehydrated as described [43]. Samples were fixed in 5% glutaraldehyde in 0.1 M sodium cacodylate buffer, pH 7.4, containing 2% Suc in a 2-ml tube at 4°C overnight and another 3 h at room temperature. Rinsed by 0.1 M sodium cacodylate buffer, they were postfixed in buffered 1% osmium tetroxide at 4°C overnight followed by dehydration in a graded series of acetone washings. The dehydrated samples were then embedded in epon resin. The 1 mm-thick sections were picked up on a glass slide, stained with methylene blue and scoped with a light microscope. For transmission electron microscopy (TEM), 90 nm-thin sections were cut on a Leica EM UC6 ultramicrotome, stained with a saturated solution of uranyl acetate and lead citrate and scoped at 80 kV with a Philips CM 12 transmission electron microscope.

Determination of starch content of developing turions

One hundred milligrams of fresh sample tissues were taken from a time course of 0 (no ABA), 1, 2, 3, 5, 7, 10, 14 days of ABA treatment and flash frozen in liquid nitrogen. Before 7 days, the whole plants including both mother and daughter fronds were collected. After 7 days, the developed turions were separated from mother fronds and collected, when they sunk to the bottom of flask (Table 2). Three biological replicates were done for each time point. The quantification of starch content was determined colorimetrically following manufacturer's protocols of a "total starch assay" procedure (amyloglucosidase/α-amylase method) (Megazyme, K-TSTA). We used water as a blank control and D-glucose as a standard. Dry weight was counted by 500 mg fresh tissue after incubation in 65°C chamber for 24 h.

Genomic DNA and total RNA isolation

Total genomic DNA was extracted from whole plant tissue by the CTAB method [44]. Considering that only daughter fronds shorter than 0.7 mm in length respond to ABA treatment and undergo turion formation after ABA treatment [16], developing turions only with specific sizes were collected at their developmental stages after 0 (no ABA addition), 1, 2, 3, 5, 7 days of ABA treatment, respectively, for quantification of *APL* gene expression (Table 2). For each time point we used again three biological replicates. High-quality total RNA was extracted with RNeasy

Table 2 Sample collection for starch analysis and *APLs* expression quantification

Days in ABA	Samples for Testing Starch Content	Size of fronds or turions (mm) for Analysis *APLs* Expression	Characterization of Developing Turions
0	Whole plants	~ 0.5-0.7	Light green
1	Whole plants	~ 1	Light green
2	Whole plants	~ 1.5	Light green
3	Whole plants	~ 2	Dark green
5	Whole plants	~ 2	Dark green
7	Only turions	~ 2	Dark green, sink at the bottom
10	Only turions	no collection	Dark green, sink at the bottom
14	Only turions	no collection	Dark green, sink at the bottom

Plant Mini Kit (Qiagen, 74904). The on-column DNase I was used to remove contaminating genomic DNA (Qiagen, 79254). The RNA quality and quantity were confirmed by analysis with Nanodrop 1000 (Nanodrop Technologies, Wilmington, DE). First-strand cDNA synthesis of all samples was generated by kit of SuperScript™ III First-Strand Synthesis System for RT-PCR (Invitrogen, 18080) using oligo-dT as primer.

Retrieval of APL genes and CDS sequence

The conserved domains of APL proteins of *Arabidopsis* were used to set up degenerate primers. Degenerate PCR reactions were done with templates of cDNA extracted from samples of 3 days of ABA treatment. The program was: 35 cycles of 94°C 30 s, 50°C 30s and 72°C 1 min. PCR products were cloned into the pGEM-T Easy Vector (Promega) and DNA fragment sequences were determined using the ABI 3730XL platform. Gene specific primers were designed based on the sequence of the cloned DNA to perform 5' and 3' RACE using the SMARTer™ RACE cDNA Amplification Kit (Clontech, 634923). The RACE-ready cDNA was also generated from total RNA of samples treated 3 days with ABA. RACE reactions were performed under the following program: 5 cycles of 94°C 30 s and 72°C 2 min; 5 cycles of 94°C 30 s, 70°C 30s and 72°C 2 min; 25 cycles of 94°C 30 s, 68°C 30 s and 72°C 2 min. The RACE products were also cloned and sequenced. The full-length cDNA was confirmed with primers designed from 5' end of the 5' RACE sequence and the 3' end of the 3' RACE sequence. The same primers were used to amplify corresponding gene sequences using genomic DNA as template. Because of the size of the genes we used Expand Long Range dNTPack (Roche, #04829042001). The thermal cycling conditions were: 10 cycles of 94°C 15 s, 55°C 30 s and 68°C 9 min; 25 cycles of 94°C 15 s, 55°C 30 s and 68°C 9 min with 10 more seconds for each cycle. Initially, primer sequences derived from APL cDNA were used to sequence genomic DNA. Subsequently, primers derived from genomic sequences were used in iterative rounds of sequencing until sufficient coverage was achieved. The sequences were assembled and analyzed with DNASTAR. MUST system, which tested the existence of a pair of terminal inverted repeats (TIRs) and a pair of direct repeats (DRs) [45] was used to predict miniature inverted-repeat transposable elements (MITEs) in *APL* introns. All successful primers were listed in Additional file 2: Table S1.

