TECHNOLOGIES FOR TRACKING GENOME VARIATIONS IN DUCKWEED AND OTHER SPECIES

BY

PHILOMENA N. CHU

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Plant Biology

Written under the direction of

Eric Lam

And approved by

_____

_____

_____

_____

New Brunswick, New Jersey

October, 2018

ABSTRACT OF THE DISSERTATION

Technologies for tracking genome variations in duckweed and other species

by PHILOMENA N. CHU

Dissertation Director:

Eric Lam

Near-term and future increases in global population overburden our energy resources as we seek solutions to unstable foreign petroleum pricing and its associated problems with climate change and ecological biosphere integrity. Renewable and sustainable biofuels have the potential to replace existing transportation fuels and dampen their related problems.

Small but versatile, duckweed is an aquatic plant that can be used in a variety of applications such as biofuels, animal feed, and wastewater remediation. This dissertation addresses the need for quick and reliable typing of duckweed species and lower taxonomic levels. First, a sequence database for two plastidic barcodes, *atpF-atpH* and *psbK-psbI* was created representing all 37 known duckweed species. Using a BLAST-based protocol, our approach can distinguish 30 out of the 37 species. To distinguish clones of the duckweed species *Spirodela polyrhiza*, a bioinformatics pipeline was developed that identifies hyperpolymorphic regions of the NB-LRR-based plant disease resistance protein-encoding genes, which can be used as genotyping markers. We demonstrate that a combination of seven hyperpolymorphic regions from six loci using fragment analysis and Sanger sequencing post-PCR can distinguish 20 out of the 23 *S. polyrhiza* clones tested. A subset of these markers can be used to clearly separate *S. polyrhiza* clones from the closely related

S. intermedia. Finally, our bioinformatics pipeline was applied to *Arabidopsis thaliana* to locate NB-LRR markers that can computationally distinguish all 1,135 *A. thaliana* accessions, and to validate the efficacy of the pipeline at identifying hyperpolymorphic genic regions. This novel method can be used to track accessions for a species of interest using polymorphisms from sequenced genomes, in addition to assisting a better understanding of the differential frequency of mutations across the genome.

Acknowledgements

I'd like to thank my advisor and committee members. I am grateful for funding from the NSF-IGERT Renewable and Sustainable Fuel Solutions for the 21st Century project for a traineeship, the Division of Life Sciences at Rutgers University for teaching assistantships, and the Eagleton Institute of Politics for the Henry J. Raimondo Legislative fellowship. Thank you to Linda Anthony, our first IGERT program coordinator. Additional thanks to current and former lab members, especially Peter, Aniça, Ryan, Chia-Hui, Sarah, and Kenny. I appreciate the helpful discussions with my fellow graduate students: Csanad, Anna, YeeChen, Gregx2, Sam, Debaleena, and Ari.

Chapter 2 is previously published: Borisjuk, N.*, Chu, P.*, Gutierrez, R., Zhang, H., Acosta, K., Friesen, N., Sree, K., Garcia, C., Appenroth, K., and Lam, E. (2015). Assessment, validation and deployment strategy of a two-barcode protocol for facile genotyping of duckweed species. *Plant Biol.*, 17:4249. doi:10.1111/plb.12229. *co-first authors. P.C. contributed to the design of the experiments, assembled the barcode sequences into a duckweed BLAST database and performed associated statistical analyses with guidance from E.L. P.C. contributed to the writing of the article.

Chapter 3 has been submitted for publication: Chu, P., Wilson, G.M., Vaiciunas, J., Honig, J., and Lam, E. Sequence-guided approach to genotyping plant clones and species using polymorphic NB-ARC-related genes. P.C., G.W., and E.L. designed the experiments and wrote the manuscript. P.C. and G.W. collaborated on the bioinformatics analysis. P.C. performed all PCR experiments.

In Chapter 4, P.C. designed the experiments and performed the analyses in collaboration with G.W. and E.L. P.C. wrote the manuscript.

Dedication

To my mom and dad.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Ethnobotanical history of transportation fuels

In the 1850s, Dr. Benjamin Silliman, Jr., a professor of chemistry at Yale University, was commissioned to perform a chemical analysis on a sample of "rock oil" collected from northwestern Pennsylvania. He concluded that this oil, or "petroleum," could be separated into distinct hydrocarbon fractions. One such fraction, gasoline, had no apparent application at the time of its discovery and was often discarded as a useless by-product. Two key innovations galvanized the transformation of gasoline into the economically important global resource it is today. First, the internal combustion engine was invented. Although many people contributed technological improvements, major advancements were made in 1885–1886 by Gottlieb Daimler and Karl Benz who developed in parallel the first automobile containing an internal combustion engine that ran on gasoline [Barnard, 2001]. The second innovation came in 1908 when Henry Ford began selling the mass-produced Model T automobile. With cheap gasoline and an affordable price tag, the internal combustion vehicle became widely adopted. As a result, gasoline consumption has steadily risen over the last century. In 2015 alone, the U.S. consumed a daily average of 384.74 million gallons [US Energy Information Administration, 2016a]. But the boom that encouraged more cars on the road came with unforeseen consequences. By the middle of the 20th century, toxic smog from vehicle emissions was becoming a major environmental and public health problem nationally. State and federal governments struggled to develop a comprehensive strategy to address the issue. Finally in 1970, Congress created the Environmental Protection Agency (EPA) and authorized it to carry out the Clean Air Act (CAA) passed that same year. Twenty years later, Congress amended the CAA with stricter regulations on air pollution from stationary and mobile sources. The

1990 amendments also promoted the expansion of cleaner transportation fuels, such as ethanol. Compared to gasoline and diesel, these alternative fuels could be generated by more renewable and domestic means, helping to strengthen the country's energy security.

1.2 Ethnobotanical use

The transportation fuel powering most cars on the road is a blend of petroleum-derived gasoline mixed with 10% ethanol, although 15% and higher ethanol blends are becoming increasingly widespread. As part of the Energy Policy Act of 2005 and the Energy Independence and Security Act of 2007, the Renewable Fuel Standard aims to shift our unsustainable consumption and reliance on petroleum products to renewable fuels, such as those derived from plant biomass. Ethanol (C_2H_5OH) is regarded as a high performance renewable fuel that boosts engine performance and lowers greenhouse gas emissions. The United States has consistently been the world's largest producer of ethanol for the past several years, with a 2014 output of over 14 billion gallons, as well as its largest consumer [US Energy Information Administration, 2016b]. Current methods of ethanol production from corn, the main biofuel feedstock in the U.S., are not sustainable. Corn has many pitfalls as an energy source. It requires intensive fertilizer and freshwater inputs and it competes with other food crops for fertile land. The most expensive cost of starch-based ethanol production, and most variable, is the price of corn [Rismiller and Tyner, 2009, McAloon et al., 2000]. Although the overall trend of corn prices has been relatively stable over the past few years, climatic and environmental events can cause dramatic fluctuations. For example, a severe drought in 2012 caused the price of corn to sharply increase. By diversifying the crops used to make ethanol and encouraging the development of more environmentally friendly ones, the U.S. can significantly bolster its energy security. A promising alternative to corn is duckweed, which has been garnering more attention as a renewable and sustainable biofuel feedstock in recent years. Duckweed is a miniature, aquatic basal monocot that floats on stationary to slow-moving freshwater. Plants get dispersed by waterfowl, promoting its widespread distribution. Located on every continent except Antarctica, duckweeds have adapted to a wide range of environments. There are 37 species of duckweed and countless clones, or natural strains, which

provide a diverse gene pool and a plethora of biochemical properties that can be exploited for commercial use. In addition, its simple body architecture and high surface area-to-volume ratio help duckweed achieve an extremely fast doubling time of approximately 24 hours to 3 days [Landolt, 1986].

