Sequence-guided approach to genotyping plant clones and species using polymorphic NB-ARC-related genes

Philomena Chu, [1]

Glen M. Wilson, [2]

Todd P. Michael, [3]

Jennifer Vaiciunas, [1]

Joshua Honig, [1]

Eric Lam, [1]✉

Phone 848 932 6351
Email ericL89@hotmail.com

[1] Department of Plant Biology, Rutgers University, 59 Dudley Rd., New Brunswick, NJ, 08901-8520 USA

[2] Department of Mathematics, University of Oslo, Postboks 1053, 0316 OSLO Blindern, Norway

[3] J. Craig Venter Institute, La Jolla, CA, 92037 USA

Received: 26 April 2018 / Accepted: 23 August 2018

Abstract

Key message

Leveraging the heightened levels of polymorphism in NB-ARC-related

protein encoding genes in higher plants, a bioinformatic pipeline was created to identify regions in this gene family from sequenced plant genomes that exhibit fragment length or single nucleotide differences in different accessions of the same species. Testing this approach with the aquatic plant *Spirodela polyrhiza* demonstrated its superior performance in comparison with currently available genotyping technologies based on PCR amplification.

Abstract

Rapid and economical genotyping tools that can reliably distinguish species and intraspecific variations in plants can be powerful tools for biogeographical and ecological studies. Clones of the cosmopolitan duckweed species, *Spirodela polyrhiza*, are difficult to distinguish morphologically due to their highly abbreviated architecture and inherently low levels of sequence variation. The use of plastidic markers and generic Amplification Fragment Length Polymorphism approaches have met with limited success in resolving clones of *S. polyrhiza* from diverse geographical locales. Using whole genome sequencing data from nine *S. polyrhiza* clones as a training set, we created an informatic pipeline to identify and rank polymorphic regions from nuclear-encoded NB-ARC-related genes to design markers for PCR, Sanger sequencing (barcoding), and fragment length analysis. With seven primer sets, we found 21 unique fingerprints from a set of 23 *S. polyrhiza* clones. However, three of these clones share the same fingerprint and are indistinguishable by these markers. These primer sets can also be used as interspecific barcoding tools to rapidly resolve *S. polyrhiza* from the closely related *S. intermedia* species without the need for DNA sequencing. Our work demonstrates a general approach of using hyper-polymorphic loci within genomes as a resource to produce facile tools that can have high resolving power for genotyping applications.

Keywords

Spirodela polyrhiza
Duckweed

Genotyping
NB-ARC-related genes
Biogeography

Introduction

DNA-based markers for classification and taxonomy can help elucidate complex evolutionary relationships while minimizing ambiguities often encountered with morphology-based methods, making them especially important tools for organisms that may have variable phenotypes in response to their environment. One such plant family, the Lemnaceae (aka duckweed), can have a range of phenotypes depending on environmental conditions (Vaughan and Baker 1994), complicating taxonomic interpretations based solely on morphological characteristics.

AQ1

Several DNA-based strategies have been successful at distinguishing duckweeds at the species level. The first attempt to apply molecular barcoding approaches to genotype duckweed combined plastid-encoded genes with non-molecular data to construct a monophyletic tree for the five genera in Lemnaceae (Les et al. 2002). Wang et al. (2010) tested the barcodes suggested by the Consortium for the Barcode of Life plant-working group to determine the optimal marker for identifying Lemnaceae species using characterized clones from 30 species. This initial effort was subsequently updated and completed by Borisjuk et al. (2015) by analyzing all 37 known duckweed species with the *atpF–atpH* and *psbK–psbI* intergenic plastidic regions to successfully resolve 30 of these species. Multiple clones for most species were included in this work in order to define the intraspecific level of sequence variation for each species with each of the barcodes used. For seven duckweed species, however, available genotyping markers are not sufficient to unambiguously separate two or more potential choices due to a high degree of similarities in their plastid genome sequences. Thus, the range of intraspecific and interspecific sequence variations in the plastid genome sequences overlap significantly in these cases.

When plastidic sequence markers fail to provide sufficient resolving power, the nuclear genome is another resource that can be mined for sequence variations. The amplified fragment length polymorphism (AFLP) technique has been shown to successfully distinguish *Lemna* and *Wolffia* species and resolve many intraspecific variations (Bog et al. 2010, 2013). AFLP was also used to separate clones into *S. polyrhiza, S. intermedia* and *Landoltia punctata* species, but did not provide sufficient variability to unequivocally distinguish clones within each species (Bog et al. 2015). Furthermore, the AFLP technique is challenging to standardize and has low reproducibility between experimental runs. In another recent study, single sequence repeat (SSR) markers were used to distinguish *S. polyrhiza* and *L. punctata* species but found to be insufficient for resolving 49 *S. polyrhiza* clones, with half of the clones originated from locales in China (Feng et al. 2017). Additional investigation with clones from more geographically diverse locations may provide further insight into the applicability of these SSR markers. Currently, it is thus difficult to find a general molecular approach for reliable duckweed identification at the subspecies level by using random primer-based approaches.

As an alternative to AFLP or SSR methods that amplify random template sites within the genome, we reasoned that a more targeted approach to identify candidate genotyping sites in the nuclear genome with hyper-polymorphic behavior can be achieved by using a genome sequence-guided pipeline. The nucleotide-binding leucine-rich repeat protein (NB-LRRs)-encoding genes could be a good source for hyper-polymorphic markers in plants to facilitate genotyping at the subspecies level. Since NB-LRRs function as conserved molecular defenders of pathogen attack (Hammond-Kosack and Jones 1997), the success and health of a plant population may require its NB-LRRs to rapidly adapt to the local pathogen population. Consistent with this expectation, studies of the NB-LRR genes in the model plant *Arabidopsis thaliana* have shown that the NB-LRR gene family is highly polymorphic in sequence and numbers as compared to other genic regions (Clark et al. 2007; Gan et al. 2011; The 1001 Genomes Consortium 2016). The highly conserved nature of the NB-LRRs across various plant lineages and their polymorphic characteristics could provide an excellent reservoir for targeted genotyping markers that can be used for species as well as clone identification in plants.

Duckweeds are increasingly important research material (Lam et al. 2014; Appenroth et al. 2015). In particular, *S. polyrhiza* is an attractive bioethanol feedstock due to its high biomass yield on wastewater in pilot-scale studies (Xu et al. 2011, 2012) and its turion (dormant frond)-forming capability. Turions have been reported to amass 60–70% starch on a dry weight basis, depending on growth conditions (Dolger et al. 1997; Wang and Messing 2012), and could provide an excellent system to understand mechanisms that drive high starch accumulation in these aquatic plants. The availability of two *S. polyrhiza* reference genomes from clones 7498 and 9509 should further aid its development as an aquatic crop (Wang et al. 2014; Michael et al. 2017). However, clones of duckweed species such as *S. polyrhiza* with differing biochemical characteristics could be difficult to distinguish morphologically. Thus, a simple and reliable molecular approach to distinguish *S. polyrhiza* clones will be useful to the duckweed research and application community. To test the approach of NB-LRR-based genotyping markers for identification of duckweed clones, a set of nine *S. polyrhiza* clones that represent a large specific turion yield (STY) distribution and diverse geographical range (Kuehdorf et al. 2014) was chosen as a training set for our informatic pipeline. The genomic data obtained by sequencing these clones were analyzed to identify seven regions among those with the greatest discriminatory power across these nine genomes, which were then tested on a total of twenty-three *S. polyrhiza* clones. Finally, we demonstrated that primer sets identified from our pipeline could be used for rapid PCR-based interspecific identification between the two closely related species of *S. polyrhiza* and *S. intermedia*.

Materials and methods

Plant material and DNA extraction

Duckweed clones used in this study were obtained from the Rutgers Duckweed Stock Cooperative (RDSC) collection maintained in the Department of Plant Biology at Rutgers University. Plants were aseptically maintained on half-strength Schenk and Hildebrandt plant nutrient media (Phytotechnology Laboratories, Shawnee Mission, KS), 0.1% sucrose and 100 mg/L cefotaxime under 16 h light/8 h dark conditions at 25 °C. A number of the *S. polyrhiza* clones used for this study were previously examined for their STY trait (Kuehdorf et al. 2014). Within our nine sequenced clones, we sampled a geographically diverse collection that encompasses a wide range of STY: three were European clones, five from Asia, and one from South America. Eleven additional clones from the Kuehdorf study (Kuehdorf et al. 2014) and three additional *S. polyrhiza* clones that were not part of the Kuehdorf study were also examined. The Landolt accession number for all *S. polyrhiza* clones used in this study are listed in Tables 4 and 5. Ten *S. intermedia* clones from the RDSC collection were also included in this work. Their Landolt numbers are provided in Fig. 5. Total DNA extractions were performed using a modified CTAB protocol (Doyle and Doyle 1987). Concentrations were diluted to 100 ng/µl for storage as stock at − 20 °C.

Identifying polymorphic NB-ARC regions in nine *S. polyrhiza* genomes

Predicted protein sequences in the reference genome of *S. polyrhiza* clone 9509 (Michael et al. 2017) that contain NB-ARC (nucleotide binding —APAF1, R genes, CED-4) domains were identified by using the Pfam family 21 NB-ARC seed profile hidden Markov model (Finn et al. 2016) with hmmsearch from HMMER 3.0 (Eddy 1998, http://hmmer.org/). All hits with an E-value of at most 10^{-4} were considered for our analysis, excluding hits located on unassembled scaffolds. We approximate the NB-LRR proteins using NB-ARC-related genes, which suffices for our downstream analyses.

