Supplementary Information

Plant evolution and environmental adaptation unveiled by long-read whole-genome sequencing of *Spirodela*

Dong An[1a], Yong Zhou[2a], Changsheng Li[2], Qiao Xiao[2], Tao Wang[2], Yating Zhang[1], Yongrui Wu[2], Yubin Li[3], Dai-Yin Chao[2], Joachim Messing[4*], Wenqin Wang[1*]

[1]School of Agriculture and Biology, Shanghai Jiao Tong University, Shanghai 200240, China

[2]National Key Laboratory of Plant Molecular Genetics, CAS Center for Excellence in Molecular Plant Sciences, Institute of Plant Physiology & Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200032, China

[3]Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing 100081, China

[4]Waksman Institute of Microbiology, Rutgers University, 190 Frelinghuysen Road, Piscataway, NJ 08854, USA

[a]DA and YZ contributed equally to this work.

*Correspondence should be addressed to: messing@waksman.rutgers.edu and wang2015@sjtu.edu.cn.

This PDF file includes:

www.pnas.org/cgi/doi/10.1073/pnas.1910401116

SUPPLEMENTARY INFORMATION

1. *Spirodela polyrhiza*

1.1. Evolution significance

Duckweeds are one of the most widely distributed and adaptable species throughout the tropical and temperate zones, belonging to early-diverging monocot of the Alismatale (Figure S1). Their reduced morphology and unique aquatic habitat attract increased interest in the studies of plant evolution and adaptation. They provide the foundation of highly productive ecosystems in cleaning and recycling fresh water. They also show potentials in the fields of high-protein animal feed, biofuel and bioreactor. An accurate genome is crucial for the advancement of molecular biology and genetic engineering by dissecting the physiological, evolutionary, and architectural traits.

1.2. History of duckweed genomes

Zostera marine, located in the same order of Alismatale as duckweeds, has many specific genome characteristics and physiological adaptations to complete its entire life cycle under seawater(1, 2). For instance,, genes for stomatal development, terpenoids biosynthesis, ethylene signaling, ultraviolet protection, and phytochromes for far-red sensing were highly contracted or absent. Genes for cell wall modifications, osmotic regulation, and salt tolerance are expanded or neofunctionalized(2). Different from seagrass and other land plants, the evolution of duckweeds has to modulate the genome to inhabit and reproduce in freshwater. Thanks to the sequencing technology improvement, the duckweed genomics has been greatly facilitated, thus extending our understanding in the evolutionary and architectural context of their specific traits(3). It was found that the gene families for aquaporins, lignin biosynthesis, phenylpropanoid, and expansins were lost, while defense-related processes were enriched in the initial genome of *Spirodela polyrhiza* 7498 (Sp7498V2). Similarly, the gene families, including cell-wall modification, nutrient absorption, and small RNA signaling pathway were shrunk in another ecotype of *Spirodela polyrhiza* 9509(4). All the genomes of *Spirodela polyrhiza* and *Lemna minor* retained high gene copy number of glutamine synthetase and glutamate synthase, consistent with the efficient nitrogen absorption and high growth rate(5). However, all the mentioned duckweed genomes are fragmented and incomplete, suffering from short-read sequencing technologies that are unable to span complex regions and repetitive sequences in genomes.

1.3. The power of long reads in genome assembly

An accurate reference genome is a prerequisite for molecular biology and comparative genomics. The explosion of plant genomes benefits from short reads generated by next-generation sequencing, whereas the reads are too short to resolve the abundant repeats and ambiguous assemblies. PacBio single-molecule real-time (SMRT) sequencing produces long and unbiased reads with an average length of 10 Kb, which could precisely delineate their locations in genomes with informative sequences(6). The long-read sequencing has facilitated many complex genome assemblies, such as maize (2,300 Mb)(7), *Oropetium thomaeum* (245 Mb)(8), and quinoa (1,482 Mb)(9). Here, we took advantage of the PacBio long-read technology to improve the genome assembly of *Spirodela* polyrhiza 7498 (Sp7498V3), resulting in the highest continuity and the least gaps among the sequenced duckweed genomes.

2. Genome sequencing and assembly

2.1. Sequencing and assembly of *Spirodela polyrhiza* 7498

We generated ~126-fold coverage of the *Spirodela* genome using 4 SMRT cells on the PacBio Sequel platform (Table S1). The resulting sequences had a polymerase read N50 length of 10.9 Kb and included 10-time coverage with read length over 20 Kb (Table S1). The raw reads longer than 8 Kb (Figure S2) were used to run FALCON correction, and the corrected reads larger than 4 Kb were further assembled through overlap detection by using stringGraph algorithm (https://github.com/PacificBiosciences/falcon) (Figure S3). The assembly generated 411 contigs with a contig N50 of 831 Kb spanning 138 Mb of genome size (Table S2), improved 44-fold contiguity than that of Sp7498V2 (a contig N50 length is 18.9 kb)(10) (Table 1). Combining BAC end sequence (BES)(10), the contigs were constructed into 227 scaffolds with an N50 of 3.3 Mb (Table 1), which were merged into 20 chromosomes with the integration of FISH and Sp7498V2(11) (Table S3). The new reference assembly (Sp7498V3) has 270 gaps with 4.6% missing sequences, compared with 13,459 gap numbers with 11.8% unknown sequences in the assembled genome size of 145 Mb (Sp7498V2) (Figure 1 and Table S3), indicating the improved genome quality and contiguity. An example was given, showing that six gaps in Sp7498V2 were filled by PacBio long reads in the 100-Kb region of chromosome 10, with the retrieval of two genes and the improvement of the other five genes (Figure S4). The comparison of 20 full-fosmid sequences with an insertion size of 40 Kb to the *Spirodela* genome assembly revealed more than 98.6% sequence identity, and all fosmids were spanned by a single contig from the long-read assembly (Table S4).

3. Genome annotation

3.1. PacBio isoform sequencing

To maximally induce gene expression, *Spirodela polyrhiza* was treated by heat, frozen temperature, desiccation, high pH value, UV exposure, heavy metal, and hormone addition. The total high-quality RNA was extracted, pooled and subjected to library construction of PacBio isoform sequencing (Iso-Seq). Iso-Seq bioinformatic analysis was conducted by the three steps of identifying full-length reads, isoform-level clustering, and final consensus polishing. A total number of 753,181 Read of Inserts (ROIs) were generated via the consensus tool. The filtered full-length reads (FLNC) were counted to 545,753 with the definition of 5'-end, 3'-end adapters and a polyA tail. The length distributions of ROIs were expected as each established library, with an average of 1,292, 2,225, 3,344 and 5,715 bp, respectively (Table S5). They were polished by mapping and error correction, resulting in 492,435 high-quality full-length cDNA sequences that were uploaded to the Genbank database with the ID of SRX5321175.

3.2. Gene model

The annotation of *Spirodela* genome was greatly improved with 492,435 full-length cDNA. Together with other homologous evidence, the new version of gene model of *Spirodela* was updated to 18,708 protein-coding genes, 74.6% of which were supported by the full-length transcripts. The reduced gene number (18,708 in Sp7498V3 vs 19,623 in Sp7498V2) was mainly due to the integration of truncated genes. For example, the genes of Spipo31G0006000, Spipo31G000610 and Spipo31G0006200 in Sp7498V2 were concatenated into one gene of Spo013477 in Sp7498V3 (Figure S5a). The genes of Spipo1G0011800, Spipo1G0011600 and Spipo1G0011500 were merged into Spo003046 (Figure S5b). The genes of Spipo3G0023100 and Spipo3G0023200 were combined into Spo000617 (Figure S5c). Given the longer exon size and more exon number within each gene, the mean size of genes was significantly elongated from 3,458 to 4,342 bp (Table 1). BUSCO (Benchmarking Universal Single-Copy Orthologs) evaluation revealed that Sp7498V3 had a much higher rate (86%) compared to 79% of

Sp7498V2 (Table S6), indicating the assembly and annotation of Sp7498V3 were much improved compared to those of Sp7498V2.

3.3. Transposable elements

Most repeats are incomplete, unassembled or highly collapsed in Sp7498V2 assembled from short reads. The newly improved genome of *Spirodela* allowed us to more accurately survey its completeness of repetitive features. We found that the repetitive elements account for a surprisingly high proportion of the Sp7498V3 genome (30.91%) compared to 17.30% in Sp7498V2 (Figure S6). There were 485 copies of helitron, and 121 copies of short interspersed elements (SINE) (Figure S6). Among the different classification of repeat elements, the long terminal repeat (LTR) retrotransposons were the most abundant class and accounted for 18.61% of the *Spirodela* genome (Table S7). We identified 1,544 intact LTRs and a large number of non-intact similar LTR sequences (35,850 LTR-like elements), which was far more than Sp7498V2 (722), but less than rice (3,663)(12) and *Brachypodium* (2,162)(13). All LTRs were classified into seven groups based on the structure features (Ty3/Gypsy, Ty1/Copia, LTR repeats with RH but without INT, LTR repeats INT but without RH, LTR repeats without both INT and RH, solo LTR, LTR-like elements) by using the names of “Snow White and the Seven Dwarfs” including LTR_Happy (370), LTR_Doc (182), LTR_Sneezy (53), LTR_Sleepy (91), LTR_Dopey (110), LTR_Bashful (738) and LTR_Grumpy (35,850), respectively (Table S8). The similarity search showed that only 79 out of 1,544 full LTRs showed significant similarity with LTR database (identity > 60% and e-value < 1e-3), indicating their quickly evolution and diversity.

To estimate LTR integration time(14-16), the sequence similarity of 3’LTR and 5’LTR were aligned(17) and the Kimura parameter distances (K) were measured(18). The divergence time (T) of 1,544 intact LTR divergence was calculated by using the formula T = K/2r (r is the neutral substitution rate of 1.3 x 10^{-8} substitutions per site per million years)(19). The top peak of periodic retrotransposition was discernible 1-2 million years ago (Mya) and the weak peak was around 13-18 Mya (Figure S7). The majority of 83.16% (1,284) LTR was inserted within 2 Mya, while only 2.78% (43) LTR was showed as older elements spanning over 15 Mya (Figure S7), indicating that *Spirodela* genome retained a distinct invasion by an explosive LTR insertion before 2 Mya.

Nested repeats are a common phenomenon in plant genomes, such as rice(20), foxtail millet(21), maize(22, 23). We found that there were 156 nested regions in *Spirodela* genome. A benchmark of the most redundant LTRs on scaffold 15 (539 Kb) was analyzed to show the *Spirodela* genome landscape (Figure 2). An island with a gene-dense region was surrounded by stretches of repeat elements at both ends, which could be the reason for failure to elongate the genome assembly (Figure 2). A close investigation revealed that there were two separate LTR regions (Zone I and IV) accounting for 56% of the sequences, one gene cluster (Zone III) and two separated simple repeat regions (Zone II and V). In Zone I, six LTRs including two Copias and two Gypsys were inserted at 2 - 49 Kb region. In Zone IV, a total of 17 LTRs were defined at the location of 266 - 522 Kb with four runs of transposition that LTRs could insert to or be inserted by other types of LTRs, showing the hot spots for TE insertion. Surprisingly, there were 23 genes observed in Zone III without any intact LTR sequences (Figure 2), including molecular functions of structural constituent of ribosome protein, enzyme activities and carbohydrate biosynthesis. The mechanism of the development of an enriched gene island is currently not understood, which could be subjected to globally transcriptional regulation and also provide novel insights into understanding both diversity and evolution of plant genomes. Taken together, the chromosomal architecture of *Spirodela* genome is characterized by 1) LTRs are the most abundant TE; 2) Nested LTR structures appeared as early as the early-diverging monocots; 3) There are hotspot regions for repeats and genes; 4) The island of LTRs are likely to the genome feature of higher plants, while the gene-dense regions are similar to prokaryote genome.