Duckweed can be used in a vast range of applications, including biofuels, fish and animal feed, as well as water quality monitoring and wastewater remediation. There are 37 species of duckweed from five genera (Spirodela, Lemna, Wolffia, Wolffiella, Landoltia) distributed globally with numerous clones of each. The large geographical diversity of duckweed plants makes it possible to find an ecotype that is adapted to diverse environments. Species and even ecotypes of the same species can vary dramatically in their physiological characteristics, some of which are desirable for certain applications. For example, high starch content is useful for ethanol production for biofuels. One such duckweed species, *Spirodela polyrhiza*, has demonstrated a potential to be an environmentally-friendly biofuel feedstock. A pilot-scale study growing *S. polyrhiza* on dilute swine wastewater conducted at North Carolina State University [Xu et al., 2011b] measured a growth rate at $12.4\,\mathrm{g/(m^2\,d)}$ of dry weight and a high starch content (31% dry weight) which can be theoretically converted to an ethanol yield of $6.42 \times 10^3\,\mathrm{L/ha}$, approximately 50% higher than that achieved with corn. Our lab is interested in exploring the ability of *S. polyrhiza* clones to accumulate varying amounts of starch when grown on different sources of wastewater. These are favorable characteristics when searching for the ideal sustainable biofuel feedstock candidate.

Once a suitable clone is found for a particular application, maintaining a pure culture is often necessary to ensure consistent results. Open ponds, commonly used in duckweed cultivation, are susceptible to contamination as waterfowl and other animals can introduce other duckweed clones. Duckweeds are relatively easy to distinguish morphologically at the genus level but are considerably more difficult to discriminate at lower taxonomic levels. Many species, and especially clones, are morphologically similar or identical. Thus a reliable, and preferably facile, method of clone identification would be required for the deployment of many duckweed-based applications, especially at the industrial scale.

1.2.1 Overview of this dissertation

I detail my scientific contribution to help create an environment for duckweed-based research and applications throughout this dissertation. First, I address the need for reliably genotyping duckweed species in Chapter 2. In Chapter 3, I focus on genotyping *Spirodela polyrhiza* for its utility as a popular duckweed model organism and its application as a biofuel feedstock. I develop a bioinformatics approach to genotype *S. polyrhiza* clones using disease resistance-related genes as molecular markers. Then to evaluate the ranking algorithms developed in Chapter 3 and demonstrate their universal applicability to other organisms with rich genome resources, I test the approach on the model plant organism, *Arabidopsis thaliana*, in Chapter 4.

Chapter 2

Assessment, validation and deployment strategy of a two-barcode protocol for facile genotyping of duckweed species

Authors: Borisjuk, N.*, Chu, P.*, Gutierrez, R., Zhang, H., Acosta, K., Friesen, N., Sree, K.S., Garcia, C., Appenroth, K.J., and Lam, E.

*these authors contributed equally to this work

Published in *Plant Biology*, [Borisjuk et al., 2015].

2.1 Introduction

Duckweeds, aquatic plants of the Lemnaceae family, are monocots that have the remarkable ability to adapt to a wide range of climatic conditions without overt changes in their morphologies. As the world's smallest angiosperms that can also grow at the fastest rate [Ziegler et al., 2015], they can be found in most parts of the planet that have significant amounts of stable water bodies such as lakes and ponds [Landolt, 1986]. The systematic collection and typing of several thousand duckweed strains for more than 50 years by Elias Landolt has provided an incomparable resource to the community. Dr. Landolt's generous attitude in sharing this resource and his knowledge on duckweed also facilitated the development of systematics and biogeographical studies of this plant family. It is thus with deep gratitude that we dedicate this report to mark Dr. Landolt's passing in 2013 after his long illness. We believe that the subject matter of duckweed species identification via genotyping befits this tribute since the collection of strains used to establish a reference sequence dataset was a foundation laid by Dr. Landolt himself.

Until the recent decade, typing of duckweed species was carried out using morphological and physiological characteristics. Later, those descriptive methods have been

supplemented by allozyme and chemical analyses that provided additional supporting evidence for species identification [McClure and Alston, 1966, Les and Sheridan, 1990, Les et al., 1997]. However, these approaches are often time-consuming and not always quantitative. The abbreviated architecture of many of the duckweed species leads to a limited number of distinguishing features that are easily scored. The application of molecular approaches for a unified protocol of species identification is thus a promising alternative. Since a review of genotyping technology development for duckweeds has already been published recently [Appenroth et al., 2013], we will not elaborate too much here on the techniques that have been reported prior to 2013. While several different types of strategies such as Random Amplified Polymorphic DNA (RAPD), Inter-Simple Sequence Repeats (ISSR) and Amplified Fragment Length Polymorphism (AFLP) have been developed and used with duckweed, the AFLP technique in particular has been shown to be capable of resolving interspecies as well as intraspecies genotypes for species in the Lemna [Bog et al., 2010] and Wolffia [Bog et al., 2013] genera. Nevertheless, these nuclear genome-based techniques can be laborious, and the results vary depending on the assay conditions as well as sample quality. However, integrative analysis of these datasets may also be challenging since the relevant software and database resources are not widely distributed. An alternative strategy exploits the variable regions of the highly conserved plastid genome to create robust barcodes as a simple tool for species identification [Chase et al., 2005, Fazekas et al., 2008, Lahaye et al., 2008b]. Two plastidic regions (*rpl16* and *rps16*) were used to investigate more than 50 clones representing all 11 species of the genus *Wolffia* Horkel [Bog et al., 2013] and six of these species could be accurately genotyped. For higher plants, the CBOL working group [Group, 2009] has screened a large number of plant species and recommended 7 plastid-encoded sequences as candidates for genotyping barcodes. Several of these were used in a study by Wang et al. [Wang et al., 2010] to create the first large reference sequence database using 97 accessions, representing 31 species, from clones of the Landolt collection. While this study showed that two barcoding regions in particular—*atpF-atpH* and *psbK-psbI*—have sufficient variations among many of the species to allow their unambiguous identification, the analysis was incomplete since 6 out of the 37 duckweed species were not represented and a single accession was used for several

species. Ultimately, sequence data from 20 out of the 31 sampled species were used for relational analysis between species of the family [Wang et al., 2010]. In the present study, we aim to provide a more complete database of these two barcodes for the Lemnaceae family by including additional reference sequences for the 6 species missing in the previous work [Wang et al., 2010]. In addition, we systematically determined the BLAST score thresholds for species distinction using a reference sequence library of these two barcodes compiled from over 300 publicly available sequences. In an effort to accurately evaluate the method as well as to resolve apparent discrepancies, additional accessions for 3 other species were also studied and their barcode sequences added to the reference database. For species that are difficult to resolve based on these two plastome-encoded barcodes using current resources, future strategies to overcome the difficulties are discussed.