Phylogenetic studies

An unrooted maximum likelihood phylogenetic tree was determined by using the MEGA 5 program [46] based on the amino acid sequence alignments under the WAG model with 1000 bootstrap replications. The corresponding subunit sequences from rice and maize were downloaded from GenBank.

Modeling of the three-dimensional structures

Sequences of the APL regulatory sites from potato and *S. polyrhiza* were aligned using MEGA 5. Homology modeling studies were performed using the Swiss Model server (http://swissmodel.expasy.org/) [47] and structures were visualized and prepared by an open source program PyMOL (The PyMOL Molecular Graphics System, Version 0_99rc6, Schrödinger, LLC.). The sequence used was that of SpAPL1, SpAPL2 and SpAPL3. The chosen suitable template was homodimeric AGPase of potato (PDB 1yp3) [22] for which X-ray structure information was available, showing more than 52% sequence homology with SpAPLs. Key proline (P44, P52 and P66) and lysine (K414 and K452) residues were numbered based on AGPase large subunit of potato (x61187) [23]. Only APL1 modeled structure was shown in Figure 7b as representative for the sake of simplicity.

Expression analysis of APL genes

Alignment of full length of cDNAs produced unique regions at the 5' UTR to design primers for qPCR (Additional file 2: Table S1). qPCR was performed for 0, 1, 2, 3, 5, 7 day of ABA treatment. All cDNAs were made with 2 μg of RNA using the SuperScript® III First-Strand Synthesis System kit (Invitrogen, 18080-051). cDNAs were diluted 20-fold and Real-time PCR

was performed by using the iQTM SYBR Green Supermix (Biorad, 170-8880) following the manufacturer's standard instructions. All qPCRs were performed in triplicates. The relative quantification of each gene expressional level was calculated by calibrating CT values normalized to a standard dilution series over all samples assayed [48].

Additional material

Additional file 1: Figure S1. Multiple alignments of the deduced amino acid sequences of APL proteins from S. polyrhiza (Sp) and Oryza sativa (Os). Dashed lines indicate gaps introduced to maximize alignment.

Additional file 2: Table S1. Primers for cloning, sequencing and quantifying APLs expressions. *Primers were cited from [49].

Acknowledgements
The research described in this manuscript was supported by the Selman A. Waksman Chair in Molecular Genetics.

Author details
[1]Waksman Institute of Microbiology, Rutgers University, 190 Frelinghuysen Road, Piscataway, NJ 08854, USA. [2]Department of Plant Biology and Pathology, Rutgers University, 59 Dudley Road, New Brunswick, NJ 08901, USA.

Authors' contributions
WW designed experiment, analyzed data and wrote the manuscript. JM supervised the work, interpreted data with WW, and revised all versions of the manuscript. All authors read and approved the final manuscript.