4. Comparative genome evolution

4.1. Phytogenic tree of flowering plants

Based on conserved single copy of orthologous protein sequences alignment within the species of *Zostera marine*(1), *Phoenix dactylifera(24)*, *Ananas comosus(25)*, *Oryza sativa*(12), *Sorghum bicolor(26)*, *Zea mays*(7)*, Arabidopsis thaliana(27)*, and *Solanum lycopersicum(28)*, a divergence tree was built to show the evolutionary time and to understand how plants evolve in different habitats including sea zone, freshwater area, or terrestrial land. *Spirodela* and *Zostera* belong to the early-diverging monocots of Alismatale, which were clustered as sister clades and diverged between 118.5 and 129.0

million years ago (Mya) (Figure S9). *Spirodela* and other terrestrial monocots were separated between 130.4 and 140.4 Mya.

4.2. Loss of gene families in *Zostera*

The gene families cross *Spirodela*, *Zostera*, *Zea*, *Oryza*, and *Arabidopsis* were compared and assigned to the phylogenetic tree with the signatures of loss and gain (Figure S10). There were 7,647 gene families shared by all 5 species. A number of 674 families were present in *Spirodela*, *Zea*, *Oryza*, and *Arabidopsis*, but lost in *Zostera*. *Zostera,* who lives in a submarine and light-attenuated environment, is absent of genes for developing stomata structure, handling UV radiation and its DNA damage(1). A circadian rhythm is a conserved biological process that displays an endogenous oscillation of ~24-hour rhythms which are driven by a circadian clock. Circadian rhythm genes were lost in *Zostera*, which may be correlated with the undersea-living habitat with low-light, whereas these genes were maintained in *Spirodela*(1, 10). The GO terms were enriched in positive regulation of stomatal complex development, response to UV, starch catabolic process, regulation of DNA repair, glyoxylate cycle, tryptophan biosynthetic process, gamma-tubulin complex localization and regulation of proteasomal ubiquitin-dependent protein catabolic process.

4.3. Gain of gene families in land plants

There were 351 gene families gained in rice, maize, and *Arabidopsis,* in comparison with *Spirodela* and *Zostera*. GO enrichment analysis indicated that the biosynthetic and metabolic processes of secondary metabolites including sesquiterpene, terpene, terpenoid, and isoprenoid were over-presented. They were a large and diverse class of naturally occurring organic chemicals, which are used for precursors for various metabolites, especially gibberellin. The expanded genes in land plants were mainly involved in bio-molecule biosynthesis, playing a key role in conquering gravity and dry land(29). We speculate that *Spirodela* and *Zostera* don't require complex metabolites to cooperate variable temperature, light and nutrition in the comparatively stable aquatic environments.

5. Root development

5.1. Functions of *Spirodela* root

Dicot plants have a taproot system including the main root, lateral roots and root hairs. Monocot plants have a fibrous root system containing adventitious roots and lateral roots(30). Roots are fundamentally vital to maintaining the duckweed buoyancy in the aquatic environment(31). The root system in *Spirodela* contains only adventitious roots without any lateral roots and root hairs (Figure 3a-b). *Wolffia* possesses a pseudo-root of ventral projection to stabilize the downward direction, but is devoid of root structure(32). When we removed roots from mother fronds, they still generated daughter fronds and kept growing. We also confirmed that there was no significant growth difference between fronds with roots and without roots, indicating the nutritional intake ability of the fronds (Figure S11). When the lower surface was pained with water-proof lanolin, the duckweeds grew slowly than control plants (https://www.mobot.org).

5.2. Anatomy and histochemistry

ARs that usually derived from shoots, stems or leaves are dominant in monocots, which is different from eudicots containing lateral roots (LRs)(30). *Spirodela* ARs are generated as many as 12 from the dorsal fronds (Figure 3a-b). The structures of epidermis, cortex and vascular tissue were shown in the cross-section of AR. The epidermis is the outermost boundary, but no root hairs (RHs) were observed. The cortex is loosely packed parenchymatous cells with large intercellular air spaces, allowing gaseous exchange and providing buoyancy (Figure 3c-f). The vascular tissue is highly primitive in *Spirodela* with a tracheary element in the middle surrounded by a ring of phloem tissue (Figure 3c-f). The simple architecture is consistent with its function of maintaining the stability of plant body. However, the layers of pith, conjunctive tissue, xylem (protoxylem and metaxylem), and phloem (protophloem and metaphloem) can be investigated in rice root, which are required to provide mechanical support and to improve water and food transportation over a short or long distance(33).

5.3. Genetic networks controlling root development

The characterization of rice and Arabidopsis mutants identified many responsive genes involved in root development. It was reported that most genes responsible for AR, LR and RH development were conserved(30). For example, OsZFP is a zinc finger protein that contains C2HC-type domains. It was found that the rice plants exhibited a lateral root defective phenotype when OsZFP or OsCYP2 was knocked out, indicating that the interaction of OsZFP and OsCYP2 played a vital role in lateral root development(34). OsORC3 was strongly expressed in the primary root tip, stem base and lateral root primordium. The rice plants of OsORC3 knock-out lacked lateral roots due to the defect of lateral root primordia initiation(35). The mutants of *lrt1*and l*rt2* were identified by their lack of LRs and were found to be less sensitive to auxin. The a*lf1*, *arm1* and *arm2* failed to develop lateral roots in rice(36).

Despite the significant advances that have been made in model dicot of *Arabidopsis* and monocot of rice, the molecular mechanisms underlying root development in *Spirodela* remains unknown. Here, we borrowed molecular regulatory mechanisms from rice and searched homologous genes in *Spirodela*. It was found that *Spirodela* shared all the genes with rice for AR's initiation and elongation (Figure 3g), indicating the conserved mechanism and the early evolution of AR in early-diverging monocots. *Spirodela* had lost gene members associated with LR initiation (ZFP, NAL2 and NAL3), as well as with LR elongation (ORC3 and SLL1). The genes responsible for RH development (RSL, WOX3A, SNDP and RHL1) were also absent, resulting in incomplete RH pathway. The loss of lateral roots and root hairs are consistent with their function as a sea-anchor, since *Spirodela* doesn't need lateral root and root hairs increase the surface area for nutrition intake.

6. Plant defense

6.1. Disease-resistant genes in tandem duplication

The improved genome made it accessible to study tandem gene duplication that was substantial in plant genome, but was challenging to recover from short-read sequencing due to the sequence similarity. *Spirodela* has 1,775 tandem duplicated genes in 692 clusters, which is higher than those of 948 genes in Sp7498V2 genome. We found that the disease-resistant genes were mostly enriched in the tandem duplication, including the gene families of antimicrobial peptides and dirigent proteins (Figure S12). Plant antimicrobial peptides (AMPs) are small defense proteins that constitute a first-line protection from pathogens(37). Compared to 46 antimicrobial peptides (AMPs) in maize(38), a number of 108 AMP members were found in *Spirodela*. There were classified as families of defensin, snaking, lipid transfer protein, hevein, and cyclotide, among which cyclotide family was most dominant (Figure S13). It is worth noting that 92 out of 108 AMP genes were tandemly clustered in up to 22 copies, most of which were supported by full-length cDNA sequences (Figure 4a). There were thionins, defensins, lipid transfer proteins, knottins, heveins, and snakins. Dirigent proteins are also involved in defense response against fungi and insects(39). They could be induced by abiotic and biotic stress such as drought stress, low temperature and abscisic acid(40). The tandemly duplicated genes generally maintain a similar function due to their adjacent location and the sharing of the same regulatory elements, which is different from dispersed gene copies that tend to evolve novel functions. It was also found that 36 out of 92 AMP genes were expressed higher than the medium level of all expressed genes from RNA-seq analysis, indicating their most constant expression to maintaining their active pathogen-resistant ability (Figure 4b). The highly expressed disease-resistant genes were consistent with the fact of low content of dcl3 and 24nt siRNAs in *Spirodela* that was involved in DNA methylation and gene silencing(41). As we took a closer investigation in the pathway of RNA-directed DNA methylation (RdDM) that is the major small RNA-mediated epigenetic pathway in plants, we found that most gene members were expressed significantly lower in *Spirodela* fronds (Table 2) than in *Arabidopsis* leaves(42). Thus, the gene redundancy may affect gene functions for increased dosage(43), conferring *Spirodela* strong resistance against pathogens and pests.

Supplementary Figures

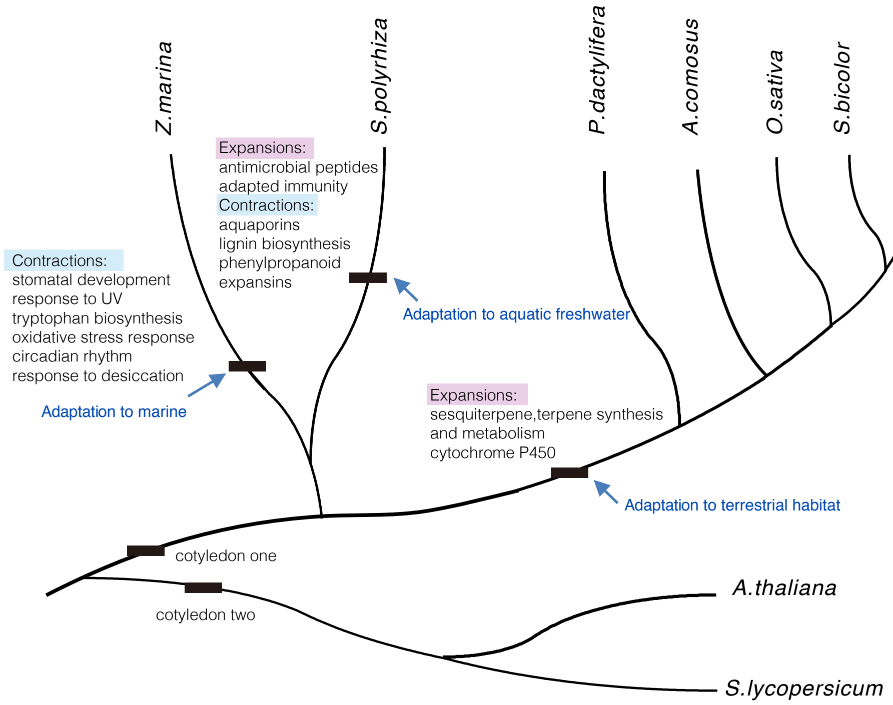


Figure S1. Evolution model of *Spirodela* and other flowering plants.

A phylogenetic tree of flowering plants was presented based on divergence time. Black boxes indicate their unique features in terms of their adaptation and evolution. Gene functions experienced expansion and contraction when compared with other species were listed.