2.2 Materials and Methods

2.2.1 Plant Material

Most duckweed accessions analyzed in this study were sampled from the Rutgers Duckweed Stock Cooperative (RDSC; `http://www.ruduckweed.org`) collection. The strains were maintained on 0.8% agar containing 0.5X Schenk and Hildebrandt (SH) salts (Sigma-Alderich) and 0.5% sucrose, pH 5.7–6.0, at 15 °C under axenic conditions. To bulk up tissue for DNA preparation, plants were propagated in liquid 0.5X SH medium with 0.5% sucrose at 25 °C for 5–7 days at 22 °C. In total, 18 ecotypes/accessions representing three *Lemna* and three *Wolffiella* species were initially used to complete the set of reference barcode sequences representing all available species of duckweed. Seven accessions of *L. japonica* and *W. borealis* were subsequently added to improve the duckweed reference sequence database for the two barcodes. Finally, seven newly isolated *W. microscopica* clones [Sree and Appenroth, 2014] were included in the final stages of the analysis as it became clear that the accession of *W. microscopica* used in the study of Wang et al. [Wang et al., 2010] is likely to be *W. elongata* instead (Personal communication of E. Landolt). To assess the utility of this approach, 19 accessions from the RDSC collection that were not previously typed were also analyzed in this study.

2.2.2 DNA extraction, fragment amplification, sequencing and alignment

Total DNA was extracted from plant tissue using a modified CTAB method [Murray and Thompson, 1980] with the following modifications: all extraction steps were performed in microcentrifuge tubes, and the CsCl ultracentrifugation step was omitted. Additionally, we investigated the utility of a simplified DNA extraction protocol [Klimyuk et al., 1993], which can be completed in approximately 30 minutes and lacks the phenol or chloroform extraction steps. Both methods used in this study gave satisfactory results for PCR amplification and downstream sequencing, although the CTAB method produced sequencing results of higher quality. For the simplified protocol, it should be noted that intact fronds of *Lemna, Spirodella, Landoltia* and *Wolffiella* species can be directly used for DNA extraction, whereas the hard, compact structure of *Wolffia* tissue necessitates slight mechanical disruption.

PCR amplification reactions were carried out as recommended by the CBOL Plant Working Group [Group, 2009], described in Wang et al. [Wang et al., 2010], using the following primers:

5'-`TTAGCATTTGTTTGGCAAG`-3' and
5'-`AAAGTTTGAGAGTAAGCAT`-3' for the *psbK-psbI* barcode;
5'-`ACTCGCACACACTCCCTTTCC`-3' and
5'-`GCTTTTATGGAAGCTTTAACAAT`-3' for the *atpF-atpH* barcode.

Following amplification, the DNA fragments were purified using the QIAEX II Gel Extraction Kit (Qiagen) or ExoSAP-IT (Affymetrix) and sent to Genewiz (South Plainfield, NJ, USA) or GenScript (Piscataway, NJ, USA) for sequencing. The raw sequences were preliminary optimized using the "Online Analysis Tools" package (`http://molbiol-tools.ca`). Multiple DNA sequence alignments were generated by ClustalW software [Thompson et al., 1997], and the alignment was subsequently corrected manually in MEGA 5 [Tamura et al., 2011].

For BLAST alignment analyses, a duckweed reference barcode set was compiled from *atpF-atpH* and *psbK-psbI* barcode sequences that were generated in this study [Borisjuk et al., 2015, Table S7] and those that were available from the NCBI database as of January 2014. The reference set was deposited into a searchable database: `http://epigenome.`

rutgers.edu/cgi-bin/duckweed/blast.cgi. Queried sequences were trimmed to include only intergenic regions and used in BLASTN (version 2.2.26+) searches to identify homologies to other barcode sequences in the set. Top hits for each query are presented in [Borisjuk et al., 2015, Supplemental Tables].

2.2.3 Phylogenetic analysis

Combined data sets (*psbK-psbI* and *atpF-atpH* spacers) were analysed by Fitch parsimony with the heuristic search option in PAUP (version 4.0b10) [Swofford, 2002] with MULTREES, TBR branch swapping, and 100 replicates of random taxon addition. The evolutionary direction of sequence changes was inferred by outgroups comparison. The consistency index (CI) of Kluge and Farris [Kluge and Farris, 1969] is presented to estimate the amount of homoplasy in the characters. Parsimony trees with equal length were summarized by the strict consensus method. Bootstrap analyses (100 replicates) were performed to assess the relative support of the clades. Bayesian analyses were implemented with MrBayes 3.1.23 [Ronquist and Huelsenbeck, 2003]. Sequence evolution models were evaluated using the Akaike Information Criterion (AIC) with the aid of Modeltest 3.7 [Posada and Crandall, 1998]. Two independent runs for each of eight chains, 10 million generations, and sampling every 100 trees were carried out. 25% of the initial trees were discarded as burn-in. The remaining 25,000 trees were combined into a single data set and a majority-rule consensus tree was obtained. Bayesian posterior probabilities were calculated for that tree in MrBayes 3.1.23.

2.3 Results

2.3.1 Completion and analysis of the duckweed species barcode database with two plastome-encoded intergenic sequences

To complete the reference barcode sequence database for all 37 known species of duckweed, we performed PCR amplification of the two plastidic intergenic spacers (*atpF-atpH* and *psbK-psbI*) with total DNA isolated from the 6 species amongst the Lemna and Wolffiella genera that were missing in the database: *L. perpusilla*, *L. tenera*, *L. yungensis*, *W.*

caudata, *W. welwitschii*, and *W. repanda*. Two to five accessions per species were used [Borisjuk et al., 2015, Table S1]. Sequences for the amplified DNA were combined with those previously reported for the other 31 species earlier by Wang et al. [Wang et al., 2010] for phylogenetic analysis of all known species in the Lemnaceae family. Since the separate analyses of *atpF-atpH* and *psbK-psbI* spacers showed no incongruent results, we therefore combined them in a joint analysis of 111 taxa, including Colocasia esculenta as an outgroup species. The combined data matrix included 1373 characters divided in two partitions: 1–684 for *atpF-atpH* and 685–1373 for *psbK-psbI*, of which 684 were constant, 142 variable characters were parsimony uninformative and 547 were parsimony informative. In this analysis, only the *atpF-atpH* spacer was used for the *Spirodela* clade. Parsimony and Bayesian analyses yielded the same topology but with lower bootstrap percentages (BP) than posterior probabilities (PP). The heuristic search found most-parsimonious trees that were 1,371 steps long (consistency index 0.7141, retention index 0.9539). The resultant dendrogram from this analysis is shown in Figure 2.1, with the species and accessions analyzed in this work shown in bold. The six *Lemna* and *Wolffiella* species that were analyzed with the two plastidic intergenic barcodes segregated into the respective *Lemna* and *Wolffiella* clades of this family. In comparison to the previously published phylogenetic analysis of *Lemna* species via AFLP [Bog et al., 2010], the placement of the species *L. perpusilla*, *L. yungensis* and *L. tenera* is also largely in agreement. Thus, *L. perpusilla* is most closely related to *L. aequinoctialis* while *L. yungensis* is very closely related to *L. valdiviana* in some cases. The relationship of *L. tenera* is somewhat less clear since it is placed closer to the *L. aequinoctialis* group in our present study whereas in the previous AFLP study, it appeared to be closer to *L. trisulca*. For the Wolffiella species, while they are clearly segregated into the Wolffiella clade and are largely grouped into distinct subgroups, some of the accessions appeared to be grouped separately with other species. This is observed for *W. repanda* 9122 and *W. welwitschii* 9381, which are structurally more related to the *W. lingulata* group. Similarly, *W. caudata* 9139 is found to be closer based on these barcode sequences to *W. oblonga* and *W. repanda*. This situation has in fact been observed previously for *S. polyrhiza* 9203 and *W. gladiata* 8350 where multiple accessions of these species were analyzed [Wang et al.,

2010]. While these "outliers" may result from hybridization between related species, a more trivial cause may be the result of mistakes in morphological typing or mislabeling during culture maintenance over the years.