Eight clones of *S. polyrhiza* were resequenced to ~ 40X coverage using the Illumina NGS platform (sequencing service provided by the Beijing Genome Institute, China) and compared with the reference genome of *S. polyrhiza* 9509 (the first nine clones listed in Table 4). Briefly, high quality genomic DNA from each of the clones were used to prepare 550 bp short-insert libraries and sequenced on the Illumina HiSeq2000 platform by BGI, generating 91 base paired-end reads. Raw sequence data were quality controlled by removing adaptors, contamination and low-quality reads. About 4.5 Gb of cleaned data, which represents roughly 40x coverage, were mapped to the *Spirodela polyrhiza* 9509 reference genome (Michael et al. 2017) using BWA (Li and Durbin 2009). Duplicate reads were removed and bam files sorted with Samtools (Li et al. 2009). Local realignment and SNP calling were conducted with GATK (McKenna et al. 2010). Resulting variant call format (VCF) files generated from GATK were analyzed by SnpEff (Cingolani et al. 2012) and used for our bioinformatic pipeline. The Please insert this sentence in front of "The": The source and VCF files have been deposited into the European Nucleotide Archive (ENA) with the accession numbers ERS2658024 and ERZ681053. 7498 genome was not included in this part of our analysis, since it was sequenced with the 454 NGS platform and BAC sequences by Sanger sequencing (Wang et al. 2014), and therefore no Illumina sequencing-based VCF file was available for this clone. Regions of the NB-ARC-related genes with no variant calls among any of the nine clones were first identified to be candidates for primer design. We define a region of length 200–900 bp in a NB-ARC-related gene of the 9509 reference genome with at least 20 conserved bases at both the 5′ and 3′ ends as a “window.” A total of 8657 windows were identified. Each window was analyzed using Primer3 version 4.0.0 to locate primers using default parameters (Untergasser et al. 2012). Primers were requested to be placed in the initial and terminal conserved regions that were identified for the window. 6576 windows with possible primer sets were produced and are then ranked using two different methods to select those windows that could best distinguish each clone from the others in the training set based on differences in PCR product length. Both methods gave similar window rankings (Table 1).

Table 1

Rankings of windows selected for length polymorphism analysis

	Sp17	**Sp13**	**Sp02b**
np	2	149	141

	Sp17	**Sp13**	**Sp02b**
q-np	8	162	68
rand	2	190	87
q-rand	7	175	75
beagle	2	170	123
q-beagle	16	161	48
max–min	36	21	303
q-max–min	2	31	65

The first six rows are rankings from the bottleneck method with either no phasing (np), random phasing (rand), or phasing with Beagle (beagle). The last two rows show the rankings for the max–min method. A prefix of “q-” indicates that the ranking was determined based on VCF files with variant calls of ~~with~~ quality less than 800 removed

The first method of ranking windows, which we refer to as the “bottleneck method,” took phased variant files (either phased with Beagle (Browning and Browning 2007) [version 27. July 2016 86a], randomly phased, or not phased), and then for each clone in our training set, calculated the difference of the length for each homologous locus from the reference window. Hence for a given window, each clone E was associated with a pair of integers (a_E, b_E) where a_E was the difference in length of the window of one chromosome from the reference and b_E was the difference in length of the window of the other chromosome from the reference. For clones E and F, the bottleneck distance of the pairs (a_E, b_E) and (a_F, b_F) was given by the following expression.

$$\min\left\{\max\left\{|a_E - a_F|, |b_E - b_F|\right\}, \max\left\{|a_E - b_F|, |b_E - a_F|\right\}\right\}$$

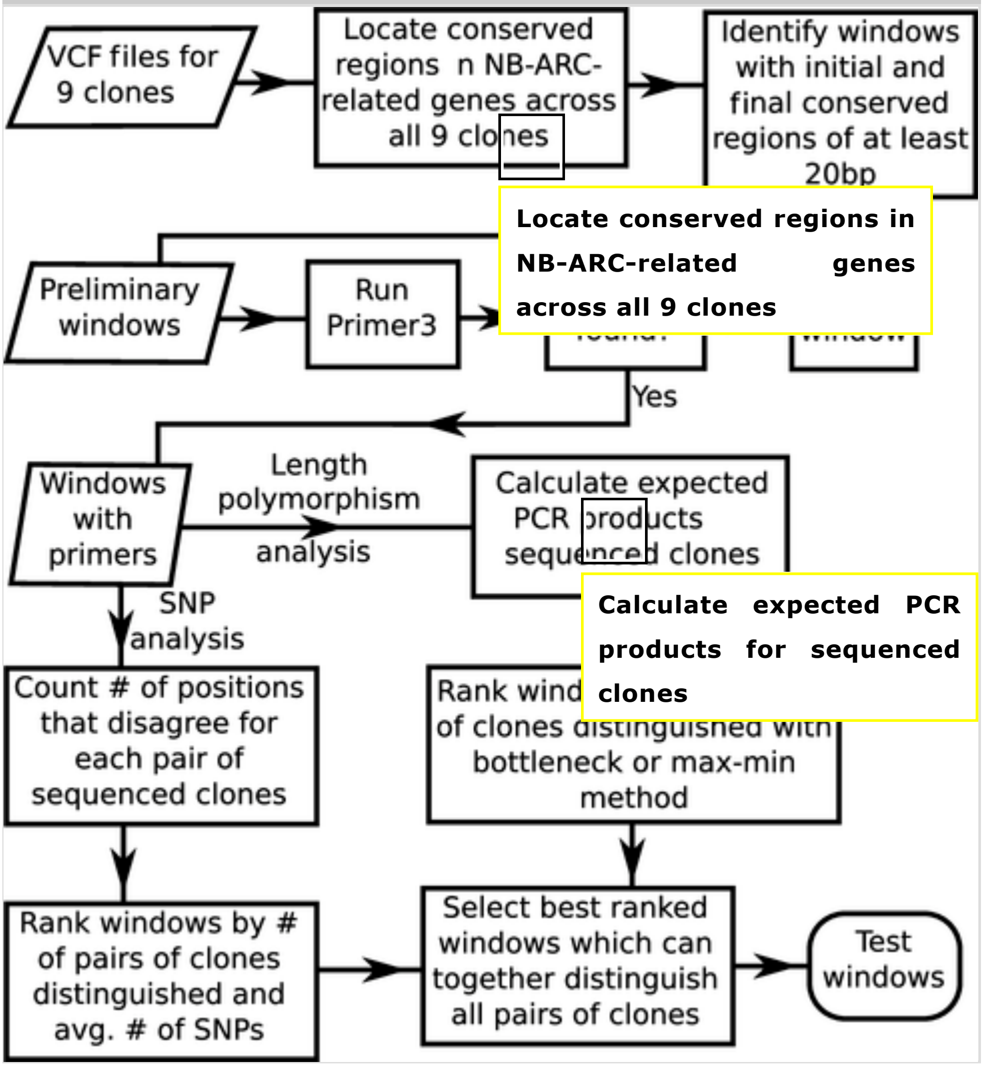
Windows were then ranked by the number of pairs of clones with bottleneck distance of at least five and the average bottleneck distance between all pairs of clones.

The second method of ranking windows, which we refer to as the “max–min” method of ranking, used unphased variant files and considered the range of possible PCR product lengths for each clone as determined by the variant files. For a given window and clone *E*, the maximum and minimum difference in length possible of homologous chromosomes from the reference sequence was calculated as follows. Two running counts were kept as the variant calls in each window were tallied, the maximum count M_E and the minimum count m_E. The length of a homozygous insertion or deletion was added or subtracted respectively from both counts, whereas the length of a heterozygous insertion was added to M_E and the length of a heterozygous deletion was subtracted from m_E. Thus for a fixed window and clone *E*, the difference in length of the PCR product from the length of the window in the reference sequence lied in the interval $[m_E, M_E]$. For a fixed window, each clone *E* was associated with an interval $[m_E, M_E]$. The likelihood that two clones *E* and *F* could be distinguished by their PCR product was measured by calculating the probability that a number from $[m_E, M_E]$ and a number from $[m_{F,} M_F]$ chosen uniformly at random would be separated by a distance of at least five; this probability was called the max–min distance of *E* and *F* with respect to the given window. The windows were then ranked by the number of pairs of clones with max–min distance of at least 0.7 (this was an ad hoc choice) and then by the average max–min distance over all pairs of clones for a window.

It was empirically determined that false positive INDEL calls were associated with quality scores of less than 800 whereas true INDEL calls were associated with quality scores of at least 800 or higher (data not shown). All INDELs from the variant files with quality score less than 800 were removed, and the bottleneck method (with no phasing, phasing by Beagle, and random phasing) and the max–min method were re-run on the revised variant files. A summarized workflow is illustrated in Fig. 1.