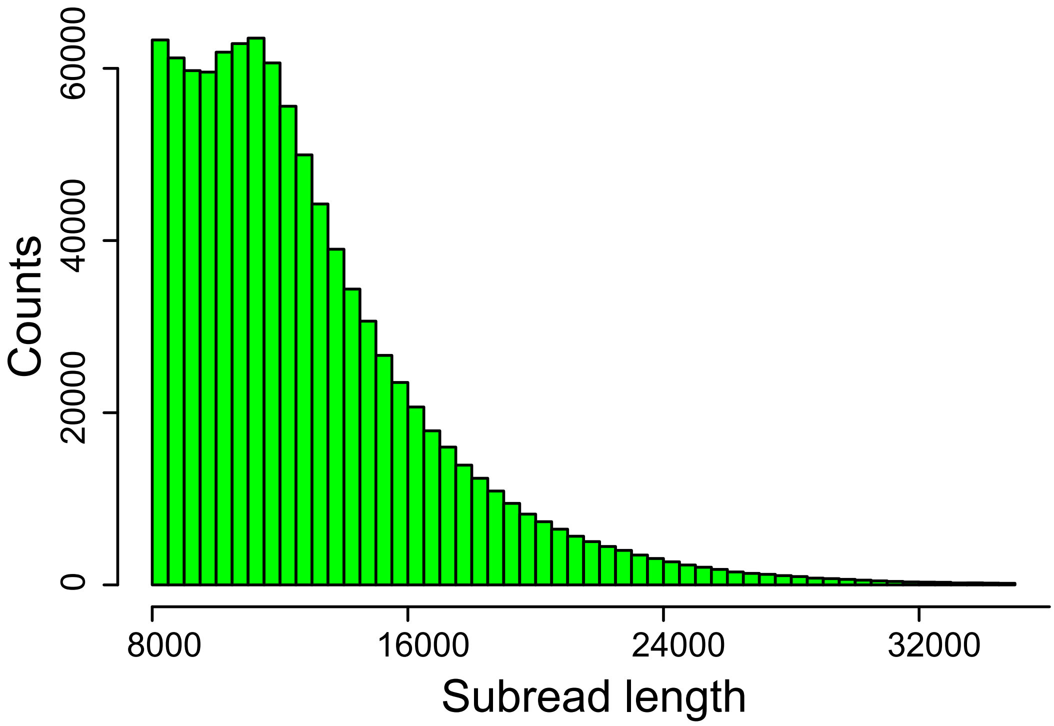


Figure S2. Length distribution of long reads used for genome assembly.

A total of 126-fold coverage of PacBio raw data was generated. The abundance of subreads longer than 8 Kb was counted and used to perform genome assembly in the downstream analysis.

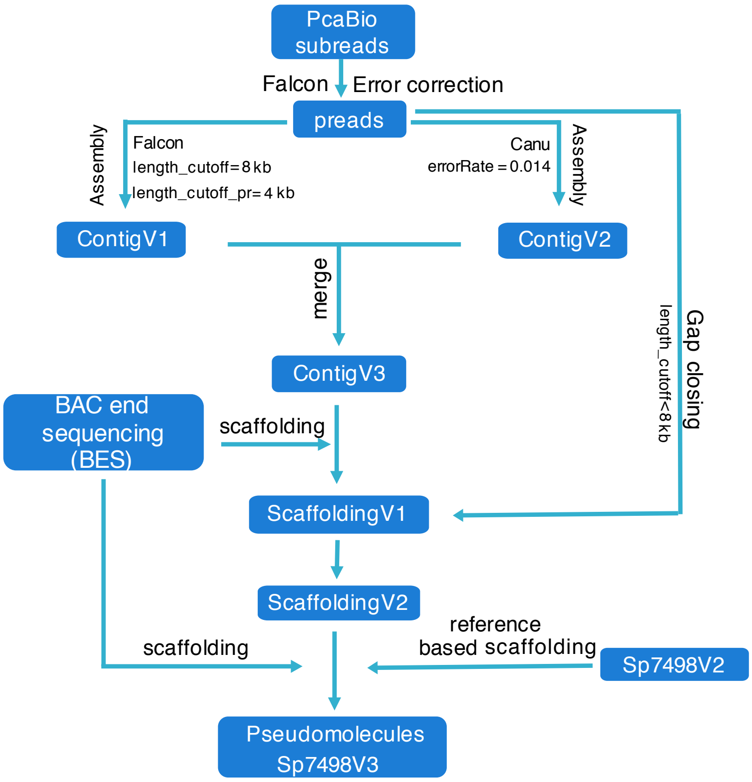


Figure S3. Genome assembly pipeline.

PacBio subreads of more than 8 Kb were corrected to preads by Falcon, and then preads longer than 4 Kb were assembled into contigs by Falcon and Canu. The contigs were polished and gap-filled with short reads (< 8 Kb). After merging these two assemblies, the scaffolds were generated by using BAC end sequencing (BES) data. The scaffolds were oriented to 20 chromosome-level pseudomolecules (Sp7498V3) after the integration of the last *Spirodela* genome version (Sp7498V2).

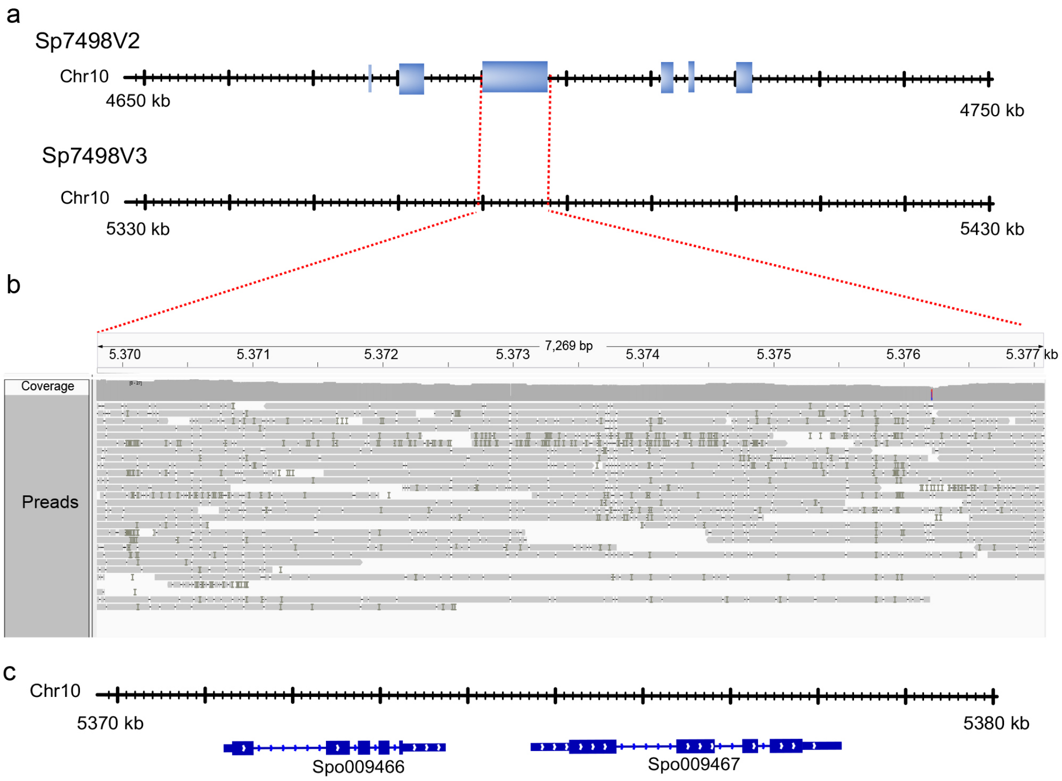


Figure S4. An example of gap filling by PacBio long reads.

(a). There were six gaps within 100 Kb of the last genome assembly from short reads (Sp7498V2), indicated by blue boxes with different sizes in chromosome 10. Gaps are filled with PacBio long reads in the current genome assembly (Sp7498V3). (b). IGV showed that the largest gap (~8 Kb) was closed by PacBio long reads. (c). Two genes (Spo009466 and Spo009467) could be retrieved in the genome assembled from long reads.

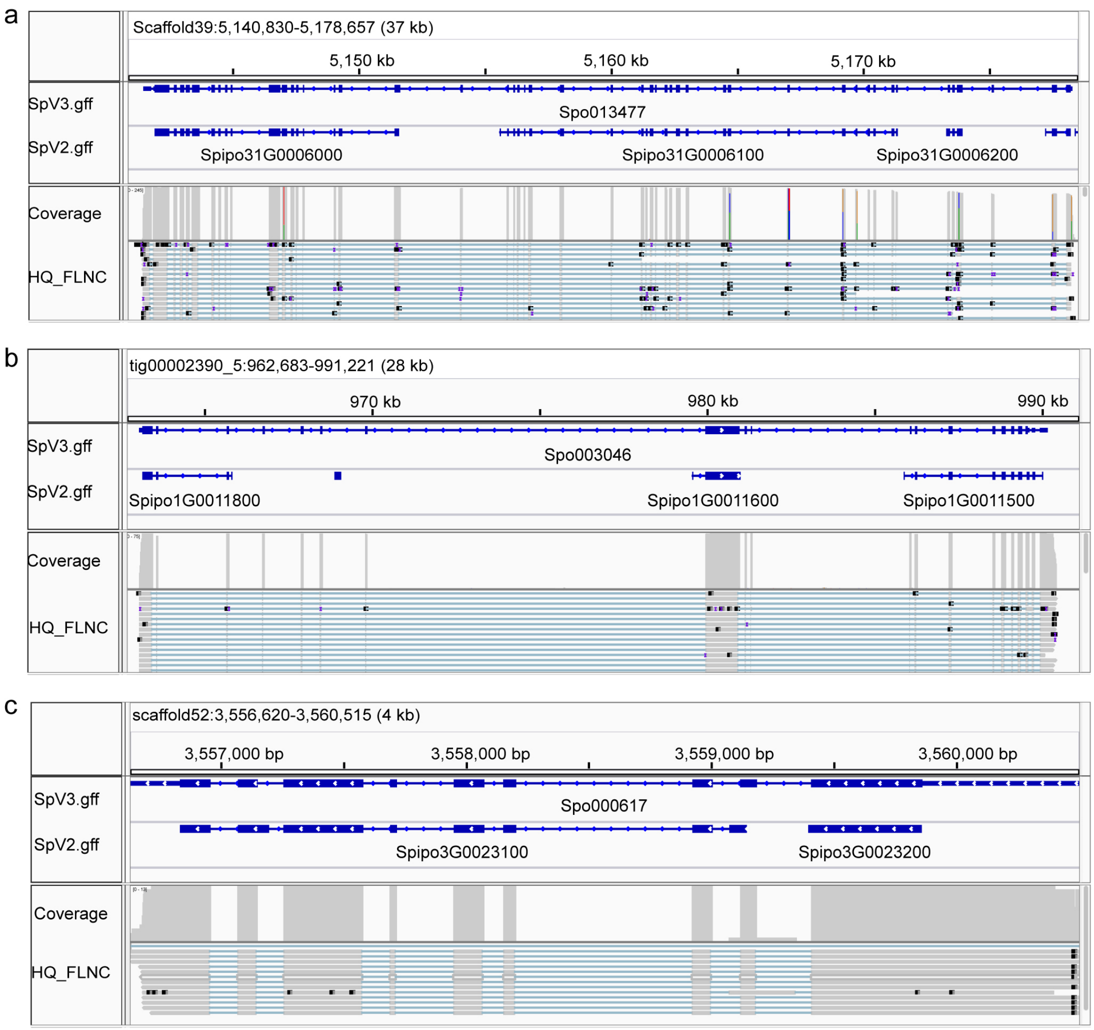


Figure S5. Improvement of gene models.