2.3.2 BLAST-based protocol for duckweed species identification using two barcodes: assessment of utility for each species and identification of problematic assignments

Although the phylogenetic analysis presented in Fig. 2.1 is a useful tool to gain a global view of the relationship between all species within a family, we seek a simpler tool to specifically query the relatedness of a particular accession of duckweed, such as a newly isolated strain, to all species for which representative barcode sequences are available. A facile and widely available bioinformatics tool that can provide quantitative descriptions of relatedness between nucleotide sequences is BLAST (Basic Local Alignment Search Tool). With the publicly available NCBI database for reported DNA sequences, the nucleotide BLAST (BLASTN) application can be used to query barcode sequences generated from genomic DNA samples of a duckweed strain to a standardized database. For this project, we first downloaded all available *atpF-atpH* and *psbK-psbI* intergenic barcode sequences from duckweed that were available in the NCBI database in January 2014 at the end of 2013. Together with the barcode sequences that we obtained for the six species described above, and several other sources, a total of 313 barcode sequences (117 for *psbK-psbI* and 196 for *atpF-atpH*) from 198 strains of duckweed were collected into a local database that we created on our own computer server [Borisjuk et al., 2015, Table S1]. For many strains, only the *atpF-atpH* barcode sequence was available, thus resulting in a lower number of total barcode sequences than the 396 that could be produced for 198 strains. These sequences were then used as the reference database for sequential BLAST analyses [Korf et al., 2003] using selected barcode sequences from each of the 37 species as query. The data from these BLAST analyses are summarized in Tables 2.1 and 2.2, and the more detailed BLAST scores and cut-offs can be found in [Borisjuk et al., 2015, Tables S2, S3].

Using the BLAST values for percent identity and bit score obtained from the systematic comparison of sequences from each species of duckweed to the total reference set, we

assessed the potential for each barcode as a tool for positive identification (ID) of that species from the other 36. Percent identity designates the level of sequence identity within a stretch of sequences that can be aligned by BLASTN whereas bit score is a log-scaled value related to the probability of finding such a match by chance. Thus, bit score depends both on sequence identity and the length of the aligned sequence but is independent of the total number of sequences in the database. By comparing the intra- and inter-species variations of these two values [Borisjuk et al., 2015, Table S4], we empirically arrived at the following cutoff values for species ID with good confidence: a difference in identity of 2% and a difference in bit score of 40. Thus for each barcode, species that showed values equal to or greater than these cutoffs for interspecies differences (shown by broken lines in [Borisjuk et al., 2015, Tables S2 and S3] would be scored as "good targets for confident ID" (✓), whereas species that failed to reach these cutoffs would be designated as either "potentially useful but use with caution" (±) or "insufficient for positive ID" (X). Using these criteria with the current reference database in NCBI, we found that 25 species of duckweed can be positively identified by one or both of the two barcodes studied here. For 5 species, one or both barcodes can potentially provide species ID but the data will need to be assessed with caution to make a final conclusion. For example, *L. perpusilla* is found to be distinguishable from its most closely related species *L. aequinoctialis* with either barcode studied here. However, with the *psbK-psbI* barcode, the two *L. perpusilla* strains are identical and thus not very informative of the full intraspecies variation, and show 99.77% (432/433) identity with *L. aequinoctialis* 7126 [Borisjuk et al., 2015, Table S2]. However, the length of alignment (433 bp) is clearly shorter than the actual barcode sequence from this strain of *L. aequinoctialis* (472 bp), thus generating a large bit score difference of 83 between the two species. In contrast, for the *atpF-atpH* barcode, the percent identity and bit score were generated from alignment over the whole length of the barcodes (420 bases), and the interspecies difference for these two values—0.95% and 24, respectively—are too low for confident ID of *L. perpusilla* using this barcode. In this case, sequencing of more strains of *L. perpusilla* and *L. aequinoctialis* in the future may help to determine whether the *psbK-psbI* barcode could be used with more confidence for *L. perpusilla* identification.

For four species, *L. minuta*, *L. valdiviana*, *W. lingulata* and *W. globosa*, data from our BLAST searches clearly indicated that neither barcode is able to resolve the intraspecies and interspecies variations. Thus other barcodes or strategies, such as AFLP, may be required to create a molecular method for their positive ID. Of the species for which only a single accession was present in our reference database, three strains—*L. japonica* 7182, *W. borealis* 9123 and *W. microscopica* 9276—produced results that raised doubts about their original species ID. This conclusion is made based on two key observations: (1) the barcodes for these strains that were present in our database comprised of a single accession, and the sequence for the two barcodes are nearly, or completely identical to those found in strains of *L. minor*, *W. globosa*, or *W. elongata*, respectively. (2) Based on the AFLP analyses of other accessions of *L. japonica* (9250, 6728 and 8695 from [Bog et al., 2010]), as well as those of *W. borealis* and *W. microscopica* (9123, 9143, 8359 and 9276 from [Bog et al., 2013]), these three species should be quite distinct from the others and are not closely related to *L. minor*, *W. globosa*, or *W. elongata* respectively. It should be pointed out that the *W. microscopica* 9276 used in the AFLP work of Bog et al. [Bog et al., 2013] could be distinct from the strain used in the barcoding study of Wang et al. [Wang et al., 2010] since these were kept at different collections for more than a decade. A similar situation is also observed for one strain of *S. polyrhiza* (7205), the barcode sequences of which clearly showed identity to those of *S. intermedia*. In the study by Wang et al. [Wang et al., 2010], in which barcodes for these four strains of "suspect" species assignment were used for comparative analyses, the apparent anomalies for *L. japonica* 7182 and *S. polyrhiza* 7205 were noted but not resolved. The possibility that *L. japonica* is a hybrid between *L. minor* and *L. turionifera* was raised as an explanation, but this hypothesis was not consistent with an AFLP-based phylogenetic analysis of the Lemna species [Bog et al., 2010]. For *S. polyrhiza* 7205, after more extensive genotyping of 34 additional *S. polyrhiza* strains and examination of five additional genotyping markers, the possibility that this strain is misidentified was raised [Wang et al., 2010]. Upon follow-up conversations and additional physiological studies by one of the authors (KA) and Elias Landolt in 2011, strain 9203 as well as strain 9428, which has yet to be sequenced, have been redesignated as *S. intermedia* since neither strain is able to produce turions

under 3 different methods of turion formation (KA, data not shown). Lastly, in the same conversation, Elias Landolt had confirmed that the *W. microscopica* 9276 strain formerly available in the RDSC collection at Rutgers (originally in the duckweed collection held at Biolex, Pittsboro, NC, USA) is in fact likely to be *W. elongata* or *W. columbiana* (personal communication to KA, 2011) [Sree and Appenroth, 2014]. These revised designations are remarkably consistent with our present BLAST results using the two barcodes. However, to directly assess the validity of reassignment for strains 7182, 9123 and 9276, we needed to generate the corresponding barcodes from additional strains of these three species. If the hypothesis of misidentity were correct, we would expect to find barcode sequences distinct from those generated by Wang et al. [Wang et al., 2010]. The addition of these sequences would also strengthen the completeness of our reference dataset as well as its assessment for species ID.