Fig. 1

Workflow to identify candidate hyper-polymorphic regions of NB-ARC-related genes. A window is defined as a region of an NB-ARC-related gene with conserved start and stop sub-domains across a set of sequenced accessions from a species of interest


VCF files for 9 clones
Locate conserved regions n NB-ARC-related genes across all 9 clones
Identify windows with initial and final conserved regions of at least 20bp
Preliminary windows
Run Primer3
Yes
Windows with primers
Length polymorphism analysis
Calculate expected PCR products sequenced clones
SNP analysis
Count # of positions that disagree for each pair of sequenced clones
Rank wind of clones distinguished with bottleneck or max-min method
Rank windows by # of pairs of clones distinguished and avg. # of SNPs
Select best ranked windows which can together distinguish all pairs of clones
Test windows

Locate conserved regions in NB-ARC-related genes across all 9 clones

Calculate expected PCR products for sequenced clones

SNP ranking analysis

The set of windows with potential primers described above were ranked to identify those windows that could effectively distinguish the nine clones from one another based on single nucleotide polymorphisms (SNPs) after sequencing the PCR product. For each window and each pair of clones, the number of positions where the clones have different bases within the window were counted. The windows were first sorted by the number of pairs of clones which the window predicted would differ by at least one base pair. The groups of windows which distinguished the same number of pairs of clones were then sorted by the average number of different bases among all clones to yield a ranking of all windows.

PCR reactions and fragment analyses for length polymorphism or SNPs

A modified method of economically labeling PCR products with fluorescent dye was performed for fragment length polymorphism after PCR amplification with genomic DNA (Schuelke 2000). The PCR components in a 20 µl final reaction volume included 4 µl 5x Phusion GC buffer containing 7.4 µmol $MgCl_2$, 0.1 µl Phusion high-fidelity DNA polymerase (Thermo Fisher Scientific, cat. #F530L), 1.6 µl dNTPs (2.5 µmol), 1 µl each of forward (5′ M13(−21) tailed) and reverse primers (10 µmol), and 100 ng DNA template. 8% DMSO was added to the reactions with primer sets Sp17, Sp02a, and Sp02b. PCR reactions with Phusion polymerase were run under the following conditions: 98 °C for 30 s, 30 (98 °C for 10 s, X °C for 30 s, 72 °C for 24 s) cycles, 72 °C for 10 min for final extension, and then 4 °C ∞ until sample analysis. X is 66 °C for Sp17, 64 °C for Sp02a, 62 °C for Sp02b, and 67 °C for Sp13. M13-labeled amplicons were then fluorescently labeled with a 6-carboxyfluorescein (FAM)-labeled-M13(−21) forward primer and marker-specific reverse primer in a similar reaction mix as in the first rounds of PCR, except 1 µl of PCR product from the first round of PCR was used as template for the second round. Fluorescent labeling was performed with eight cycles of 98 °C for 10 s, 53 °C for 45 s, and 72 °C for 24 s, then final extension at 72 °C for 10 min and reaction terminated at 4 °C ∞.

Aliquots of the reaction samples were analyzed by agarose gel electrophoresis to check the quality of the amplification reactions before running on an Applied Biosystems 3500xl Genetic Analyzer (Thermo Fisher Scientific, USA) at the Rutgers Plant Biology DNA Barcoding Core Facility using 2 µl of PCR product into 40 µl of sterile dH_2O. GeneScan 1200 LIZ (Thermo Fisher Scientific, cat. #4379950) was used as a size standard. Fragments with peaks above 200 relative fluorescence units (RFU) were identified using the two-surrounding-peaks sizing method in the microsatellite plug-in from Geneious software (version 10). For each primer set, experimentally observed fragment combinations were grouped and assigned “bin” numbers. M13-labeled fragments amplified from clone 9509 were Sanger sequenced to confirm the predicted sequence in the amplified products. When necessary, PCR amplification products were cloned using the Zero Blunt TOPO cloning kit for sequencing (Thermo Fisher Scientific, cat. #450245). Colonies were screened with M13(-20) forward and reverse primers, amplicons were treated with ExoSAP-IT (Thermo Fisher Scientific, cat. #78200.200.ul) and Sanger sequenced. Sequences were blasted using BLASTN to the 9509 Chromosome Assembly version 3.0 and TBLASTX to the 9509 Annotated Genes v3.5 on the Epigenome server at http://epigenome.rutgers.edu/cgi-bin/duckweed/blast.cgi hosted locally in the Lam laboratory.

SNP Sequencing analysis

Reactions were run using a 5′ M13(−21)-tailed forward primer and the corresponding reverse primer for 30 cycles. PCR components and conditions were conducted as described above for fragment analysis; 8% DMSO was added to reactions. Annealing temperatures for Sp17-SNP were 68 °C, 68 °C for Sp02-SNP and 67 °C for Sp13-SNP. After agarose gel electrophoresis, M13-tailed amplicons were sequenced in both directions using M13F(−21) and the marker-specific reverse primer. When necessary, PCR products were cloned with the Zero Blunt TOPO cloning kit, colony PCR was performed, and amplicons were then Sanger sequenced by Genewiz (South Plainfield, New Jersey).

Sequences were analyzed using a heterozygous base calling program and multiple sequence alignments were made using Seqman Pro (version 10.1). Chromatograms that contained double peaks for the marker Sp17-SNP were also analyzed using Poly Peak Parser (Hill et al. 2014).

Amplification from *S. intermedia* species

PCR reactions for 10 *S. intermedia* clones (listed in Fig. 5) were conducted using the same components and conditions carried out for *S. polyrhiza* for a particular primer set. PCR products were analyzed using agarose gel electrophoresis.

Cluster analysis

Fingerprint and SNP data were analyzed using single-linkage cluster analysis to evaluate how effective the markers distinguish the *S. polyrhiza* clones from one another. A distance matrix was calculated from the results of the length polymorphism and SNP determinations. The distance between two clones is the percentage of markers that have distinct values between the two clones. The single-linkage cluster analysis was performed with SciPy (Jones et al. 2001). In the dendrograms shown in Figs. 6 and 7, clones in clusters formed at 25% dissimilarity or less are similarly colored.

Data availability

The computer code for the length and SNP polymorphism analysis and the rankings for each analysis are available at https://github.com/glenwilson/variant_analysis. Additional data are available upon request.

Results

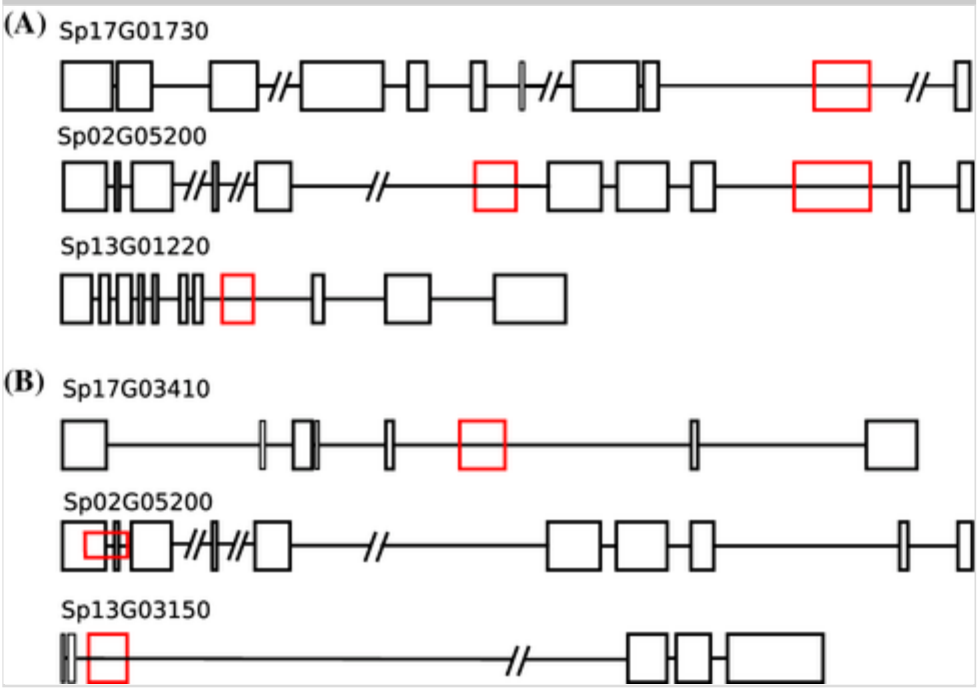
Design of a genotyping approach based on a set of polymorphic loci that are highly conserved in plant genomes

In the *S. polyrhiza* 9509 genome, 53 NB-ARC-related genes were identified through our Pfam-based HMM search (Michael et al. 2017) and were used for downstream analysis. Those genic regions were analyzed across Illumina resequenced genomes of eight additional clones to identify conserved regions for primer design with PCR products ranging from 200 to 900 bp in length. 6576 potential PCR regions were ranked according to the bottleneck and max–min methods (see “Materials and methods” section) to maximize length polymorphism across the nine clones. After filtering out low confidence INDELs, potential primer sets were re-analyzed using the bottleneck and max–min methods. Four windows were then chosen based on their rankings, the quality of their primer sets, and their ability to collectively distinguish all nine sequenced clones under consideration (Fig. 2a). Three of these regions were based on our criteria for markers that allow for lengths 200–900 bp: Sp17G01730 from chromosome position 1085546–1086393 (Sp17) between exons 9–10, Sp02G05200 from positions

3706657–3707464 (Sp02b) between exons 5–6, and Sp13G01220 from 1104684 to 1105262 (Sp13) between exons 7–8. Sp17G01730 and Sp02G05200 had putative disease resistance protein annotations in the 9509 reference genome, and Sp13G01220 was annotated as an AAA-type ATPase (Michael et al. 2017). The four windows were chosen primarily based on their ranking with the max–min method of ranking windows (Table 1). Many overlapping windows were ranked similarly and were effectively identical in the clones that they were predicted to distinguish. The most significant difference amongst overlapping windows was the quality of their primer sets. For example, window Sp17 was effectively the best window from a group of 13 overlapping windows ranked 1–13. Sp13 belonged to the third best cluster of overlapping windows, and Sp02b was identified from the sixth best cluster of overlapping windows. Of the 6576 windows analyzed with the q-max–min method, only the first 841 windows were able to distinguish any pairs of windows with our threshold score of 0.7. An additional window from Sp02G05200 3701793–3702886 (Sp02a) between exons 8 and 9 was identified in a preliminary run of the window selection pipeline that allowed windows to have PCR product lengths up to 1200 bp. This window is ranked 69th out of 8037 windows with length 200–1200 bp using the max–min method without additional filtering steps.