Three genes shown as examples with a length of 37 Kb (a), 28 Kb (b), and 4 Kb (c) respectively were misannotated as four, four, and two isolated genes in Sp7498V2. With the evidence of high-quality of full-length cDNA (HQ-FLNC) sequenced by PacBio isoform sequencing, they were re-annotated as single gene models in Sp7498V3.

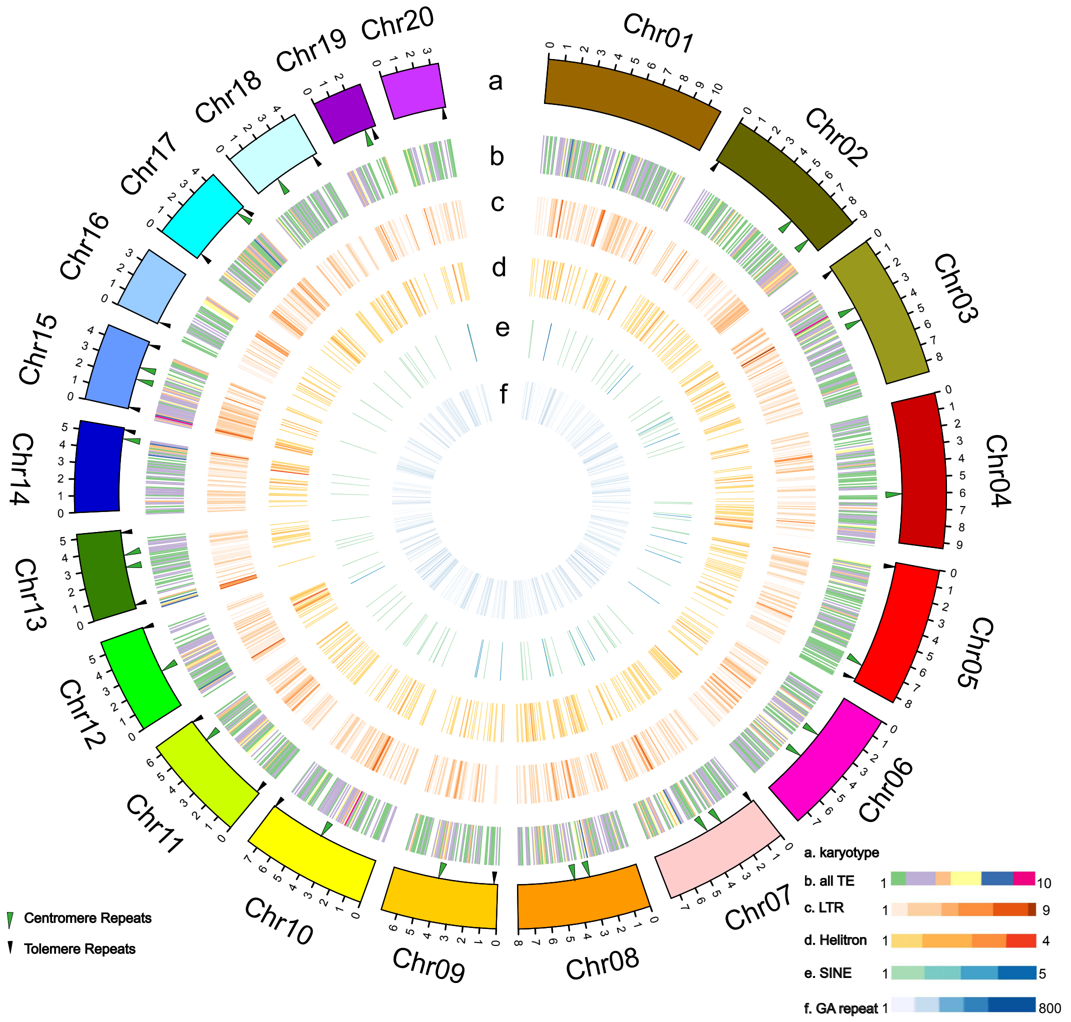


Figure S6. Repeat distribution.

The a (karyotype), b (TE), c (LTR), d (Helitron), e (SINE) and f (GA repeat) were drawn as circles. The predicted centromere and telomere regions were labeled in green and black triangles.

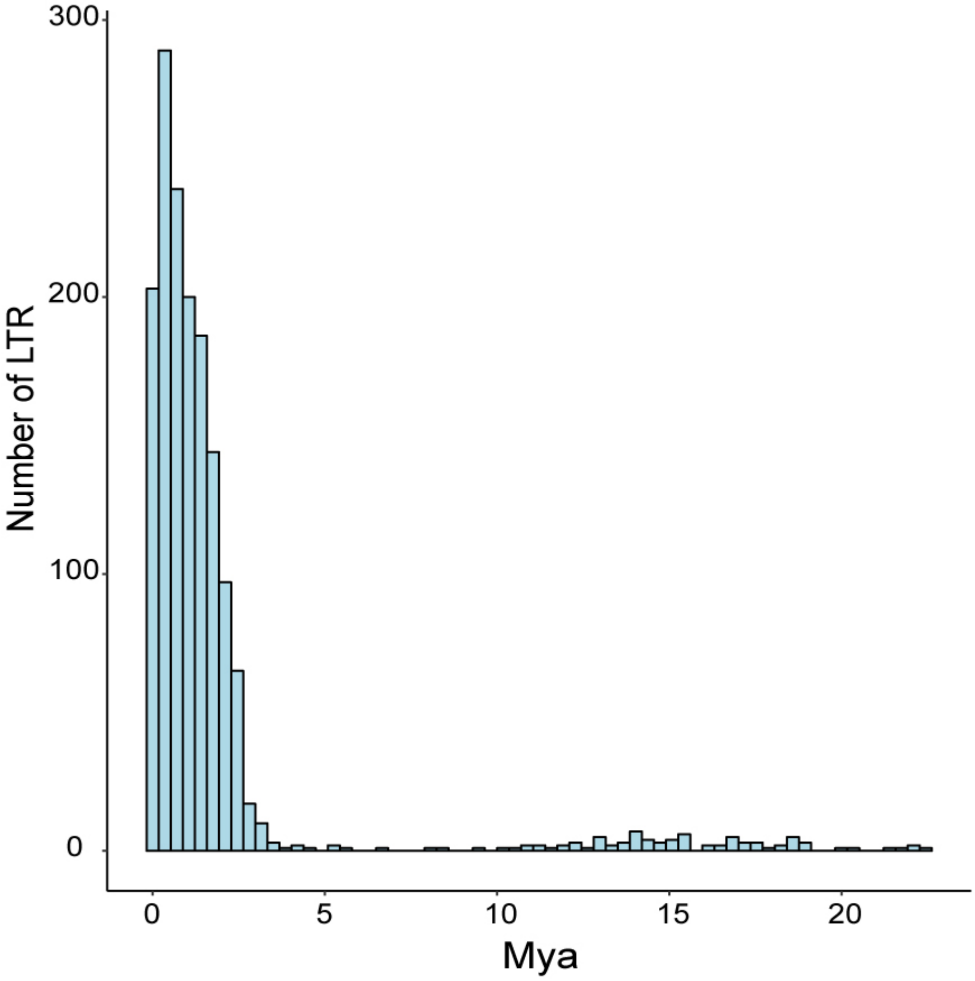


Figure S7. LTR age distribution.

The insertion time of intact LTR retrotransposons was calculated using sequence similarity between 3'LTRs and 5'LTRs. Most insertions occurred before two million years ago (MYA).

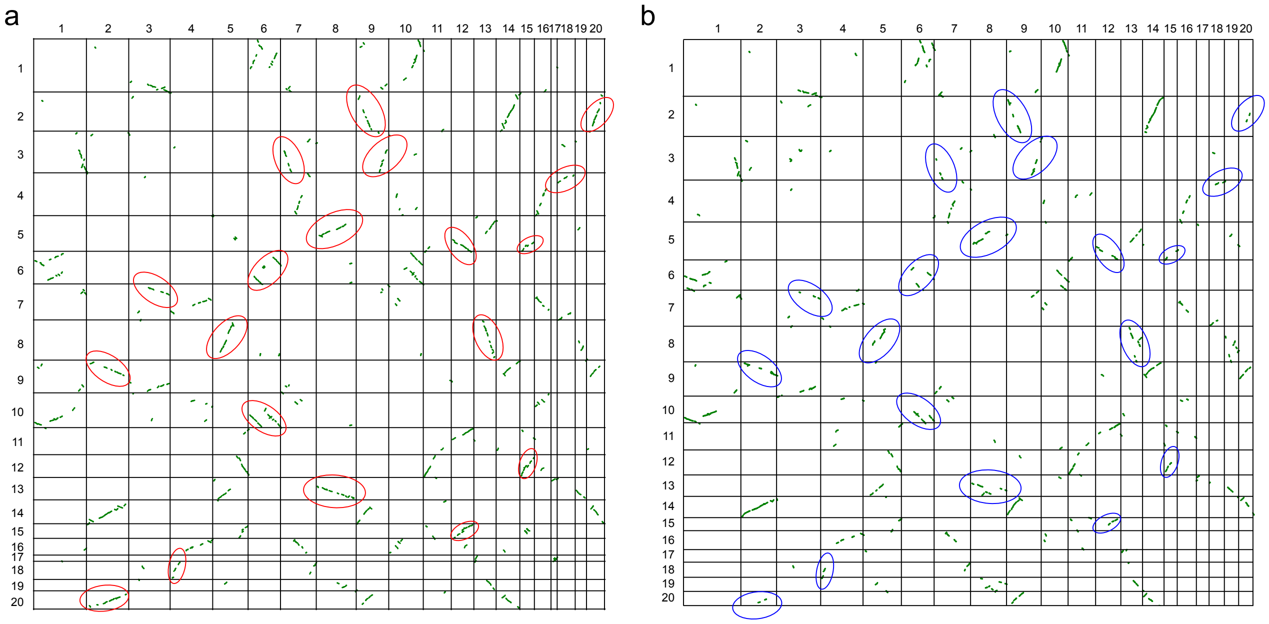


Figure S8. Syntenic dotplot of intragenomic duplications in SpirodelaV2 and V3. (a). The syntenic dotplot of Sp7498V3 versus itself; (b). The syntenic dotplot of Sp7498V2 versus itself. The top hits for each gene of an intra-genomic comparison were plotted. The duplicated regions were shown in green.