In analyzing the information summarized in Table 2.1, we also sought to compare the relative ability of the two intergenic barcodes to serve as useful markers for species ID with duckweed. In the previous work by Wang et al. [Wang et al., 2010], this analysis was done by finding the "best match" of each barcode sequence within a dataset of 84 accessions (which excluded species with only a single strain) and calculating the success rate of matching strains of the same species for each intergenic region. The authors concluded that the *atpF-atpH* barcode is clearly superior to the other markers for duckweed, including the *psbK-psbI* barcode. This notable finding contradicted previous work that indicated the *psbK-psbI* barcode is more useful than the *atpF-atpH* marker for species ID of plants in general [Kress et al., 2005, Lahaye et al., 2008a]. One explanation for this apparent discrepancy is the uneven and incomplete representation of species in the "best match" analysis used in the work of Wang et al. [Wang et al., 2010]. Since 18 species were not represented in the comparative analysis, it is not clear whether the conclusion is applicable to all or even most species of duckweed. To minimize contributions from these factors, we used the data from our BLASTN-based approach and compared the ability of each of the two barcodes to positively identify each of the duckweed species. Excluding the species where the two barcodes worked similarly, we found that the *atpF-atpH* marker worked better in three species whereas the *psbK-psbI* marker is superior for ten species,

with six instances where the *psbK-psbI* barcode clearly gave much better resolving power than the *atpF-atpH* marker. We thus believe that these two intergenic markers are complementary and should be used together for the positive ID of duckweed species whenever possible.

2.3.3 Further barcoding efforts for three species of duckweed to complete an improved database: *L. japonica*, *W. borealis* and *W. microscopica*

In an effort to complete the reference dataset for the two barcodes used in this study, we carried out additional sequence analyses from five strains of *L. japonica* and two strains of *W. borealis* that are available in the RDSC collection. In addition, the recent isolation of seven new accessions of *W. microscopica* [Sree and Appenroth, 2014] provided an opportunity to complete the reference library of these two barcodes for duckweed. Barcode sequences from these 14 duckweed strains were generated and added to our local database. Our BLAST analyses using the enhanced dataset are summarized in [Borisjuk et al., 2015, Table S4, S5] for these three species. Our results showed that all seven strains of *W. microscopica* are identical for the two barcodes that we examined and that these sequences are readily distinguished from the other species of duckweed. Importantly, the previously barcoded *W. microscopica* 9276 strain [Wang et al., 2010] is less closely related to the newly isolated *W. microscopica* clones than either *W. brasiliensis* (*psbK-psbI*) or *W. arrhiza* (*atpF-atpH*). These results support the re-designation of strain 9276 in the RDSC as *W. elongata* while indicating that the new strains of *W. microscopica* may be highly related. In contrast to the situation with *W. microscopica*, barcode sequences from additional strains of *L. japonica* and *W. borealis* showed that they do not provide better separation from other species. Thus, the *psbK-psbI* and *atpF-atpH* barcodes of *L. japonica* and *W. borealis* strains are too similar to those from other duckweed species and thus are insufficient for positive ID [Borisjuk et al., 2015, Table S4, S5]. In addition, since barcode sequences from four of the additional *L. japonica* strains were indistinguishable from that of *L. minor*, *L. minor* would also need to be removed from the list of species that can be positively identified (Table 2.1). In conclusion, using the present reference

database for the two plastidic intergenic sequences, we can potentially identify 30 out of the 37 known species of duckweed. Additional genotyping strategies, such as AFLP, or exploration of other barcode candidates from the nuclear or mitochondrial genomes can be alternative methods to resolve the identity of the more problematic species.

2.3.4 Application test for species identification: case studies at the RDSC

One of the main drivers for our work is to provide a reliable and facile protocol for positive species ID of newly isolated duckweed strains, since species ID is one of the criteria for assignment of a unique 4-digit registration number that is recognized by the duckweed community. To test the approach outlined in the present work, we carried out barcode analyses for 15 strains from the RDSC collection that had not been positively typed [Borisjuk et al., 2015, Table S6]. In addition, we also included a strain of duckweed called *L. aoukikosa* that has been used in a number of bioremediation studies by the Morikawa lab in Japan [Yamaga et al., 2010]. Using only the *atpF-atpH* barcode, we were able to ID the species for nine of the RDSC strains with the remaining strains narrowed to a limited set of potential species (Table 2.3). Future barcode analysis with the *psbK-psbI* intergenic region may help to further resolve the species ID of the Wolffia strain in this study. For the *L. aoukikosa* strain, we found that it is identical to the barcode sequences of *L. aequinoctialis*. At this point, we see no reason to designate it as a new species of *Lemna*. The nine newly typed RDSC duckweed strains have thus been assigned new 4-digit codes for their maintenance and tracking in the RDSC as well as their future use in publications, and we recommend renaming the *L. aoukikosa* strain as *L. aequinoctialis*.

2.4 Discussion

In an effort to complete the cpDNA barcoding of the whole Lemnaceae family, we have undertaken analysis of the *atpF-atpH* and *psbK-psbI* spacer regions for 16 accessions from the RDSC collection representing six duckweed species that were not covered in previous studies: *L. perpusilla*, *L. tenera*, *L. yungensis*, *W. caudata*, *W. repanda*, and

W. welwitschii. Sequence alignment with the *atpF-atpH* intergenic sequences using the MEGA online program package resulted in a close grouping of ecotypes of the same species and species placement consistent with their morphological classification. The only exception was one of the accessions of *W. welwitschii* grouped with *W. repanda*. The same *W. welwitschii* ecotype was also grouped with *W. repanda* based on alignment of the *psbK-psbI* sequences. Therefore, these results suggest that the identity of the accession should be re-examined in light of these molecular data. In contrast to the results derived from the *atpF-atpH* spacer, the alignment of *psbK-psbI* sequences grouped two additional accessions (*L. yungensis* 9210 and *W. caudata* 9139) with those of other species.

To assess the potential of using a publicly accessible software, such as BLAST, and open-source databases, such as NCBI GenBank, as a facile platform for barcoding of duckweed species, we performed a systematic analysis of these two plastidic barcodes for their potential to identify each of the 37 known species of duckweed. With the generation of additional barcode sequences for 14 strains that comprised three species with questionable assignments, we were able to determine whether these species have sufficient interspecies variation for their positive identification. In summary, our work has now completed the first comprehensive survey of two plastidic barcodes and produced a reference sequence dataset that will allow the identification of 30 species of duckweed using a PCR-Sequence-BLAST protocol. For the seven species that cannot be identified unambiguously with these two barcodes, this resource can nevertheless provide a smaller subset of two to four different candidate species for further resolution using other approaches, such as more classical physiological/morphological typing. Alternatively, additional molecular techniques such as AFLP and barcoding strategies using polymorphic nuclear loci could be future resources to complement the current protocol.

2.5 Acknowledgements

This paper is dedicated to Elias Landolt (1926–2013), whose work on curating and sharing the first comprehensive duckweed collection has laid the cornerstone for this model system to understand natural diversity and adaptation. Duckweed research in the Lam laboratory is supported in part by a Hatch grant from NIFA-USDA through the New

Jersey Agricultural Experiment Station (project #12116), support from the School of Environmental and Biology Sciences at Rutgers University, as well as funding through an NSF-IGERT project on Renewable and Sustainable Fuel Solutions (DGE-0903675).