Fig. 2

Selected windows for polymorphism analysis. Red boxes denote candidate windows from each NB-ARC-related gene that were identified from the workflow shown in Fig. 1. Black boxes represent exons (displayed 5′–3′ for each gene). Scale is uniform throughout all genes, however, some introns were truncated as indicated by hash marks. **a** Locations of windows for length polymorphism analysis. For Sp02G05200, Sp02a is further 3′ than Sp02b. **b** Locations of windows for SNP analysis. Red box representing candidate region in Sp02-SNP is modified slightly in size for clarity.



The preliminary window analysis without additional variant filtering identified windows that lacked length polymorphisms when aliquots of the PCR products were assayed on agarose gel electrophoresis. Three regions were selected for experimentation based on the presence of several diagnostic SNPs observed from Sanger sequencing (Fig. 2b): Sp17G03410 chromosome positions 2322464–2323187 (Sp17-SNP), Sp13G03150 positions 3078156–3078820 (Sp13-SNP), and Sp02G05200 positions 3718402–3718995 (Sp02-SNP). These three genes were annotated as being “disease resistance protein-related” in the 9509 reference genome assembly (Michael et al. 2017). Although these windows were not selected using the SNP ranking described in the methods section, Sp13-SNP and Sp02-SNP ranked highly at 292 and 474 respectively out of the 6576 windows. Sp13-SNP was predicted to distinguish all of the nine clones in the training set, whereas Sp02-SNP was predicted to distinguish all but 9509 from 9511. Sp17-SNP was ranked at 2563, but still expected to distinguish 27 out of 45 pairs from the training data. These polymorphic NB-ARC-related genes for SNP occurrence reside in the same three chromosome models as the ones obtained for the length polymorphism screen, suggesting that these three loci in the *S. polyrhiza* genome could be hotspots for polymorphisms. Sp02-SNP is at a different location from the two windows identified for fragment length polymorphism in the gene Sp02G05200, while Sp17-SNP and Sp13-SNP are NB-ARC-related genes on the same chromosomal regions as Sp17 and Sp13, respectively. The design of all chosen primer sets is diagrammed in Table 2.

Table 2

PCR primer sequences

Primer set	Forward/reverse primer sequence
Sp17	CTTCCCTATTCCTCCCACGC/CTGGCTTCTTCTCCACCTCG
Sp02a	TTTTCAGTGTTGATGGCAGC/GCAATCAAGATGCCCTGCAA
Sp02b	TGTGTTCGACTAGTATTGGACCT/CTCGTTGACTACCGCACAGT
Sp13	AAGCCACAATCCTTCCGGAG/GCCTTCTCAGGGGCTTTCAG
Sp17-SNP	GCTTTGAATCCACCGTTCGG/TGGCAGCAACAACCTACGTT
Sp02-SNP	GCCTCTCTTCTCTCCTCTGC/GTTCTGAGCACCTTCCCACA
Sp13-SNP	CCGGGAATGGTATCTCGCAA/ACGCTGTCCCCAAAAAGACA

The first group of primer sets were used for fragment length polymorphism analysis, while the second group (-SNP) were used for SNP polymorphism analysis. Forward primers have M13(−21) sequence added to the 5′ end (not shown)

Fingerprinting of *S. polyrhiza* clones

The four NB-ARC-related windows Sp17, Sp02a, Sp02b and Sp13 were tested for fragment length polymorphisms using PCR and capillary electrophoresis. We experimentally observed two possible PCR fragments for the Sp17 window resulting in four possible fragment bins, eight possible fragments in Sp02a for ten bins, four possible fragments for window Sp02b with four fragment bins, and four replace "four" with "three" fragments for Sp13 with four bins (Table 3). Bins 7–10 with Sp02a primer set were added as new bins for additional fragment patterns observed with the inclusion of the unsequenced clones to the data set in subsequent studies. In some instances, the fragment analysis from capillary electrophoresis output and the agarose gel electrophoresis image are inconsistent because of the 1200 bp size limitation with our capillary electrophoresis system (Figs. 3, 4, 5).

Table 3

Capillary electrophoresis output for each fragment length polymorphism primer set

Primer set	Bin	Fragment combination
Sp17	0	None
	1	777
	2	777, 765
	3	765
Sp02a	0	None
	1	397, 863, 1093
	2	397, 863, 1072, 1093, 1110, 1140
	3	431, 1072, 1093, 1110, 1140
	4	1072
	5	1072, 1110
	6	1093, 1110, 1121, 1140
	7	397
	8	397, 863, 1110
	9	431, 1072
	10	1072, 1093, 1110, 1140
Sp02b	1	572

Experimentally observed PCR products were grouped into primer set-specific bins. For each primer set, the Fragment combination column lists the number of PCR products and their lengths for each bin determined by capillary electrophoresis. This table serves as a k~~K~~ey for the bin assignment used in Table 4

Primer set	Bin	Fragment combination
	2	582
	3	572, 582
	4	554, 564, 572, 582
Sp13	1	419
	2	446
	3	419, 446
	4	430, 446

Experimentally observed PCR products were grouped into primer set-specific bins. For each primer set, the Fragment combination column lists the number of PCR products and their lengths for each bin determined by capillary electrophoresis. This table serves as a k~~K~~ey for the bin assignment used in Table 4

Fig. 3

Fragment comparison between ten clones of *S. polyrhiza* using primer set Sp17. PCR amplification from Sp17G01730 with a 66 °C annealing temperature. The lane headings correspond to assigned numbers for each sequenced clone in Table 4. M: 1 kb ladder as size marker (GoldBio), numbers to each side correspond to the size of DNA in base pairs. N: negative control without template (water added)

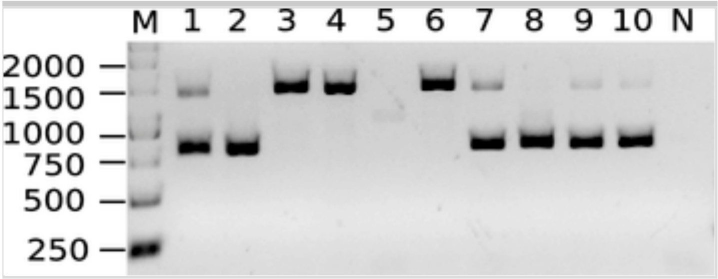


Fig. 4

Fragment length polymorphism between 10 clones of *S. polyrhiza* using

Sp02a. Fragments amplified from the Sp02G05200 window using annealing temperature at 67 °C. Lane headings correspond to the same clones used in Fig. 3. M = 1 kb ladder (GoldBio), N = negative control (water)

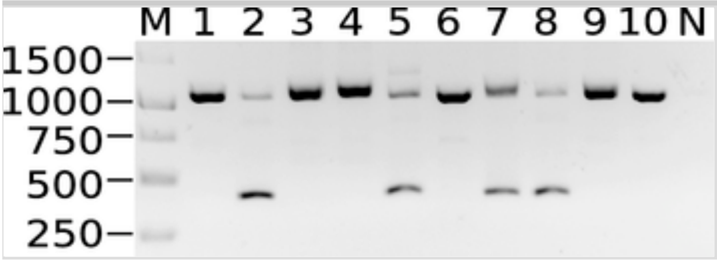


Fig. 5

Use of NB-ARC primer set for interspecific barcoding application. **a** replace "a" with "A" The fragment length polymorphism primer set Sp17 (from Sp17G1730) is used for PCR amplification using gDNA of non-reference *S. polyrhiza* and *S. intermedia* clones as templates with annealing temperature set at 66 °C. A 0.8% agarose gel was run at 50 V for 2 h. **b** replace "b" with "B" SNP primer set Sp13 (Sp13G01220) using gDNA of non-reference *S. polyrhiza* and *S. intermedia* clones as templates with annealing temperature set at 67 °C. A 2% agarose gel was run at 50V for 2 h. M = 1 kb ladder (GoldBio), N = negative control (water), lane headings indicate the clone numbers in the Landolt collection