Received: 3 October 2011 Accepted: 11 January 2012
Published: 11 January 2012

References

1. Stomp AM: **The duckweeds: a valuable plant for biomanufacturing.** *Biotechnol Annu Rev* 2005, **11**:69-99.
2. Rusoff LL, Blakeney EW, Culley DD: **Duckweeds (*Lemnaceae* family): a potential source of protein and amino acids.** *J Agric Food Chem* 1980, **28(4)**:848-850.
3. Maheshwari SC, Venkataraman R: **Induction of flowering in a duckweed – *Wolffia microscopica* – By a new kinin, zeatin.** *Planta* 1966, **70(3)**:304-306.
4. Blazey EB, McClure JW: **The distribution and taxonomic significance of lignin in the *Lemnaceae*.** *Am J Bot* 1968, **55(10)**:1240-1245.
5. Landolt E, Kandeler R: **The family of *Lemnaceae* - a monographic study.** Veroffentlichungen des Geobotanischen Institutes der Eidgenossischen Technischen Hochschule, Stiftung Rubel; 1987**2.**
6. Landolt E: **The family of *Lemnaceae* - a monographic study.** Veroffentlichungen des Geobotanischen Institutes der Eidgenossischen Technischen Hochschule, Stiftung Rubel; 1986**1.**
7. Smart CC, Trewavas AJ: **Abscisic-acid-induced turion formation in *Spirodela polyrrhiza* L. II. Ultrastructure of the turion; a stereological analysis.** *Plant Cell Environ* 1983, **6(6)**:515-522.
8. Gray KA, Zhao L, Emptage M: **Bioethanol.** *Curr Opin Chem Biol* 2006, **10(2)**:141-146.
9. Wyman CE: **Potential synergies and challenges in refining cellulosic biomass to fuels, chemicals, and power.** *Biotechnol Prog* 2003, **19(2)**:254-262.
10. Cheng JJ, Stomp AM: **Growing duckweed to recover nutrients from wastewaters and for production of fuel ethanol and animal feed.** *CLEAN-Soil, Air, Water* 2009, **37(1)**:17-26.
11. Perry TO, Byrne OR: **Turion induction in *Spirodela polyrrhiza* by abscisic acid.** *Plant Physiol* 1969, **44(5)**:784-785.
12. Reid MS, Bieleski RL: **Response of *Spirodela oligorrhiza* to phosphorus deficiency.** *Plant Physiol* 1970, **46(4)**:609-613.
13. Smart CC, Fleming AJ, Chaloupkova K, Hanke DE: **The physiological role of abscisic acid in eliciting turion morphogenesis.** *Plant Physiol* 1995, **108(2)**:623-632.
14. Appenroth K, Teller S, Horn M: **Photophysiology of turion formation and germination in *Spirodela polyrhiza*.** *Biol Plant* 1996, **38(1)**:95-106.
15. Appenroth K-J, Nickel G: **Turion formation in *Spirodela polyrhiza*: The environmental signals that induce the developmental process in nature.** *Physiol Plant* 2010, **138(3)**:312-320.
16. Smart CC, Trewavas AJ: **Abscisic-acid-induced turion formation in *Spirodela polyrrhiza* L. I. Production and development of the turion.** *Plant Cell Environ* 1983, **6(6)**:507-514.
17. Georgelis N, Braun EL, Shaw JR, Hannah LC: **The two AGPase subunits evolve at different rates in angiosperms, yet they are equally sensitive to activity-altering amino acid changes when expressed in Bacteria.** *Plant Cell* 2007, **19(5)**:1458-1472.
18. James MG, Denyer K, Myers AM: **Starch synthesis in the cereal endosperm.** *Curr Opin Plant Biol* 2003, **6(3)**:215-222.
19. Emes MJ, Bowsher CG, Hedley C, Burrell MM, Scrase-Field ES, Tetlow IJ: **Starch synthesis and carbon partitioning in developing endosperm.** *J Exp Bot* 2003, **54(382)**:569-575.
20. Ouyang S, Buell CR: **The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants.** *Nucleic Acids Res* 2004, , **32 Database:** D360-363.
21. Comparot-Moss S, Denyer K: **The evolution of the starch biosynthetic pathway in cereals and other grasses.** *J Exp Bot* 2009, **60(9)**:2481-2492.
22. Jin X, Ballicora MA, Preiss J, Geiger JH: **Crystal structure of potato tuber ADP-glucose pyrophosphorylase.** *EMBO J* 2005, **24(4)**:694-704.
23. Hwang S-K, Hamada S, Okita TW: **Catalytic implications of the higher plant ADP-glucose pyrophosphorylase large subunit.** *Phytochemistry* 2007, **68(4)**:464-477.
24. Greene TW, Woodbury RL, Okita TW: **Aspartic acid 413 is important for the normal allosteric functioning of ADP-glucose pyrophosphorylase.** *Plant Physiol* 1996, **112(3)**:1315-1320.
25. Henssen A: **Die Dauerorgane von *Spirodela polyrrhiza* (L.) Schleid. in physiologischer Betrachtung.** *Flora* 1954, **141**:523-566.
26. Ley S, Dolger K, Appenroth KJ: **Carbohydrate metabolism as a possible physiological modulator of dormancy in turions of *Spirodela polyrhiza* (L.) Schleiden.** *Plant Sci* 1997, **129**:1-7.
27. Appenroth KJ, Ziegler P: **Light-induced degradation of storage starch in turions of *Spirodela polyrhiza* depends on nitrate.** *Plant Cell Environ* 2008, **31(10)**:1460-1469.
28. Pankey RD, Draudt HN, Desrosier NW: **Characterization of the starch of *Spirodela polyrrhiza*.** *J Food Sci* 1965, **30(4)**:627-631.
29. Fujita M, Mori K, Kodera T: **Nutrient removal and starch production through cultivation of *Wolffia arrhiza*.** *J Biosci Bioeng* 1999, **87(2)**:194-198.
30. Lin Y, Tanaka S: **Ethanol fermentation from biomass resources: current state and prospects.** *Appl Microbiol Biotechnol* 2006, **69(6)**:627-642.
31. Gupta M, Bawa AS, Semwal AD: **Morphological, thermal, pasting, and rheological properties of barley starch and their blends.** *Int J Food Prop* 2009, **12(3)**:587-604.
32. Appenroth K-J, Keresztes A, Krzysztofowicz E, Gabrys H: **Light-induced degradation of starch granules in turions of *Spirodela polyrhiza* studied by electron microscopy.** *Plant Cell Physiol* 2011, **52(2)**:384-391.
33. Franco CML, Ciacco CF, Tavares DQ: **The structure of waxy corn starch: effect of granule size.** *Starch-Stärke* 1998, **50(5)**:193-198.
34. Smith A, Quinton-Tulloch J, Denyer K: **Characteristics of plastids responsible for starch synthesis in developing pea embryos.** *Planta* 1990, **180(4)**:517-523.
35. Burgess D, Penton A, Dunsmuir P, Dooner H: **Molecular cloning and characterization of ADP-glucose pyrophosphorylase cDNA clones isolated from pea cotyledons.** *Plant Mol Biol* 1997, **33(3)**:431-444.
36. Ohdan T, Francisco PB, Sawada T, Hirose T, Terao T, Satoh H, Nakamura Y: **Expression profiling of genes involved in starch synthesis in sink and source organs of rice.** *J Exp Bot* 2005, **56(422)**:3229-3244.
37. Smart CC, Trewavas AJ: **Abscisic-acid-induced turion formation in *Spirodela polyrrhiza* L III. Specific changes in protein synthesis and translatable RNA during turion development.** *Plant Cell Environ* 1984, **7(2)**:121-132.
38. Ruuska SA, Girke T, Benning C, Ohlrogge JB: **Contrapuntal networks of gene expression during *Arabidopsis* seed filling.** *Plant Cell* 2002, **14(6)**:1191-1206.