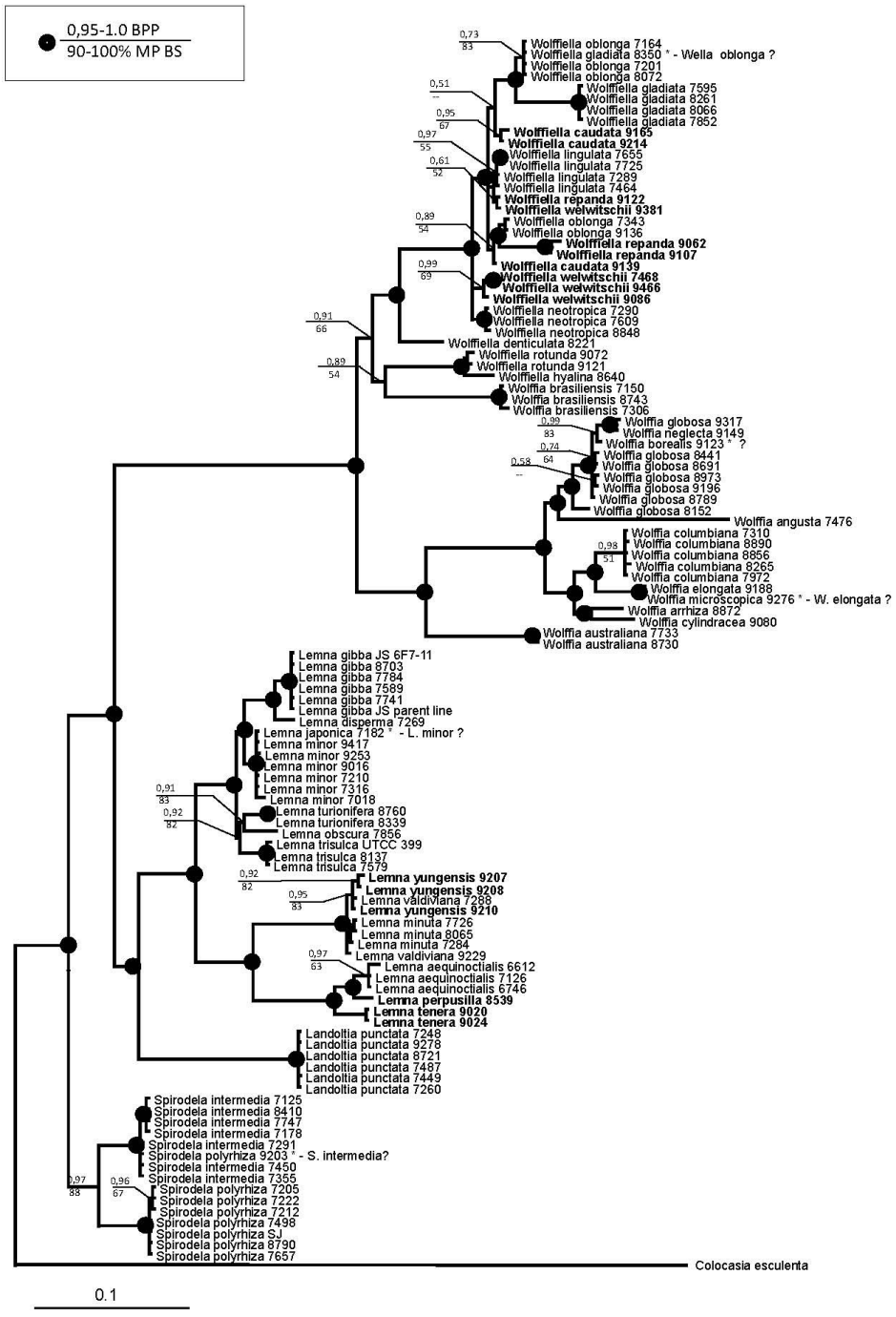

Figure 2.1: Bayesian consensus tree from the analysis of the combined cpDNA dataset (*atpF-atpH* and *psbK-psbI* intergenic spacers) of all Lemnaceae taxa and Colocasia esculenta as outgroup. Bayesian posterior probabilities (B PP) and Maximum parsimony bootstrap values (MP BS) are shown on the branches. Strongly supported clades (MP BS > 90% and BPP > 0,95) are indicates with black points.

Species	*psbK-psbI*	*atpF-atpH*
Spirodela intermedia	✓	✓
Spirodela polyrhiza	✓	✓
Landoltia punctata	✓	✓
Lemna aequinoctialis	X	✓
Lemna disperma	✓	✓
Lemna gibba	✓	✓
Lemna japonica*	$U \rightarrow X$	$U \rightarrow X$
Lemna minor	✓$\rightarrow X$	✓$\rightarrow X$
Lemna minuta	X	X
Lemna obscura	✓	✓
Lemna perpusilla	$\pm$	$\pm$
Lemna tenera	✓	✓
Lemna trisulca	✓	$\pm$
Lemna turionifera	✓	✓
Lemna valdiviana	X	X
Lemna yungensis	$\pm$	X
Wolffiella caudata	$\pm$	X
Wolffiella denticulata	✓	✓
Wolffiella gladiata	X	✓
Wolffiella hyalina	✓	X
Wolffiella lingulata	X	X
Wolffiella neotropica	✓	X

Table 2.1: Summary of BLASTN results from sequence comparison of two barcodes for 37 species of duckweed. To determine the fidelity of this approach for resolving the identity of each species, the degree of interspecies sequence similarities were tested by selected representative(s) from each of the 37 species that are used in constructing a database comprised of 200 strains. The number of accessions for each species that have barcode sequences currently deposited in NCBI are listed on [Borisjuk et al., 2015, Table S1], along with the reference for their origin. From the BLASTN search results, the level of intraspecies variation (in species with data from multiple accessions) as well as the interspecies similarity levels are examined in the BLASTN report as shown in [Borisjuk et al., 2015, Tables S2, S3], for the *psbK-psbI* and *atpF-atpH* barcodes respectively. The Bit Score, which is a normalized value for the degree of sequence similarity used in BLAST analyses that takes into account of sequence length as well as nucleotide position, was found to be a convenient indicator for determining a cut-off. By examining the intraspecies and interspecies variations, we found that a difference of 40 for the Bit Score between two species is a sufficient gap that usually provides a window of separation, while a percent identity difference of less than 2% would be considered a weak gap for confident separation of two species. ✓: barcode sufficient for species identification (ID); X: barcode does not have sufficient sequence specificity for species ID; $\pm$: barcode may be potentially useful for species ID but will require caution; U: undetermined, likely issue in current species ID and need additional data from additional strains for resolution. Strain designations for four species were subsequently revised ($\rightarrow$) based on additional sequence information.

Species	*psbK-psbI*	*atpF-atpH*
Wolffiella oblonga	X	✓
Wolffiella repanda	±	*X*
Wolffiella rotunda	✓	*X*
Wolffiella welwitschii	±	*X*
Wolffia angusta	✓	✓
Wolffia arrhiza	✓	✓
Wolffia australiana	✓	✓
Wolffia borealis*	$U \rightarrow X$	$U \rightarrow X$
Wolffia brasiliensis	✓	✓
Wolffia columbiana	✓	±
Wolffia cylindracea	✓	✓
Wolffia elongata	✓	✓
Wolffia globosa	*X*	*X*
Wolffia microscopica*	$U \rightarrow$✓	$U \rightarrow$✓
Wolffia neglecta	±	*X*

Table 2.2: Continued: Summary of BLASTN results from sequence comparison of two barcodes for 37 species of duckweed.