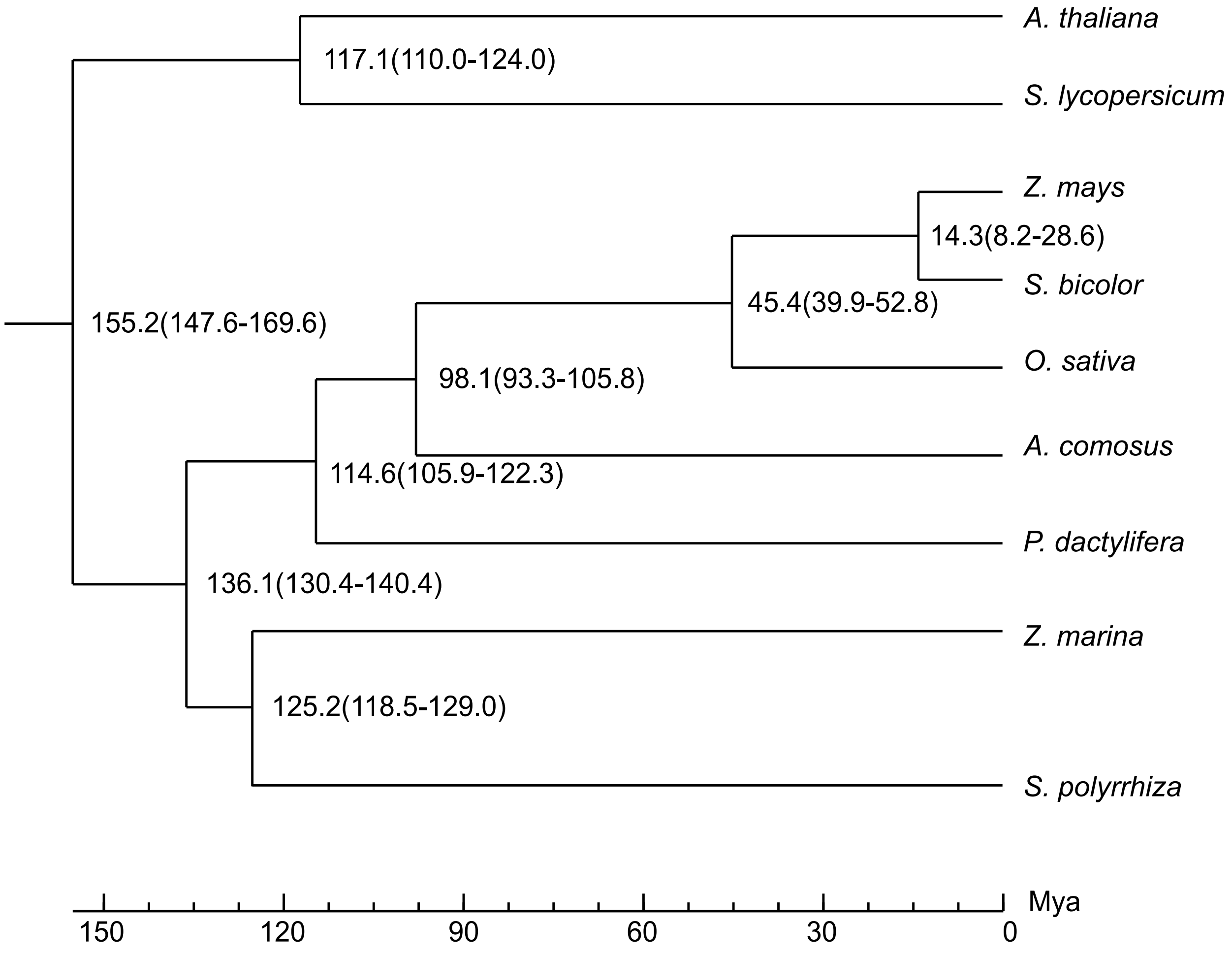


Figure S9. Divergence tree of flowering plants.

A phylogenetic tree of flowering plants with their divergence time was shown with their nodes. Divergence times (million years ago, Mya) were shown in parenthesis.

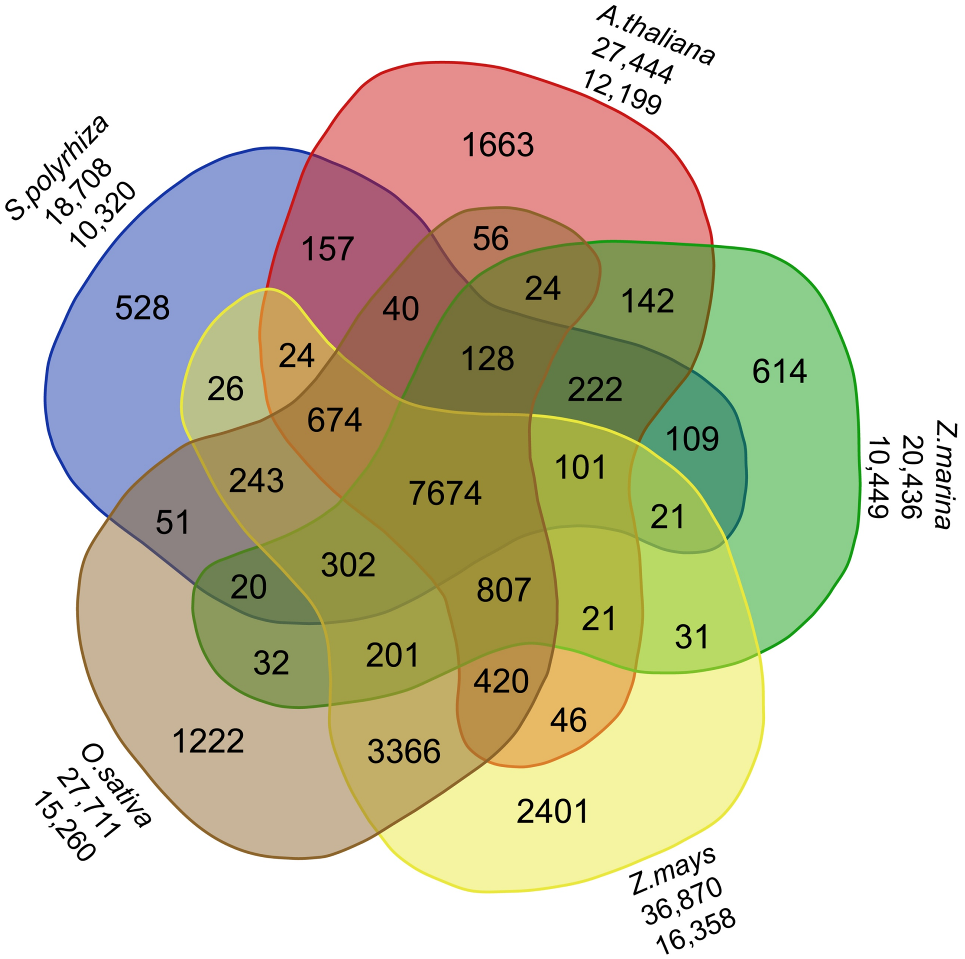


Figure S10. Gene family analysis.

The Venn diagram illustrated shared and distinct gene families via orthoMCL analysis of plant proteomes of *Arabidopsis thaliana*, *Zea mays*, *Spirodela polyrhiza*, *Zostera marina*, and *Oryza sativa*. Numbers below each species are total genes and orthoMCL clusters.

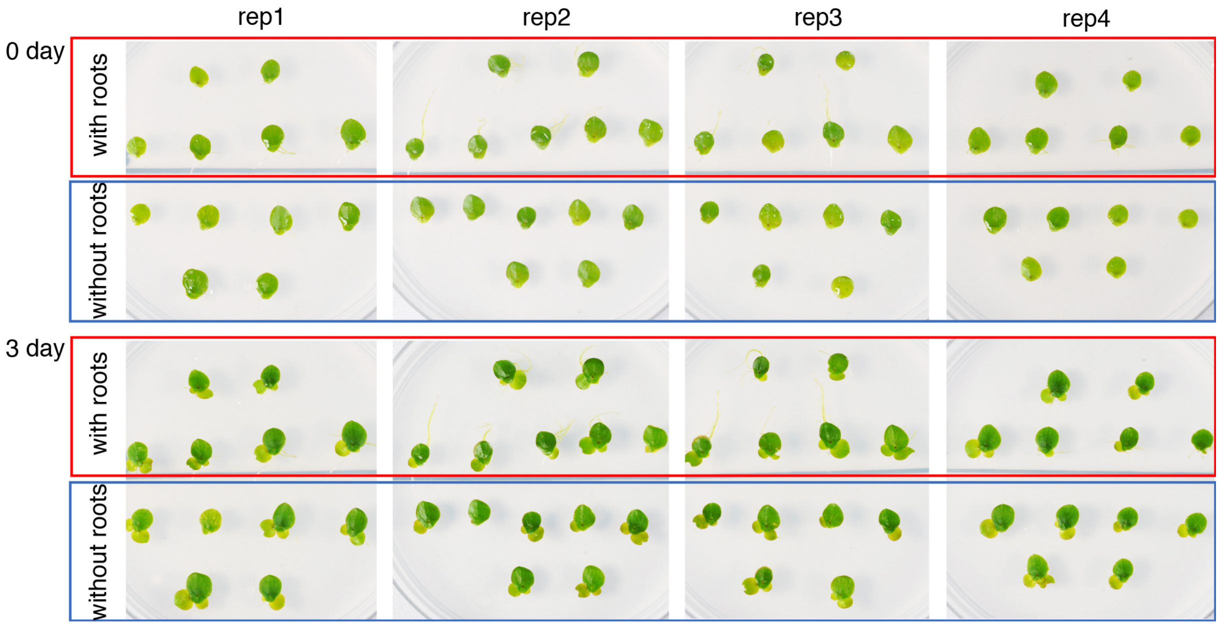


Figure S11. The growth comparison in *Spirodela* with and without roots.

The mother fronds were removed roots and kept in the agar plates. The control fronds with roots under the same growth conditions were also cultured. The growth results of four replicates were investigated after three days.

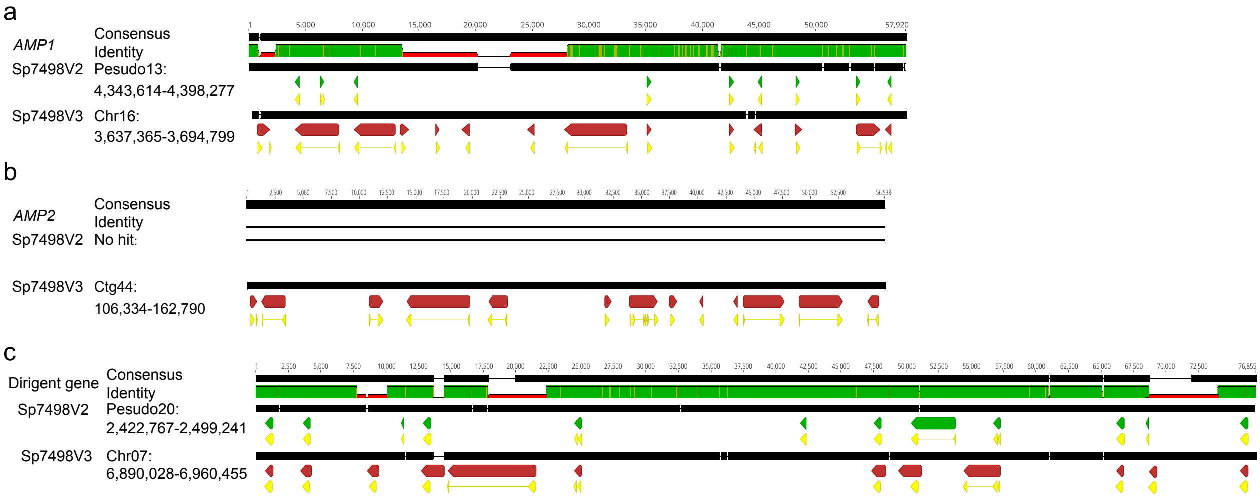


Figure S12. The sequence alignment in tandem duplicated gene regions between Sp7498V2 and Sp7498V3.