39. Bhave MR, Lawrence S, Barton C, Hannah LC: **Identification and molecular characterization of shrunken-2 cDNA clones of maize.** *Plant Cell* 1990, **2(6)**:581-588.
40. Smith-White BJ, Preiss J: **Comparison of proteins of ADP-glucose pyrophosphorylase from diverse sources.** *J Mol Evol* 1992, **34(5)**:449-464.
41. Figueroa CM, Esper MC, Bertolo A, Demonte AM, Aleanzi M, Iglesias AA, Ballicora MA: **Understanding the allosteric trigger for the fructose-1,6-bisphosphate regulation of the ADP-glucose pyrophosphorylase from *Escherichia coli*.** *Biochimie* 2011, Corrected Proof.
42. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, *et al*: **The *Sorghum bicolor* genome and the diversification of grasses.** *Nature* 2009, **457(7229)**:551-556.
43. Wu Y, Messing J: **RNA interference-mediated change in protein body morphology and seed opacity through loss of different zein proteins.** *Plant Physiol* 2010, **153(1)**:337-347.
44. Murray MG, Thompson WF: **Rapid isolation of high molecular weight plant DNA.** *Nucleic Acids Res* 1980, **8(19)**:4321-4325.
45. Chen Y, Zhou F, Li G, Xu Y: **MUST: A system for identification of miniature inverted-repeat transposable elements and applications to *Anabaena variabilis* and *Haloquadratum walsbyi*.** *Gene* 2009, **436(1-2)**:1-7.
46. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S: **MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods.** *Mol Biol Evol* 2011.
47. Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T: **Protein structure homology modeling using SWISS-MODEL workspace.** *Nat Protocols* 2008, **4(1)**:1-13.
48. Zdepski A, Wang W, Priest HD, Ali F, Alam M, Mockler TC, Michael TP: **Conserved daily transcriptional programs in *Carica papaya*.** *Trop Plant Biol* 2008, **1(3-4)**:236-245.
49. Park SW, Chung WI: **Molecular cloning and organ-specific expression of three isoforms of tomato ADP-glucose pyrophosphorylase gene.** *Gene* 1998, **206(2)**:215-221.

doi:10.1186/1471-2229-12-5
Cite this article as: Wang and Messing: Analysis of ADP-glucose pyrophosphorylase expression during turion formation induced by abscisic acid in *Spirodela polyrhiza* (greater duckweed). *BMC Plant Biology* 2012 **12**:5.