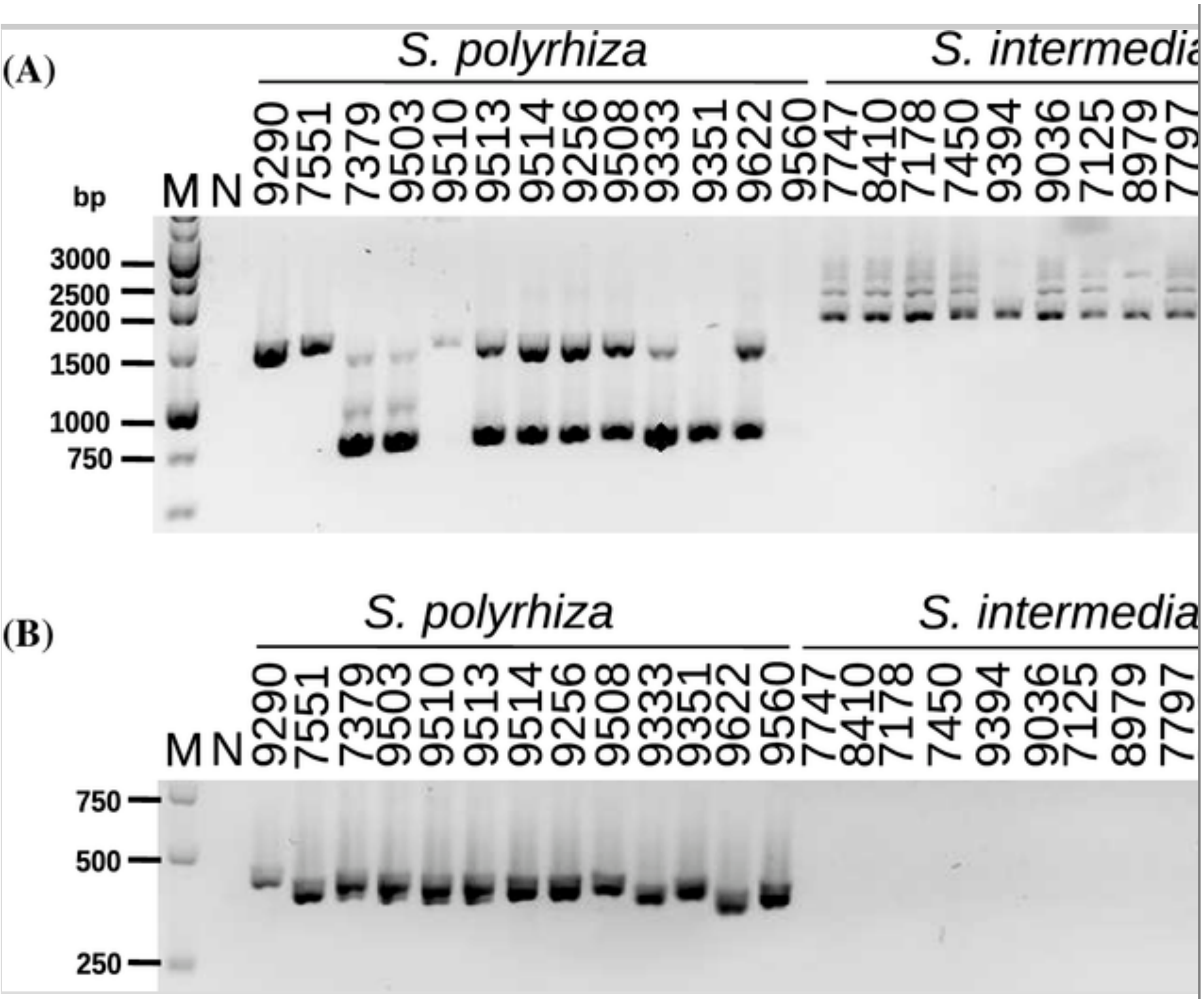


Table 4

NB-ARC-related fragments from capillary sequencer results for tested *S. polyrhiza* clones

Lane	Clone	Origin	Sp17	Sp02a	Sp02b	Sp13
1	9509	Germany	1	4	2	4
2	9316	India	2	1	1	1

Each clone was assigned a bin number according to its fragment combination output from the particular primer set listed in Table 3. Each clone’s four-digit Landolt collection number is listed along with its locale of origin. The “Lane” column refers to sequenced clones in Figs. 3 and 4. CZ is an abbreviation for Czech Republic and MZ is an abbreviation for Mozambique

Lane	Clone	Origin	Sp17	Sp02a	Sp02b	Sp13
3	9501	Albania	0	3	4	3
4	9502	Ireland	0	6	4	3
5	9504	India	0	1	1	4
6	9511	Russia	0	4	2	1
7	9512	Russia	3	2	4	3
8	9506	India	2	1	1	3
9	9242	Ecuador	3	5	1	1
10	7498	USA	3	4	1	1
	7379	India	2	1	1	3
	9290	India	0	5	1	1
	9503	India	2	1	1	3
	7551	Australia	0	2	4	3
	9333	China	3	8	3	1
	9351	Vietnam	3	7	1	1
	9256	Finland	1	2	4	2
	9508	Poland	1	9	1	2
	9510	MZ	3	10	4	3
	9513	CZ	1	10	3	2
	9514	Austria	1	10	4	2
	9560	Hungary	1	0	1	3
	9622	Germany	1	10	4	2

Each clone was assigned a bin number according to its fragment combination output from the particular primer set listed in Table 3. Each clone’s four-digit Landolt collection number is listed along with its locale of origin. The “Lane” column refers to sequenced clones in Figs. 3 and 4. CZ is an abbreviation for Czech Republic and MZ is an abbreviation for Mozambique

All 9509 PCR products, except those from Sp17, amplified the expected target sequence. 9509 gDNA PCR product from the Sp17 primer set contained an additional fragment of approximately 1500 bp in size (Fig. 3) that matched to Sp17G03530 by BLAST, which was annotated as "protein of unknown function" in the 9509 Annotated Gene set (version 3.5, http://epigenome.rutgers.edu/cgibin/duckweed/blast.cgi). When multiple PCR bands amplified from the 9509 gDNA template with a primer set, those sequences were cloned and sequenced. PCR products from the other clones that share the same fragment size as in clone 9509 are assumed to likely correspond to the same loci as that amplified from the 9509 template since the genomes are highly conserved (Michael et al. 2017).

Each of the nine *S. polyrhiza* clones from the training genome sequence dataset had a unique fingerprint based on the binning with the four fragment length primer sets, validating our pipeline's accuracy. The inclusion of clone 7498, which also has been sequenced (Wang et al. 2014), and additional unsequenced clones resulted in finding three Indian clones 9503, 9506, and 7379 sharing the same fingerprint pattern based on fragment length polymorphism markers, and also clones 9622 and 9514 being indistinguishable from each another. However, 18 out of the 23 clones examined had unique fingerprint patterns using the four primer combinations (Table 4).

Genotyping with SNP primers

The 23 *S. polyrhiza* clones are distinguishable from one another using SNP combinations from Sp17-SNP, Sp02-SNP, and Sp13-SNP primer sets, except for the three Indian clones 9503, 9506 and 7379. The combination of both fragment length and SNP data still did not improve the distinction amongst the three Indian clones.

The 9509 amplicon length (minus the forward and reverse primer sequences) was 588 bp, 546 bp, and 493 bp for Sp17-SNP, Sp02-SNP, and Sp13-SNP target sites, respectively. Double peaks were seen with this clone downstream of a heterozygous INDEL (9 bp) in chromatograms from Sp17-SNP genomic DNA amplicons, requiring additional analyses to determine the full sequence. Sp02-SNP genomic DNA chromatograms from the unsequenced *S. polyrhiza* clones contained complicated traces, possibly due to increase in copy number of this NB-ARC-related locus. Affected PCR products were cloned and sequenced. The Sanger-sequenced colony results take precedent when a discrepancy arises amongst the Illumina, genomic DNA or Sanger sequencing results. Heterozygous SNPs/INDELs found in the different clones are illustrated in Table 5. The average number of SNPs from all nine clones in the three NB-ARC-related loci were 0.98% in Sp17-SNP, 1.32% in Sp02-SNP, and 0.38% for Sp13-SNP, all higher than the genome-wide averages reported from comparing clone 7498 sequencing reads versus the 9509 reference assembly (0.33%) (Michael et al. 2017).

Table 5

Representative SNPs for Sp02-SNP, Sp13-SNP, and Sp17-SNP

~~Clone~~Clone	Sp02-SNP										Sp13-S	
	99	**184**	**216**	**250**	**268**	**274**	**288**	**299**	**380**	**498**	**111**	**11**
9509	A	A	G	A	A	A	C	C	T	G	T	G
9316	R	R	R	M	R	R	Y	Y	C	A	W	R
9501	R	A	R	A	R	R	Y	Y	Y	R	W	G
9502	R	R	R	A	G	G	T	T	C	A	W	G
9504	G	A	R	C	G	G	T	T	C	A	A	A
9511	A	A	G	A	A	A	C	C	T	G	W	G
9512	R	R	R	A	G	G	T	T	C	A	A	G
9506	R	R	R	M	R	R	Y	Y	C	A	W	R
9242	G	A	R	M	R	G	Y	Y	Y	R	T	G
7498	G	A	R	M	R	G	Y	Y	C	R	T	G

For each primer set (bolded and underlined), PCR fragment(s) from each clone were
headings (bolded) correspond to the SNP's nucleotide position in clone 9509's full l
codes are used for heterozygous loci. NA indicates no amplification. A 9 bp INDEL
this INDEL corresponds to an insertion of a 9 bp sequence relative to the 9509 refer
for the insertion relative to the 9509 reference sequence, a "0" indicates the clone is
reference sequence, and "0,~~-~~9" indicates it is heterozygous for the INDEL