RDSC Serial No.	Previous Strain #	Location of Origin	Species ID	Conf.[1]	Registration Number[2]
7	DWC021	San Diego, CA, USA	Landoltia punctata	***	2010
241	DWC136	NJ/DE, USA	Lemna minor/japonica[3]	*	
428	DWC313	Not known	Wolffia globosa/neglecta/ borealis/angusta[3]	*	
612	EL004	Wellesley Island, NY, USA	Lemna minor/japonica[3]	*	
616	EL008	Princeton Meadows Wastewater Treatment Plant, NJ, USA	Lemna minor/japonica[3]	*	
617	n/a	Passion Puddle, Rutgers University, New Brunswick, NJ, USA	Lemna minor/japonica[3]	*	
618	EL010	Pinelands Nursery, Columbus, NJ USA	Lemna obscura	**	2011
619	EL011	Princeton Meadows Wastewater Treatment Plant, NJ, USA	Landoltia punctata	***	2012
620	EL012	Linneas Garden, Uppsala, Sweden	Lemna minor/japonica[3]	*	
639	EL017	Kunming, Yunnan, China	Lemna aequinoctialis	**	2013
643	EL018	Kunming, Yunnan, China	Spirodela polyrhiza	***	2014
695	EL028	Recife, Brazil	Lemna minor/japonica[3]	*	
696	EL029	Palace of Versailles, Paris, France	Lemna aequinoctialis	**	2015
700	EL031	Mangueira, Recife, Brazil	Lemna aequinoctialis	**	2016
707	EL034	Port Douglas, Australia	Landoltia punctata	***	2017
n/a	*L. aoukikosa*		Lemna aequinoctialis	**	2018

Table 2.3: Species identification of 15 new duckweed strains based on the *atpF-atpH* barcode using the reference dataset for BLAST comparison. [1] Confidence of species identification based on BLAST search results: ***, very strong; **, strong; *, weak.
[2] As recommended by the International Steering Committee on Duckweed Research and Application, all clones that have undergone species identification should be given a unique 4-digit number assigned by the Rutgers Duckweed Stock Cooperative.
[3] Species identification could not be determined but was narrowed down to 2–4 possible species.

Chapter 3

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

Authors: Philomena Chu, Glen M. Wilson, Jennifer Vaiciunas, Joshua Honig, and Eric Lam

Submitted.

3.1 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 (i.e., duckweed), can have varying phenotypes depending on environmental conditions [Vaughan and Baker, 1994], complicating taxonomic interpretations based on solely morphological characteristics.

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 *rbcL* and *matK* chloroplast genes and *trnK* and *rpl16* introns with non-molecular data to construct a monophyletic tree for the five genera of Lemnaceae [Les et al., 2002]. In *Wolffia*, six of the 11 known species could be distinguished by using the plastidic *rpl16* and *rps16* markers [Bog et al., 2013]. More recently, the plastidic *rpl16, rps16,* and *atpF-atpH* markers were used to distinguish species in the *Landoltia* and *Spirodela* genera [Bog et al., 2015]. Wang et al. [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. [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 strains for most species were included in this work in order to define the intraspecific level of sequence variation for each species barcode 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 variation. The Amplified Fragment Length Polymorphism (AFLP) technique has been shown to successfully distinguish *Lemna* and *Wolffia* species and resolve most intraspecific variations [Bog et al., 2010, Bog et al., 2013]. AFLP was also used to separate clones into *Spirodela polyrhiza*, *S. intermedia* and *Landoltia punctata* species [Bog et al., 2015]. However, the AFLP technique itself is more challenging to standardize and AFLP banding patterns did not provide sufficient variability to unequivocally characterize *L. punctata*, *S. polyrhiza* or *S. intermedia* clones. In fact, although AFLP demonstrated greater genetic diversity than plastidic regions in *S. polyrhiza*, only one or four of 24 clones could be separated, depending on the type of analysis that was conducted [Bog et al., 2015]. This highlights the difficulty of finding a suitable molecular approach to duckweed identification at the subspecies level.

A recent paper used three pairs of single sequence repeat (SSR) markers to resolve 49 *S. polyrhiza* clones into 11 haplotypes, with about half of the clones represented from China [Feng et al., 2017]. While the markers chosen for this study were able to distinguish *L. punctata* from *S. polyrhiza* species, each haplotype may be represented in multiple clones, so it is not clear how those clones that share the same haplotype can be further resolved. Additional investigation with clones from more geographically diverse locations could provide a better idea of how applicable these SSR markers could be.

As an alternative to AFLP or SSR approaches that target random template sites within the genome for amplification, we want to first identify candidate sites amenable to

genotyping by using a genome sequence-guided pipeline. We reason that the nucleotide-binding leucine-rich repeat protein-encoding genes (NB-LRRs) could be a good source for hyper-polymorphic markers in plants to facilitate genotyping efforts. 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 that its NB-LRRs can be rapidly adapted 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]. Specifically, this study found that the NB-LRRs of *A. thaliana* were located in regions of high nucleotide diversity distributed throughout the five chromosomes in pericentrometric regions [Clark et al., 2007]. In twenty *A. thaliana* populations, NB-LRRs were found to display the greatest percentage of amino acid pairwise distances in the major protein families [Gan et al., 2011], representing an extreme divergence of NB-LRR encoding genes at the protein level. Further analysis of 1,135 *A. thaliana* populations confirmed that the NB-LRRs are the most highly diverged class of protein families across accessions [Consortium, 2016]. We thus expect that the highly conserved nature of the NB-LRRs across various plant lineages and their polymorphic characteristics could provide an excellent target as a genotyping marker that can be used for species as well as clone identification in plants.

Duckweeds are finding greater use as a model organism [Appenroth et al., 2015, Lam et al., 2014]. In particular, *S. polyrhiza* is an attractive bioethanol feedstock for its high biomass yield on wastewater in pilot-scale studies [Xu et al., 2011a, Xu et al., 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]. Two publicly available *S. polyrhiza* reference genomes of clones 7498 and 9509 should further aid its development as an aquatic crop [Wang et al., 2014, Michael et al., 2017].

Clones of a particular duckweed species that have been adapted to diverse locales 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 demonstrate the feasibility of using NB-LRR loci as genotyping markers, we first chose to focus on a set of nine *S. polyrhiza* clones that represent a large specific turion yield (STY) distribution and diverse geographical range [Kuehdorf et al., 2014]. STY was used as a proxy for climatic and environmental adaptations to each population's local conditions. The non-reference clones from this subset were previously sequenced with the Illumina NGS platform and mapped against the reference genome produced with the 9509 clone from Jena, Germany [Michael et al., 2017].

In our effort to use NB-LRR genes as markers for distinguishing *S. polyrhiza* clones, we first approximate the NB-LRR genes in the 9509 genome with those genes that contain an NB-ARC subdomain—we refer to such genes as NB-ARC-related genes (ARC is an abbreviation of APAF-1, *R* gene products and CED-4). We leveraged the available genomic data from nine *S. polyrhiza* clones to analyze 53 NB-ARC-related regions across these nine genomes to identify those regions that have the highest discriminatory power. Four NB-ARC-related loci were identified for their hyper-polymorphic fragment length variations and an additional three NB-ARC-related loci were identified for their high concentration of SNPs. We then tested these markers on 23 *S. polyrhiza* clones: nine from our training set, 11 additional ones from the prior STY study [Kuehdorf et al., 2014] including clone 7498, and three additional clones selected for increased geographical diversity. Finally, we also tested and validated 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*.

3.2 Materials and methods

3.2.1 Plant materials 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 hr light/8 hr dark conditions at 25 °C. Several of the

S. polyrhiza clones for this study were previously examined for their STY [Kuehdorf et al., 2014], a likely climatic adaptation trait that is potentially important for biomass production. 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 numbers for all *S. polyrhiza* clones used in this study are listed in Tables 3.3 and 3.5. Ten *S. intermedia* clones from the RDSC collection were also included in this work. Their Landolt numbers are provided in Fig. 3.5. 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.