(a). Tandem duplicated AMP genes shown in Chr16 region. A 1.3-Kb and a 14.5-Kb gap were filled in Sp7498V3 compared with Sp7498V2. (b). Tandem duplicated AMP genes shown in Ctg44 region. Homologous sequences of Sp7498V3 Ctg44 could not be found in Sp7498V2. (c). Tandem duplicated dirigent genes shown in the Chr07 region. Three gaps were filled where 12 dirigent genes were found. The red lines were gaps in Sp7498V2. The green and red arrows indicate gene structure of Sp7498V2 and Sp7498V3, respectively. The yellow track represents CDS. Genes are drawn to scale.

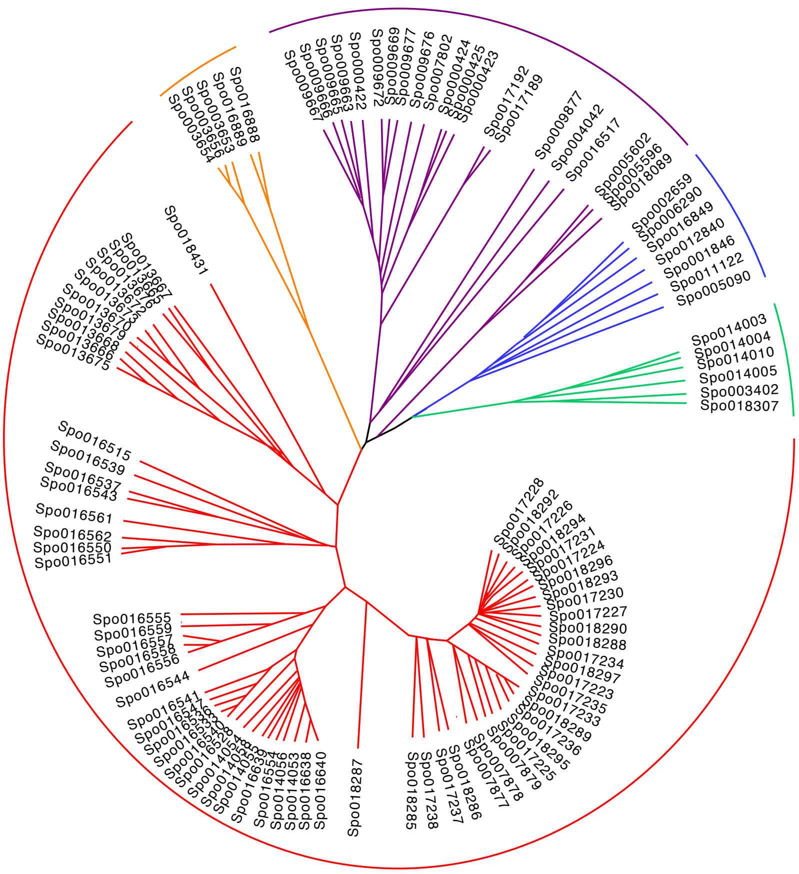


Figure S13. A phylogenetic tree for genes of antimicrobial peptides.

Supplementary Tables

Table S1. A summary of genomic sequencing by PacBio long reads.

SMRT cells were sequenced by using the PacBio Sequel platform, generating 126-fold coverage of the *Spirodela* genome.

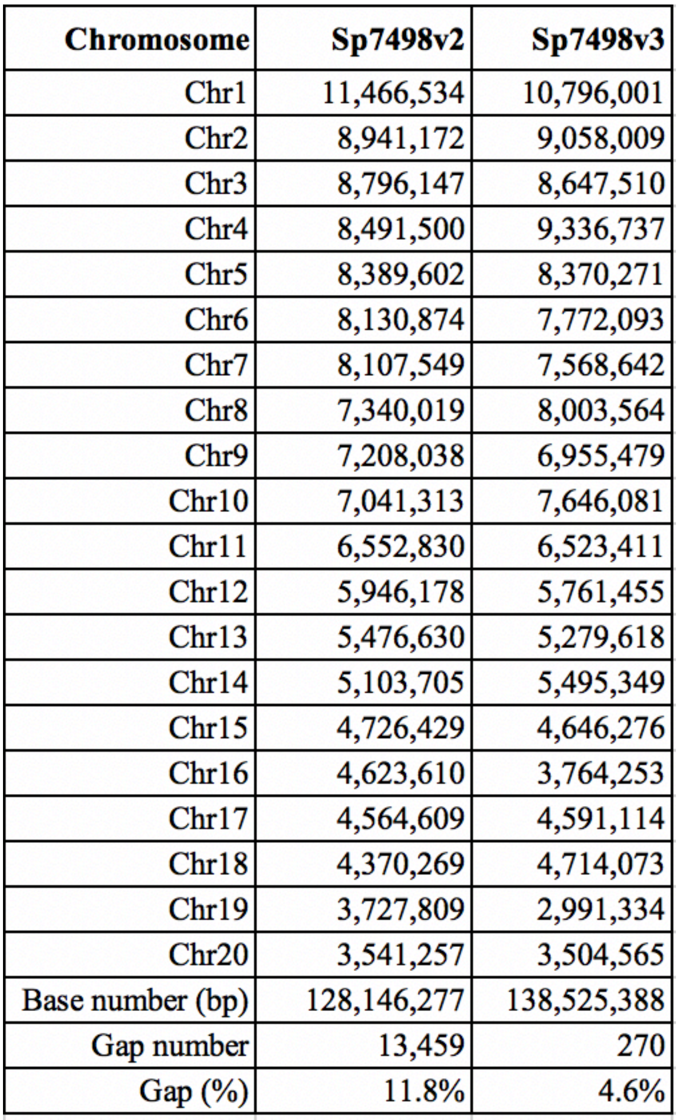
Sequel cell	Total bases (bp)	Number of subread	Mean length of subread (bp)	Maximum subread length (bp)	Subread N50 (bp)	Coverage				
						All	>5 kb	>8 kb	>10 kb	>20 kb
G01	4,145,468,180	638,627	6,491	56,290	10,666	29	24	19	16	2
A12	5,808,236,819	842,595	6,893	89,690	11,109	40	34	28	23	3
C12	4,985,990,428	783,354	6,365	60,240	10,560	34	29	23	19	2
B01	3,283,591,581	462,051	7,107	60,020	11,302	23	19	16	14	2
Total	**18,223,287,008**	**2,726,627**	**6,713**	**66,560**	**10,905**	**126**	**106**	**86**	**71**	**10**

Table S2. 411 contigs assembled from PacBio long reads.

The assembly generated 411 contiguous contigs with a contig ND50 of 831 Kb.

Contig	Length (bp)	Contig	Length (bp)	Contig	Length (bp)	Contig	Length (bp)	Contig	Length (bp)	Contig	Length (bp)	Contig	Length (bp)	Contig	Length (bp)	Contig	Length (bp)
1	3,236,248	51	838,997	101	508,859	151	289,125	201	146,577	251	62,636	301	22,930	351	10,017	401	2,404
2	3,151,984	52	834,086	102	502,491	152	282,084	202	146,554	252	58,707	302	22,903	352	10,014	402	2,217
3	2,316,792	53	831,105	103	491,780	153	281,377	203	146,205	253	58,072	303	21,709	353	10,014	403	2,215
4	2,188,489	54	810,677	104	490,987	154	281,184	204	145,585	254	55,501	304	20,793	354	10,011	404	2,019
5	2,178,241	55	806,904	105	486,873	155	275,171	205	143,542	255	53,779	305	20,655	355	10,004	405	1,809
6	1,957,632	56	806,771	106	472,469	156	270,410	206	140,791	256	53,300	306	19,575	356	9,990	406	1,807
7	1,943,408	57	778,466	107	468,574	157	261,555	207	140,048	257	51,127	307	19,485	357	9,990	407	1,805
8	1,868,646	58	768,557	108	466,059	158	258,864	208	138,149	258	49,593	308	19,470	358	9,989	408	1,514
9	1,833,038	59	763,345	109	461,040	159	257,207	209	134,957	259	48,915	309	18,995	359	9,988	409	1,362
10	1,786,847	60	756,141	110	459,245	160	256,531	210	134,936	260	48,467	310	17,418	360	9,986	410	1,271
11	1,645,206	61	745,450	111	457,592	161	256,218	211	134,936	261	46,463	311	17,357	361	9,927	411	501
12	1,527,618	62	743,613	112	447,156	162	249,693	212	134,389	262	45,514	312	17,219	362	9,898	Sum	138,535,128
13	1,419,464	63	736,004	113	442,622	163	242,888	213	131,397	263	44,669	313	17,013	363	9,733		
14	1,383,338	64	726,824	114	436,156	164	241,101	214	131,222	264	44,278	314	15,756	364	9,619		
15	1,359,755	65	717,200	115	435,955	165	238,848	215	130,909	265	43,943	315	15,195	365	9,581		
16	1,357,755	66	716,921	116	434,249	166	229,632	216	129,080	266	43,631	316	14,375	366	9,575		
17	1,353,359	67	716,602	117	417,964	167	227,081	217	128,709	267	42,233	317	12,160	367	9,529		
18	1,285,807	68	708,613	118	415,651	168	225,156	218	124,648	268	40,839	318	10,940	368	9,403		
19	1,264,517	69	696,578	119	408,488	169	224,128	219	123,822	269	39,567	319	10,749	369	9,382		
20	1,238,260	70	683,482	120	404,445	170	223,681	220	122,639	270	38,540	320	10,748	370	9,303		
21	1,237,583	71	683,265	121	398,354	171	223,458	221	122,626	271	38,485	321	10,736	371	9,274		
22	1,215,352	72	681,950	122	395,461	172	222,429	222	122,295	272	38,130	322	10,718	372	9,135		
23	1,213,227	73	679,215	123	383,823	173	222,244	223	120,631	273	37,362	323	10,707	373	9,060		
24	1,152,096	74	678,986	124	379,733	174	219,025	224	119,223	274	35,587	324	10,527	374	9,046		
25	1,112,722	75	671,788	125	373,194	175	215,495	225	118,403	275	35,520	325	10,516	375	9,012		
26	1,111,170	76	667,767	126	367,486	176	211,735	226	114,702	276	34,670	326	10,503	376	8,961		
27	1,101,444	77	650,359	127	366,364	177	201,952	227	112,372	277	34,210	327	10,499	377	8,852		
28	1,096,951	78	639,144	128	363,957	178	199,102	228	108,196	278	33,902	328	10,498	378	8,764		
29	1,075,504	79	637,419	129	363,524	179	195,047	229	106,695	279	33,822	329	10,497	379	8,709		
30	1,070,080	80	636,390	130	362,064	180	193,667	230	106,199	280	33,573	330	10,495	380	8,701		
31	1,063,430	81	622,019	131	362,019	181	187,771	231	106,020	281	33,492	331	10,494	381	8,671		
32	1,060,458	82	612,776	132	357,841	182	184,228	232	105,296	282	33,340	332	10,489	382	8,592		
33	1,052,932	83	608,797	133	356,009	183	182,448	233	103,192	283	33,132	333	10,485	383	8,450		
34	1,041,627	84	605,375	134	347,668	184	181,512	234	100,714	284	32,522	334	10,480	384	8,396		
35	1,024,991	85	603,471	135	345,610	185	180,247	235	98,548	285	31,979	335	10,478	385	8,380		
36	1,022,380	86	600,221	136	337,398	186	178,695	236	95,287	286	30,998	336	10,467	386	8,364		
37	999,097	87	596,397	137	332,553	187	177,874	237	93,438	287	30,838	337	10,460	387	8,356		
38	992,088	88	572,289	138	332,172	188	177,357	238	93,293	288	30,570	338	10,456	388	8,330		
39	978,300	89	565,279	139	327,672	189	172,360	239	90,324	289	29,445	339	10,455	389	8,142		
40	958,463	90	561,940	140	327,080	190	172,304	240	88,913	290	29,342	340	10,403	390	8,052		
41	947,601	91	558,996	141	324,559	191	171,426	241	82,608	291	28,496	341	10,372	391	8,024		
42	938,767	92	556,624	142	321,964	192	170,246	242	80,871	292	28,278	342	10,341	392	7,786		
43	934,869	93	550,557	143	313,328	193	167,937	243	79,365	293	27,265	343	10,331	393	7,213		
44	929,637	94	544,896	144	311,684	194	160,437	244	78,050	294	26,308	344	10,318	394	6,595		
45	903,496	95	544,323	145	305,576	195	160,324	245	76,297	295	25,499	345	10,308	395	5,357		
46	895,269	96	542,739	146	304,852	196	159,729	246	74,798	296	25,280	346	10,296	396	5,317		
47	874,956	97	539,265	147	295,341	197	153,022	247	72,097	297	25,055	347	10,087	397	4,523		
48	871,865	98	534,911	148	294,995	198	151,446	248	68,194	298	24,079	348	10,078	398	4,290		
49	858,369	99	515,180	149	292,025	199	150,417	249	64,603	299	23,662	349	10,073	399	4,277		
50	856,255	100	513,368	150	290,874	200	149,973	250	63,457	300	23,472	350	10,044	400	2,576		