~~Clone~~Clone	**Sp02-SNP**										**Sp13-S**	
	99	**184**	**216**	**250**	**268**	**274**	**288**	**299**	**380**	**498**	**111**	**11**
7379	R	R	R	M	R	R	Y	Y	C	A	W	R
9290	G	A	R	M	R	G	Y	Y	C	R	G	T
9503	R	R	R	M	R	R	Y	Y	C	A	W	R
7551	R	R	R	A	G	G	T	T	C	A	A	G
9333	G	A	R	A	R	G	Y	Y	Y	R	A	G
9351	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	A	R
9256	G	A	R	A	R	G	Y	Y	Y	R	G	A
9508	R	A	G	A	A	R	C	C	T	G	T	G
9510	R	A	R	A	R	R	Y	Y	Y	R	W	G
9513	A	R	G	A	R	R	Y	Y	Y	R	A	G
9514	R	A	R	A	R	R	Y	Y	Y	R	A	G
9560	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	A	A
9622	R	R	G	A	R	G	Y	Y	Y	R	T	G

For each primer set (bolded and underlined), PCR fragment(s) from each clone were
headings (bolded) correspond to the SNP's nucleotide position in clone 9509's full l
codes are used for heterozygous loci. NA indicates no amplification. A 9 bp INDEL
this INDEL corresponds to an insertion of a 9 bp sequence relative to the 9509 refer
for the insertion relative to the 9509 reference sequence, a "0" indicates the clone is
reference sequence, and "0,~~.~~9" indicates it is heterozygous for the INDEL

Testing hyper-polymorphic NB-ARC derived markers on *S. intermedia* clones for interspecific genotyping

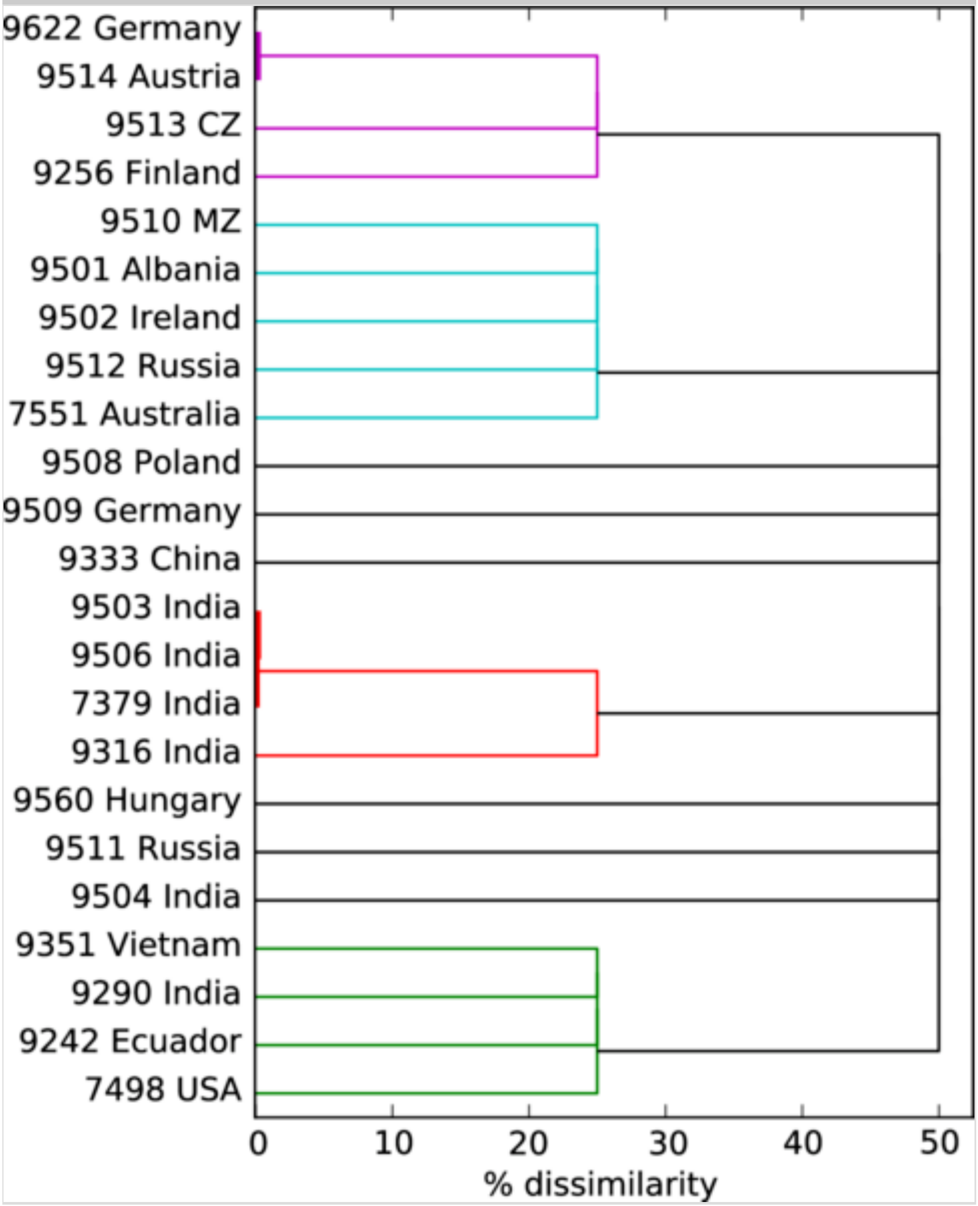
We predicted that the NB-ARC-related markers would also provide interspecific genotyping capability since a greater degree of sequence divergence would be expected when comparing between species. *S. intermedia* has recently been demonstrated by cytogenetic approaches to be closely related in sequence to *S. polyrhiza* (Phuong and Schubert 2017). We tested all four fragment length-based PCR markers on 10 *S. intermedia* clones from the RDSC collection. Only one primer set out of the four, Sp17, amplified *S. intermedia* DNA templates under the PCR conditions used for the *S. polyrhiza* templates (Fig. 5a). The amplified fragment from all ten tested *S. intermedia* clones migrated at a larger apparent size than the fragments observed with *S. polyrhiza* clones using this primer set. This clear and consistent difference in fragment pattern thus provides a simple genotyping tool between the two *Spirodela* species that can obviate any need for DNA sequencing such as those required for plastidic barcodes. Addition of one of the other fragment length primer sets that only amplifies from *S. polyrhiza* templates should provide additional support for the species identification (Fig. 5b). In sum, the combined use of the Sp17 and Sp13 primer sets defined from this study can be deployed as a simple genotyping tool to positively distinguish between the two *Spirodela* species by a simple PCR assay. This will be far superior in ease and economy than previous barcoding or AFLP strategies.

Distance Analysis of *S. polyrhiza* clones using targeted fragment length and SNP polymorphism data

Using only length polymorphism data from four markers, 18 out of 23 *S. polyrhiza* clones displayed unique fingerprints by clustering analysis (Fig. 6). The German 9622 and Austrian 9514 clones share the same fingerprint pattern, while the three Indian clones 9503, 9506 and 7379 were indistinguishable from one another as shown on the dendrogram. Clones within a clustering of ≤ 25% dissimilarity differ from clones outside of the cluster by at least two markers.

Fig. 6

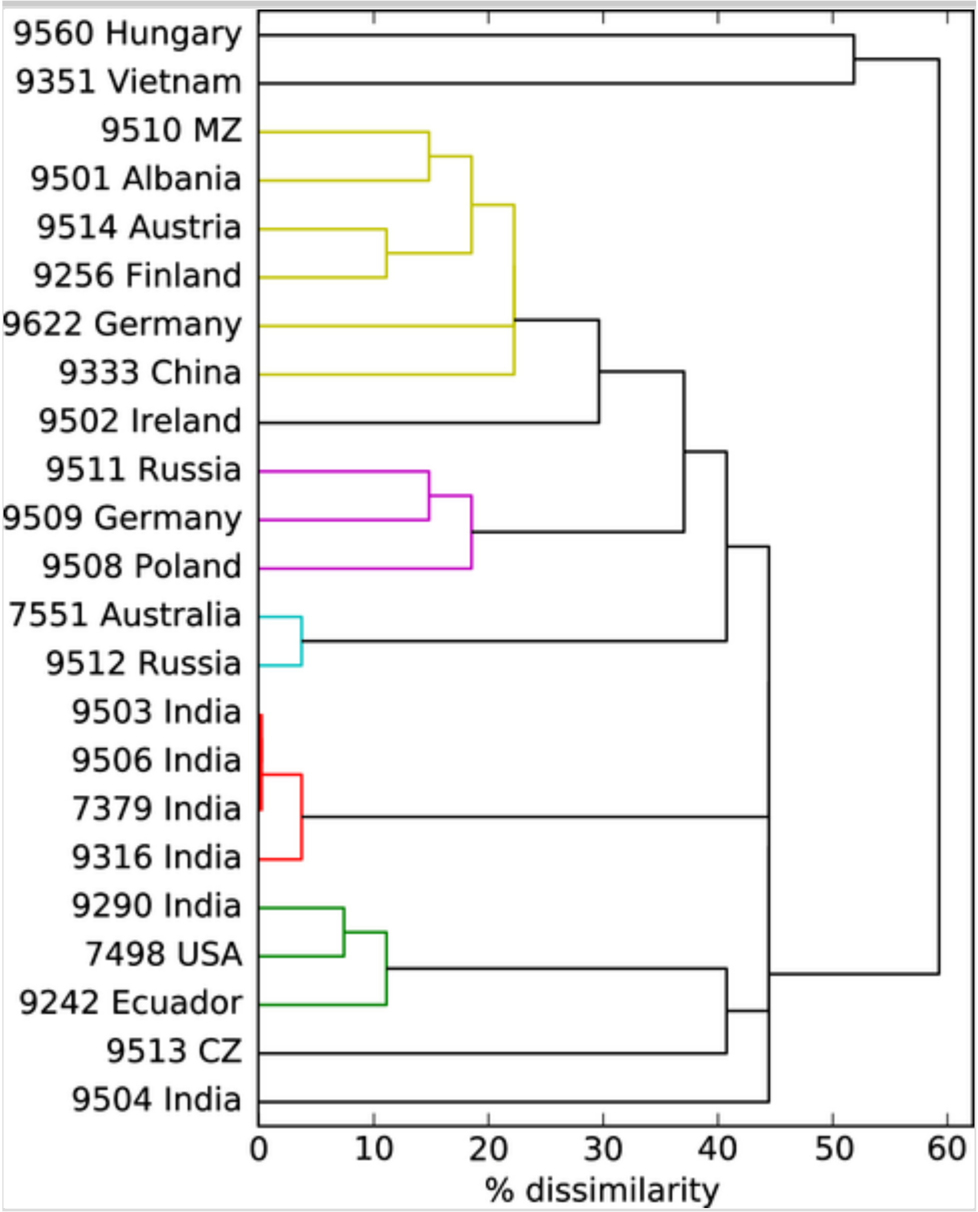
Dendrogram of 23 *S. polyrhiza* clones based on fragment length polymorphism markers. X-axis represents distance between clusters. *S. polyrhiza* clones listed along the Y-axis. Clades formed at 25% or less dissimilarity are similarly colored