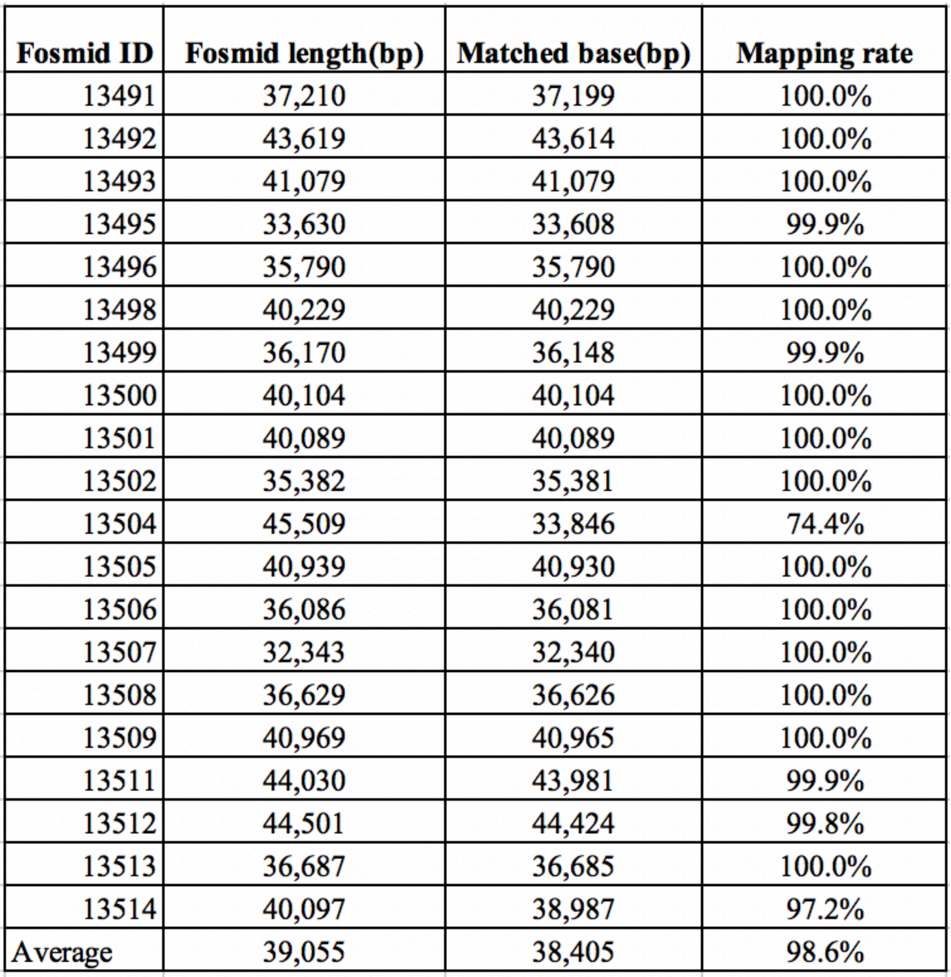
Table S3. Summary of assembly statistics of Sp7498V3 compared to Sp7498V2.



Chromosome	Sp7498v2	Sp7498v3
Chr1	11,466,534	10,796,001
Chr2	8,941,172	9,058,009
Chr3	8,796,147	8,647,510
Chr4	8,491,500	9,336,737
Chr5	8,389,602	8,370,271
Chr6	8,130,874	7,772,093
Chr7	8,107,549	7,568,642
Chr8	7,340,019	8,003,564
Chr9	7,208,038	6,955,479
Chr10	7,041,313	7,646,081
Chr11	6,552,830	6,523,411
Chr12	5,946,178	5,761,455
Chr13	5,476,630	5,279,618
Chr14	5,103,705	5,495,349
Chr15	4,726,429	4,646,276
Chr16	4,623,610	3,764,253
Chr17	4,564,609	4,591,114
Chr18	4,370,269	4,714,073
Chr19	3,727,809	2,991,334
Chr20	3,541,257	3,504,565
Base number (bp)	128,146,277	138,525,388
Gap number	13,459	270
Gap (%)	11.8%	4.6%

Table S4. Evaluation of genome assembly using fosmid sequences.

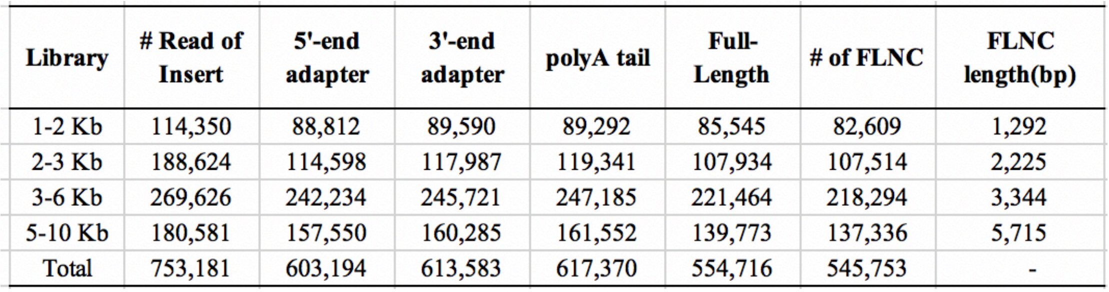
The genome assembly was evaluated with individually 40-Kb sequenced fosmids.



Fosmid ID	Fosmid length(bp)	Matched base(bp)	Mapping rate
13491	37,210	37,199	100.0%
13492	43,619	43,614	100.0%
13493	41,079	41,079	100.0%
13495	33,630	33,608	99.9%
13496	35,790	35,790	100.0%
13498	40,229	40,229	100.0%
13499	36,170	36,148	99.9%
13500	40,104	40,104	100.0%
13501	40,089	40,089	100.0%
13502	35,382	35,381	100.0%
13504	45,509	33,846	74.4%
13505	40,939	40,930	100.0%
13506	36,086	36,081	100.0%
13507	32,343	32,340	100.0%
13508	36,629	36,626	100.0%
13509	40,969	40,965	100.0%
13511	44,030	43,981	99.9%
13512	44,501	44,424	99.8%
13513	36,687	36,685	100.0%
13514	40,097	38,987	97.2%
Average	39,055	38,405	98.6%

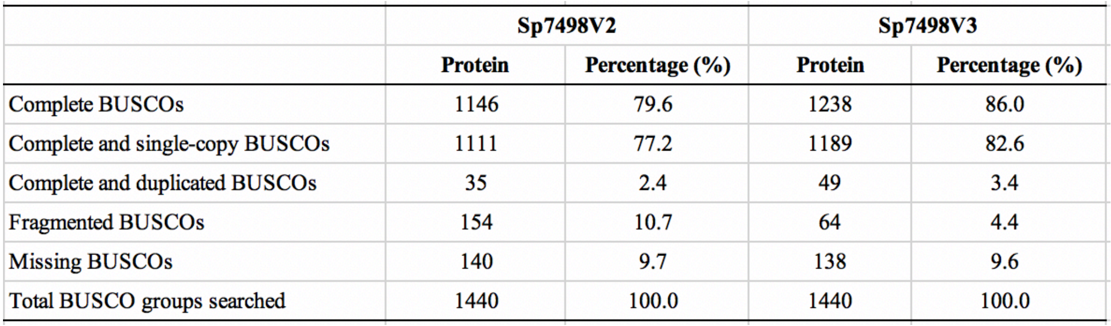
Table S5. Summary of PacBio isoform sequencing.

Four libraries of 1-2 Kb, 2-3 Kb, 3-6 Kb and 5-10 Kb were constructed. Read of Inserts (ROIs) were generated from PacBio raw data. Full-length non-chimeric reads (FLNC) were defined with 5'-end, 3'-end adapters and polyA tails.



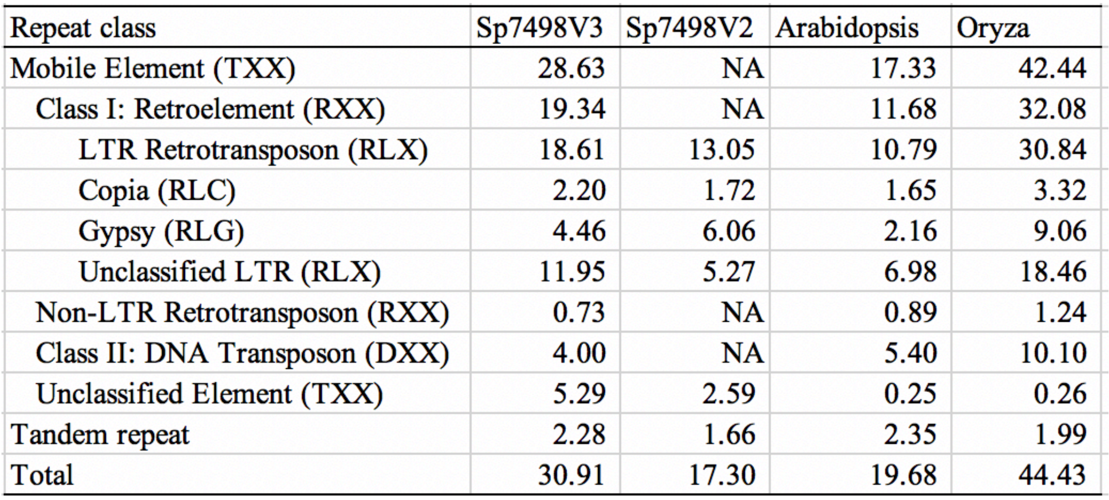
Library	# Read of Insert	5'-end adapter	3'-end adapter	polyA tail	Full-Length	# of FLNC	FLNC length(bp)
1-2 Kb	114,350	88,812	89,590	89,292	85,545	82,609	1,292
2-3 Kb	188,624	114,598	117,987	119,341	107,934	107,514	2,225
3-6 Kb	269,626	242,234	245,721	247,185	221,464	218,294	3,344
5-10 Kb	180,581	157,550	160,285	161,552	139,773	137,336	5,715
Total	753,181	603,194	613,583	617,370	554,716	545,753	-

Table S6. BUSCO evaluation of genome annotation.