The combined SNP and fragment length polymorphism dendrogram further revealed unique fingerprints in 20 out of 23 clones under investigation (Fig. 7). Clones that are separated in clusters by ≥ 25% dissimilarity differ from one another in at least six markers. Surprisingly, although the STY study (Kuehdorf et al. 2014) demonstrated that clones 9506, 9503, and 7379 have distinct quantitative phenotypes, we could not find differences for these three clones using the seven hyper-polymorphic markers, suggesting that they have very low sequence variations.

Fig. 7

Dendrogram of 23 *S. polyrhiza* clones based on SNPs, INDELs, and length polymorphisms from all markers. Distance between clusters on the X-axis, *S. polyrhiza* clones represented along Y-axis. Clones in clusters of less than 25% dissimilarity are similarly colored. These clusters consist of clones that are differentiated from one another by only a few markers



Discussion

To date, the most sensitive molecular method for distinguishing duckweed clones is AFLP (Appenroth et al. 2013), but this technique was met with limited success in a recent study of 24 *S. polyrhiza* clones (Bog et al. 2015). The lack of a sensitive and reliable method to distinguish *S. polyrhiza* clones from one another is an obstacle for the development and application of this interesting and important duckweed species. The recent availability of high quality genome data for this species (Michael et al. 2017) can be leveraged to address this need at the intraspecific level. The advantages of working with a known set of polymorphic loci rather than randomly amplifying sections of the genome, such as in SSR or AFLP approaches, is that background non-informative (hypo-polymorphic) targets can be minimized while the known target loci with high levels of polymorphism can be accentuated. In addition, the demographic data that can be coupled to the particular variations in the different target genes being queried can further inform us of potential biological significance of these loci. Here, we presented a novel informatics-driven and PCR-based pipeline to examine and select length polymorphisms and SNPs from NB-ARC-related loci to generate markers that can discriminate among a set of nine *S. polyrhiza* clones selected from a previous STY study (Kuehdorf et al. 2014) and which have been resequenced. Further analysis with an additional 14 clones illustrated the efficacy of these primer sets to distinguish 20 out of 23 clones. Resequencing the five clones of *S. polyrhiza* that were not resolved by our fragment length markers and include them into our informatics pipeline together with the nine previously sequenced genomes from our training set could further improve the resolving power of our approach. Thus we expect the inclusion of a larger, and potentially more diverse genome dataset from additional clones to our training set for marker selection should lead to more powerful genotyping primer sets with greater resolution power.

We selected and tested seven regions in NB-ARC-related genes on chromosomes 2, 13 and 17 from our first chromosome model for clone 9509 (Michael et al. 2017). The selected windows in this family of genes that resulted from our analysis mostly fell in intronic regions, a good source of sequence and length polymorphisms (Morello and Breviario 2008). Three primer sets were designed within the Sp02G05200 gene, suggesting that this gene is more polymorphic than other NB-ARC-related genes. Initial analysis of Sp02G05200 in the genome of clone 9509 revealed its location in a cluster of three other genes that also have putative disease resistance gene annotations. Its polymorphic nature may be related to more frequent occurrence of micro-rearrangements, gene duplications and recombination events commonly seen in clustered NB-LRR genes (Meyers et al. 2003). While the windows generated from our pipeline were intended to amplify from one locus (mapped uniquely to the 9509 genome by BLASTN), it is nevertheless possible that some of the alleles may come from off-target NB-ARC-related loci, especially if there are more than two amplicons per sample. This is the case for primer sets Sp02a and Sp02b, with Sp02a amplifying an especially high number of fragments (nine different fragments) as compared to the other windows. One possibility for this observation is that NB-LRRs and NB-ARC-related genes are a large gene family and are thus difficult to accurately map short Illumina sequencing reads to these regions. Improvements to the genomic assembly using long-read sequencing technology platforms such as PacBio and Oxford Nanopore may help resolve this issue in the near future. While the genomic origin of the amplified bands may be uncertain with clones that have not been sequenced, they nevertheless serve their fingerprint function as a genotyping tool.

Fragment analysis using an automated sequencer is a more sensitive technology than agarose gel electrophoresis. Since it has a resolution limit of 1 bp, fragment analysis can detect amplicons that otherwise cannot be easily resolved on an agarose gel. However, this technology is also subject to external factors such as temperature and humidity, so results from identical samples may differ by one bp from run-to-run. Additionally, the 1200 bp size limitation of our capillary electrophoresis machine prevented the detection of the approximate 1500 bp fragment amplified in multiple *S. polyrhiza* clones and in all *S. intermedia* clones amplified using the Sp17 primer set (Figs. 3, 5; Table 4). These caveats notwithstanding, it is a rapid and inexpensive technology platform that can provide quantitative fingerprinting results compared to SNP-enabled barcoding approaches.

Indian clones 9316, 9503, and 9504 all originated from Rajasthan, a northern Indian state. Both 9503 and 9504 were collected from the same bird sanctuary in Bharatpur, while 9316 was collected from Ajmer lake, approximately 322 km away. Our length polymorphism data and combined data sets suggest that 9503 is more closely related to 9506 from Hyderabad, in Andhra Pradesh state and 7379 (from Pondicherry, Tamil Nadu), than to clone 9316 from Rajasthan. This is surprising since 9503 and 9316 are both from Rajasthan and thus suggests that they may have arrived to this locale separately and never hybridized. 9504 can be separated from the other Indian clones by at least half of the length polymorphism markers and about half of the combined length polymorphism and SNPs. Despite our markers' ability to distinguish pairs of the Indian clones, the exceptions are 7379, 9506, and 9503 which had a medium–high, medium–low, and low STY, respectively, that cannot be resolved by our markers. This suggests that these three clones may be more similar to each other than compared to the other Indian clones in spite of their apparent distance in locale. One possible explanation that these three clones diverged from the other Indian clones in our analysis may be related to differences in pathogen pressures in their immediate locales; however, testing this hypothesis would require further research. Alternatively, the varying STY traits of these three clones may result from differences at the epigenetic level and thus their respective phenotype could potentially arise from changes independent of DNA sequence per se. The NB-ARC derived markers were trained on nine clones which included 9506 but not 7379 nor 9503 since these have not been resequenced. It is plausible that if 7379 and 9503 were included in the original training set to discover polymorphic NB-ARC-related genes across multiple clones, our analysis pipeline could have identified length polymorphisms or SNPs to distinguish between pairs of 9506, 7379, and 9503 clones if these sequence differences exist.

German clones 9509 and 9622 might have been expected to have similar fingerprints. Clone 9509 was originally collected from Stadtroda and 9622 originated in Baden-Wurttemberg, approximately 426 km apart. However, 9509's fingerprint is more similar to 9508 from Krakau, Poland, while 9622 grouped with 9256 (Uusimaa, Pukila, Finland), and 9514 (Wien, Austria). A possibility is that clones 9509 and 9622 may have spread into Germany via different ancestors originating from different countries in Europe.

Hungarian clone 9560 was successfully amplified from only four of the seven markers tested, and is unique from all other tested clones. A possible explanation for this observation may be its adaptation to local pathogen pressure and potential geographic isolation, stemming from Hungary's mountaineous border regions. Further testing of local clones from Hungary would provide more evidence to support this observation. Clone 9560's barcode was verified using the *psbK–psbI* barcoding marker to ensure that the current observations were not a result of clone mistyping or contamination (data not shown). In addition, the Vietnamese clone 9351 did not amplify using Sp02-SNP, but amplified with the other markers, also suggesting that it may have diverged more from the other tested clones. Lastly, the improved cluster analysis using combined length polymorphism and SNP data suggests that the Russian clone from Moscow (9511) is more similar to European clones from Germany and Poland than to the other Russian clone (9512) originally collected approximately 5200 km east of Moscow and which clustered most closely with a clone from Australia (7551). This further supports our observation that country borders are somewhat arbitrary when it comes to dispersal of this tiny, aquatic plant.

Sp17 was the sole marker that amplified *S. intermedia* clones under these PCR conditions, demonstrating its double utility as an interspecific genotyping marker. Future availability of a *S. intermedia* reference genome can help elucidate possible INDELS or rearrangements that may have occurred to explain our observation. To further probe any intergenus similarities, these primers were also tested on *Landoltia punctata* clones using the same PCR conditions conducted with *S. polyrhiza*. However, the reactions failed to amplify (data not shown), further demonstrating the specificity of the primer sequence and the greater divergence of the NB-ARC-related loci in this species from a separate genus. Data generated from our approach can thus help inform future biogeographical studies aimed at tracking the worldwide dispersion of *S. polyrhiza* clones, its evolutionary history, and divergence of NB-ARC-related genes as they evolve in different locales. We propose that a similar pipeline could be applied to other plant species with genome information of sufficient depth. Furthermore, it is likely that a comparable application of our pipeline to examine non-plant species can also be carried out using similar logic and informatics workflow for the selection of target loci. As an example, the Major Histocompatibility Complex in the human genome may also be used for high-sensitivity genotyping marker identification since it is also known to have a higher than average recombination rate compared to other nuclear loci (Sommer 2005).

Acknowledgements

This work was supported by the National Science Foundation Integrative Graduate Research and Education Traineeship Renewable and Sustainable Fuel Solutions for the twenty-first Century [DGE-0903675]; and Hatch grant from the New Jersey Agricultural Experimental Station at Rutgers University [project #12116]. We would like to thank Ryan Gutierrez and Christine Kubik for technical support, Csanad Gurdon for helpful advice, and members of the Lam lab for their feedback and support. We especially want to acknowledge Nikolai Borisjuk (presently working in Huaian, China) for his initial work at the Lam lab in the very beginning of this project. His early mentorship of P.C. in this endeavor is gratefully acknowledged.

Author Contributions

PC, GW, and EL conceived and designed the experiments, and wrote the manuscript. TPM carried out sequence analysis of the resequenced *Spirodela polyrhiza* genomes and produced the VCF files for the bioinformatics pipeline. GW and PC wrote the computer code for the bioinformatics analysis. PC performed all PCR experiments. JV and JH ran samples on the capillary electrophoresis DNA sequencer and collected fragment length data. All authors provided input to the manuscript and agreed to its final version for submission.

References

Appenroth KJ, Borisjuk N, Lam E (2013) Telling duckweed apart: genotyping technologies for the Lemnaceae. Chin J Appl Environ Biol 19(1):1–10

Appenroth KJ, Sree KS, Fakhoorian T, Lam E (2015) Resurgence of duckweed research and applications: report from the 3rd International

Duckweed Conference. Plant Mol Biol 89(6):647–654

Bog M, Baumbach H, Schween U, Hellwig F, Landolt E, Appenroth KJ (2010) Genetic structure of the genus *Lemna* L. (Lemnaceae) as revealed by amplified fragment length polymorphism. Planta 232:609–619. https://doi.org/10.1007/s00425-010-1201-2

Bog M, Schneider P, Hellwig F, Sachse S, Kochieva EZ, Martyrosian E, Landolt E, Appenroth K (2013) Genetic characterization and barcoding of taxa in the genus *Wolffia* Horkel ex Schled. (Lemnaceae) as revealed by two plastidic markers and amplified fragment length polymorphism (AFLP). Planta 237:1–13. https://doi.org/10.1007/s00425-012-1777-9

Bog M, Lautenschlager U, Landrock MF, Landolt E, Fuchs J, Sree KJ, Oberprieler C, Appenroth KJ (2015) Genetic characterization and barcoding of taxa in the genera *Landoltia* and *Spirodela* (Lemnaceae) by three plastidic markers and amplified fragment length polymorphism (AFLP). Hydrobiologia 749:169–182. https://doi.org/10.1007/s10750-014-2163-3

Borisjuk N, Chu P, Gutierrez R, Zhang H, Acosta K, Friesen N, Sree KS, Garcia C, Appenroth KJ, Lam E (2015) Assessment, validation and deployment strategy of a two-barcode protocol for facile genotyping of duckweed species. Plant Biol 17:42–49

Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am J Hum Genet 81:1084–1097

Cingolani P, Platts A, Wang le L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) 6:80–92. https://doi.org/10.4161/fly.19695

Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, Warthmann N, Gu TT, Fu G, Hinds DA et al (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. Science 317:338–342

Dolger K, Tirlapur UK, Appenroth K (1997) Phytochrome-regulated starch degradation in germinating turions of *Spirodela polyrhiza*. Photochem Photobiol 66(1):124–127

Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 19:11–15

Eddy SR (1998) Profile hidden Markov models. Bioinformatics 14:755–763

Feng B, Fang Y, Xu Z, Xiang C, Zhou C, Jiang F, Wang T, Zhao H (2017) Development of a New Marker System for Identification of *Spirodela polyrhiza* and *Landoltia punctata*. Int J Genom. https://doi.org/10.1155/2017/5196763

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A (2016) The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res Database Issue 44:D279–D285

Gan X, Stegle O, Behr J, Steffen J, Drewe P, Hildebrand K, Lyngsoe R, Schultheiss SJ, Osborne EJ, Sreedharan KL et al (2011) Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. Nature 477:419–423. https://doi.org/10.1038/nature10414

Hammond-Kosack KE, Jones JDG (1997) Plant disease resistance genes. Annu Rev Plant Physiol Plant Mol Biol 48:575–607

Hill JT, Demarest BL, Bisgrove BW, Su YC, Smith M, Yost HJ (2014) Poly Peak Parser: method and software for identification of unknown indels using Sanger Sequencing of PCR products. Dev Dyn 243(12):1632–1636

Jones E, Oliphant T, Peterson P (2001) SciPy: open source scientific tools for Python. http://www.scipy.org/

Kuehdorf K, Jetschke G, Ballani L, Appenroth KJ (2014) The clonal dependence of turion formation in the duckweed *Spirodela polyrhiza*—an ecogeographical approach. Physiol Plant 150(1):46–5424

Lam E, Appenroth KJ, Michael T, Mori K, Fakhoorian T (2014) Duckweed in bloom: the 2nd international conference on duckweed research and applications heralds the return of a plant model for plant biology. Plant Mol Biol 84(6):737–742

Les DH, Crawford DJ, Landolt E, Gabel JD, Kimball RT (2002) Phylogeny and systematics of Lemnaceae, the duckweed family. Syst Bot 27(2):221–240

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

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303. https://doi.org/10.1101/gr.107524.110

Meyers BC, Kozik A, Griego A, Kuang H, Michelmore RW (2003) Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. Plant Cell 15:809–834

Michael TP, Bryant D, Gutierrez R, Borisjuk N, Chu P, Zhang H, Xia J, Zhou J, Peng H, El Baidouri M et al (2017) Comprehensive definition of genome features in *Spirodela polyrhiza* by high-depth physical mapping and short-read DNA sequencing strategies. Plant J 89:617–635

Morello L, Breviario D (2008) Plant spliceosomal introns: not only cut and paste. Curr Genom 9:227–238

Phuong TNH, Schubert I (2017) Reconstruction of chromosome rearrangements between the two most ancestral duckweed species *Spirodela polyrhiza*. and *S. intermedia*. Chromosoma 126(6):729–739

Schuelke M (2000) An economic method for the fluorescent labeling of PCR fragments. Nat Biotechnol 18:233–234

Sommer S (2005) The importance of immune gene variability (MHC) in evolutionary ecology and conservation. Front Zool. https://doi.org/10.1186/1742-9994-2-16

The 1001 Genomes Consortium (2016) 1135 Genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell 166:481–491. https://doi.org/10.1016/j.cell.2016.05.063

Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG (2012) Primer3—new capabilities and interfaces. Nucleic Acids Res 40(15):e115

Vaughan D, Baker RG (1994) Influence of nutrients on the development of gibbosity in fronds of the duckweed *Lemna gibba* L. J Exp Bot 45:129–133

Wang W, Messing J (2012) Analysis of ADP-glucose pyrophosphorylase expression during turion formation induced by abscisic acid in *Spirodela polyrhiza* (greater duckweed). BMC Plant Biol 12:5. https://doi.org/10.1186/1471-2229-12-5

Wang W, Wu Y, Ermakova M, Kerstetter R, Messing J (2010) DNA barcoding of the Lemnaceae, a family of aquatic monocots. BMC Plant Biol 10:205. https://doi.org/10.1186/1471-2229-10-205

Wang W, Haberer G, Gundlach H, Glaesser C, Nussbaumer T, Luo MC, Lomsadze A, Borodovsky M, Kerstetter RA, Shanklin J et al (2014) The *Spirodela polyrhiza* genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. Nat Commun 5:3311. https://doi.org/10.1038/ncomms4311

Xu J, Cui W, Cheng JJ, Stomp AM (2011) Production of high-starch duckweed and its conversion to bioethanol. Biosyst Eng 110:67–72

Xu J, Cheng JJ, Stomp AM (2012) Growing *Spirodela polyrrhiza* in swine wastewater for the production of animal feed and fuel ethanol: a pilot study. Clean Soil Air Water 40:760–765