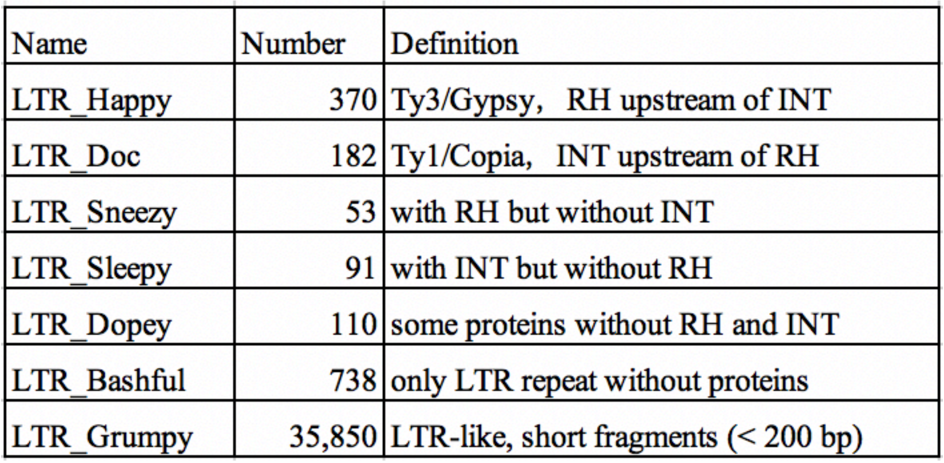
	Sp7498V2		Sp7498V3	
	Protein	Percentage (%)	Protein	Percentage (%)
Complete BUSCOs	1146	79.6	1238	86.0
Complete and single-copy BUSCOs	1111	77.2	1189	82.6
Complete and duplicated BUSCOs	35	2.4	49	3.4
Fragmented BUSCOs	154	10.7	64	4.4
Missing BUSCOs	140	9.7	138	9.6
Total BUSCO groups searched	1440	100.0	1440	100.0

Table S7. Repeat content of Sp7498V3 compared to Sp7498V2, Arabidopsis, and rice.



Repeat class	Sp7498V3	Sp7498V2	Arabidopsis	Oryza
Mobile Element (TXX)	28.63	NA	17.33	42.44
Class I: Retroelement (RXX)	19.34	NA	11.68	32.08
LTR Retrotransposon (RLX)	18.61	13.05	10.79	30.84
Copia (RLC)	2.20	1.72	1.65	3.32
Gypsy (RLG)	4.46	6.06	2.16	9.06
Unclassified LTR (RLX)	11.95	5.27	6.98	18.46
Non-LTR Retrotransposon (RXX)	0.73	NA	0.89	1.24
Class II: DNA Transposon (DXX)	4.00	NA	5.40	10.10
Unclassified Element (TXX)	5.29	2.59	0.25	0.26
Tandem repeat	2.28	1.66	2.35	1.99
Total	30.91	17.30	19.68	44.43

Table S8. LTR retrotransposon composition.



Name	Number	Definition
LTR_Happy	370	Ty3/Gypsy， RH upstream of INT
LTR_Doc	182	Ty1/Copia， INT upstream of RH
LTR_Sneezy	53	with RH but without INT
LTR_Sleepy	91	with INT but without RH
LTR_Dopey	110	some proteins without RH and INT
LTR_Bashful	738	only LTR repeat without proteins
LTR_Grumpy	35,850	LTR-like, short fragments (< 200 bp)

References

1. Olsen JL*, et al.* (2016) The genome of the seagrass Zostera marina reveals angiosperm adaptation to the sea. *Nature* 530(7590):331-334.
2. Lee H*, et al.* (2016) The Genome of a Southern Hemisphere Seagrass Species (Zostera muelleri). *Plant Physiol* 172(1):272-283.
3. An D, Li C, Zhou Y, Wu Y, & Wang W (2018) Genomes and Transcriptomes of Duckweeds. *Front Chem* 6:230.
4. Michael TP*, et al.* (2017) Comprehensive definition of genome features in Spirodela polyrhiza by high-depth physical mapping and short-read DNA sequencing strategies. *Plant J* 89(3):617-635.
5. Van Hoeck A*, et al.* (2015) The first draft genome of the aquatic model plant Lemna minor opens the route for future stress physiology research and biotechnological applications. *Biotechnol Biofuels* 8:188.
6. Li C, Lin F, An D, Wang W, & Huang R (2017) Genome Sequencing and Assembly by Long Reads in Plants. *Genes (Basel)* 9(1).
7. Jiao Y*, et al.* (2017) Improved maize reference genome with single-molecule technologies. *Nature* 546(7659):524-527.
8. VanBuren R*, et al.* (2015) Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. *Nature* 527(7579):508-511.
9. Zou CS*, et al.* (2017) A high-quality genome assembly of quinoa provides insights into the molecular basis of salt bladder-based salinity tolerance and the exceptional nutritional value. *Cell Res* 27(11):1327-1340.
10. Wang W*, et al.* (2014) The genome of the primordial monocotyledonous Spirodela polyrhiza: neotenous reduction, fast growth, and aquatic lifestyle. *Nature Communications* 5:3311-3311.
11. Cao HX*, et al.* (2016) The map-based genome sequence of Spirodela polyrhiza aligned with its chromosomes, a reference for karyotype evolution. *New Phytol* 209(1):354-363.
12. International_Rice_Genome_Sequencing (2005) The map-based sequence of the rice genome. *Nature* 436(7052):793-800.
13. International_Brachypodium_Initiative (2010) Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature* 463(7282):763-768.

14. Bowen NJ & McDonald JF (2001) Drosophila euchromatic LTR retrotransposons are much younger than the host species in which they reside. *Genome Res* 11(9):1527-1540.

15. Jiang N, *et al.* (2002) Dasheng: a recently amplified nonautonomous long terminal repeat element that is a major component of pericentromeric regions in rice. *Genetics* 161(3):1293-1305.

16. SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, & Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. *Nat Genet* 20(1):43-45.

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

18. Rice P, Longden I, & Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 16(6):276-277.

19. Ma J & Bennetzen JL (2004) Rapid recent growth and divergence of rice nuclear genomes. *Proc Natl Acad Sci U S A* 101(34):12404-12410.

20. Kuykendall D, Shao J, & Trimmer K (2009) A Nest of LTR Retrotransposons Adjacent the Disease Resistance-Priming Gene NPR1 in Beta vulgaris L. U.S. Hybrid H20. *Int J Plant Genomics* 2009:576742.

21. Zhang G, *et al.* (2012) Genome sequence of foxtail millet (Setaria italica) provides insights into grass evolution and biofuel potential. *Nat Biotechnol* 30(6):549-554.

22. Dooner HK & He L (2008) Maize genome structure variation: interplay between retrotransposon polymorphisms and genic recombination. *Plant Cell* 20(2):249-258.

23. He L & Dooner HK (2009) Haplotype structure strongly affects recombination in a maize genetic interval polymorphic for Helitron and retrotransposon insertions. *Proc Natl Acad Sci U S A* 106(21):8410-8416.

24. Al-Dous EK, *et al.* (2011) De novo genome sequencing and comparative genomics of date palm (Phoenix dactylifera). *Nat Biotechnol* 29(6):521-527.

25. Ming R, *et al.* (2015) The pineapple genome and the evolution of CAM photosynthesis. *Nat Genet* 47(12):1435-1442.

26. Paterson AH, *et al.* (2009) The Sorghum bicolor genome and the diversification of grasses. *Nature* 457(7229):551-556.

27. Lamesch P*, et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* 40(Database issue):D1202-1210.

28. Sato S*, et al.* (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* 485(7400):635-641.

29. Mizutani M & Ohta D (2010) Diversification of P450 genes during land plant evolution. *Annual Review of Plant Biology* 61:291-315.

30. Bellini C, Pacurar DI, & Perrone I (2014) Adventitious roots and lateral roots: similarities and differences. *Annu Rev Plant Biol* 65:639-666.

31. Landolt E (1986) *The family of Lemnaceae - a monographic study, Vol 1* (Veroffentlichungen des Geobotanischen Institutes der Eidgenossischen Technischen Hochschule, Stiftung Rubel).

32. Sree KS*, et al.* (2015) The duckweed Wolffia microscopica: A unique aquatic monocot. *Flora - Morphology, Distribution, Functional Ecology of Plants* 210:31-39.

33. Chhun T, Taketa S, Tsurumi S, & Ichii M (2003) Interaction between two auxin-resistant mutants and their effects on lateral root formation in rice (Oryza sativa L.). *J Exp Bot* 54(393):2701-2708.

34. Cui P*, et al.* (2017) A zinc finger protein, interacted with cyclophilin, affects root development via IAA pathway in rice. *J Integr Plant Biol* 59(7):496-505.

35. Chen X*, et al.* (2013) OsORC3 is required for lateral root development in rice. *Plant J* 74(2):339-350.

36. Meng F, Xiang D, Zhu J, Li Y, & Mao C (2019) Molecular Mechanisms of Root Development in Rice. *Rice (N Y)* 12(1):1.

37. Hammami R, Ben Hamida J, Vergoten G, & Fliss I (2009) PhytAMP: a database dedicated to antimicrobial plant peptides. *Nucleic Acids Res* 37(Database issue):D963-968.

38. Noonan J, Williams WP, & Shan X (2017) Investigation of antimicrobial peptide genes associated with fungus and insect resistance in maize. *Int J Mol Sci* 18(9).

39. Li N*, et al.* (2017) A novel soybean dirigent gene GmDIR22 contributes to promotion of lignan biosynthesis and enhances resistance to Phytophthora sojae. *Front Plant Sci* 8:1185.

40. Ralph S, Park JY, Bohlmann J, & Mansfield SD (2006) Dirigent proteins in conifer defense: gene discovery, phylogeny, and differential wound- and insect-induced expression of a family of DIR and DIR-like genes in spruce (Picea spp.). *Plant Mol Biol* 60(1):21-40.

41. Fourounjian P*, et al.* (2019) Post-transcriptional adaptation of the aquatic plant Spirodela polyrhiza under stress and hormonal stimuli. *Plant J.*

42. Ferreira-Saab M*, et al.* (2018) Compounds Released by the Biocontrol Yeast Hanseniaspora opuntiae Protect Plants Against Corynespora cassiicola and Botrytis cinerea. *Front Microbiol* 9:1596.

43. Kondrashov FA (2012) Gene duplication as a mechanism of genomic adaptation to a changing environment. *Proc Biol Sci* 279(1749):5048-5057.