3.2.2 Finding 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 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. This set of proteins is a first approximation of the NB-LRR proteins and should suffice for our purpose of intraspecific comparison; we refer to these as NB-ARC-related genes.

Eight clones of *S. polyrhiza* with varying STY were previously resequenced to $\sim$ 40X coverage using the Illumina NGS platform and compared with the reference genome of *S. polyrhiza* 9509 (the first nine clones listed in table 3.3). Variant call format (VCF) files produced from the resequencing of these eight clones and the 9509 VCF were used for our study. The 7498 genome was not included in this part of our analysis because no Illumina sequencing-based VCF file was available for this clone. Rather, clone 7498 was sequenced using the 454 NGS platform and BAC-end sequences from Sanger sequencing, which is difficult to directly use with the other eight sequenced genomes. 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 refer to a region of length 200–900bp 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 8,657 windows were identified. Each window was then analyzed using Primer3 version 4.0.0 to locate primers [Untergasser et al., 2012]; Primer3 returned 6,576 windows with possible primer sets. For each window, Primer3 was run with default parameters and requested primers to be placed in the initial and terminal conserved regions that were identified for the window.

The windows with possible primer sets were then ranked with two different methods to select those windows that could best distinguish each clone based on differences in PCR product length. Both methods gave similar rankings of windows (Table 3.1).

The first method of ranking windows, which we refer to as the "bottleneck method," takes 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, calculates the difference of the length of each homologous chromosome from the reference window. Hence for a given window, each clone E is associated to a pair of integers (a_E, b_E) where a_E is the difference in length of the window of one chromosome from the reference and b_E is 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) is given by the following expression.

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

Windows were then ranked by the number of pairs of clones with bottleneck distance of at least 5 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, uses unphased variant files and considers the range of possible PCR product lengths for each clone as determined by the variant files. For a given window and clone E, we calculate the maximum and minimum difference in length possible of homologous chromosomes from the reference sequence as follows. Two running counts are kept as we go through the variant calls in the window under consideration, the maximum count M_E and the minimum count m_E. The length of a homozygous insertion or deletion is

added or subtracted respectively from both counts, whereas the length of a heterozygous insertion is added to M_E and the length of a heterozygous deletion is 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 will lie in the interval $[m_E, M_E]$.

For a fixed window, we now have for each clone E an associated interval $[m_E, M_E]$. We measure the likelihood that two clones E and F can be distinguished by their PCR product by calculating the probability that a number from $[m_E, M_E]$ and a number from $[m_F, M_F]$ chosen uniformly at random will be separated by a distance of at least 5; call this probability the max-min distance of E and F with respect to the given window. We then ranked windows 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 less than 800 whereas true INDEL calls were associated with quality scores of at least 800 (data not shown). We then removed all INDELs from the variant files with quality score less than 800 and ran the bottleneck method (with no phasing, phasing by Beagle, and random phasing) and the max-min method on these new variant files. A summarized workflow is illustrated in Figure 3.1.

3.2.3 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 that the window predicted would differ by at least one base pair. The groups of windows that 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.

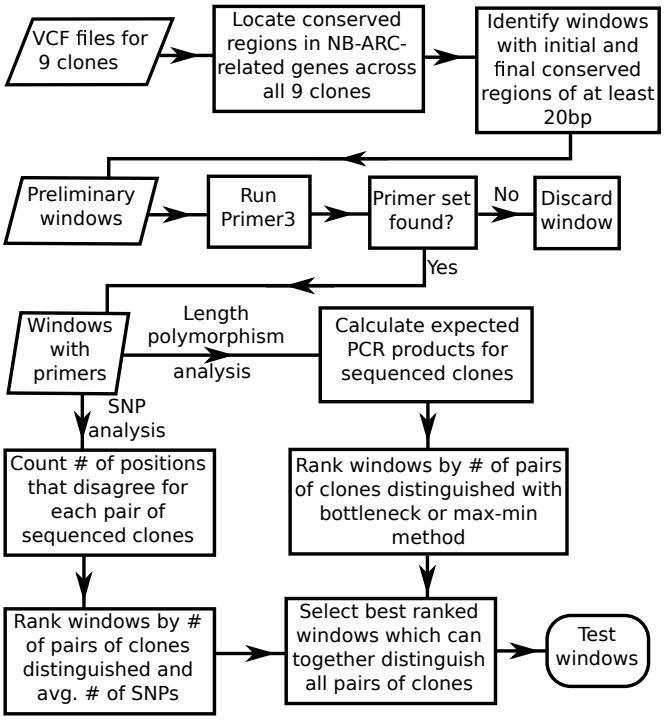

Figure 3.1: Workflow for identification of candidate hyper-polymorphic regions of NB-ARC-related genes. A window is a region of an NB-ARC-related gene with conserved start and stop sub-domains. VCF files for the 9509 reference genome and the 8 non-reference genomes were used for downstream analysis.

3.2.4 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]. Briefly, the first 30 cycles of PCR were conducted using a forward primer with a 5' M13(-21) tail and marker-specific reverse primer to incorporate the M13 sequence into the amplicons. Those amplicons are used as template for the next eight cycles of PCR with a 6-carboxyfluorescein (FAM)-labeled-M13(-21) forward primer and marker-specific reverse primer.

The PCR components for the first 30 cycles were done in 20 µl total reaction volume with 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 and reverse primers (10 µmol), and 100 ng DNA template. 8% DMSO

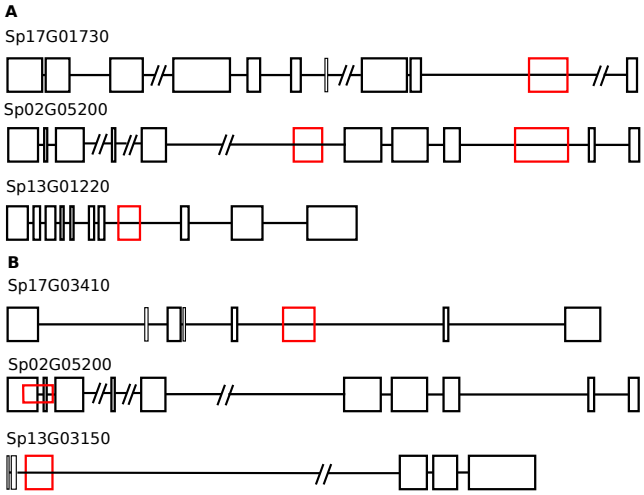

Figure 3.2: The locations of selected windows for polymorphism analysis. Red boxes illustrate windows identified from workflow of Figure 3.1. Black boxes represent exons (displayed 5' to 3' for each gene). Scale is uniform throughout all genes, however, some introns were truncated as indicated by hash 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 for clarity.

was added to the reactions with primer sets Sp17, Sp02a, and Sp02b. PCR reactions with Phusion were run under the following conditions: 98 °C for 30 seconds, 30 (98 °C for 10 seconds, X for 30 seconds, 72 °C for 24 seconds), 72 °C for 10 minutes, 4 °C ∞. X is 66 °C for Sp17, 64 °C for Sp02a, 62 °C for Sp02b, 67 °C for Sp13 or 67 °C for Sp12.

M13-labeled amplicons were then fluorescently labeled in similar reactions as in the first round 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 seconds, 53 °C for 45 seconds, 72 °C for 24 seconds, then final extension at 72 °C for 10 minutes and 4 °C ∞.

Aliquots of the reaction samples were assayed by agarose gel electrophoresis to check the quality of the amplification reactions and then were submitted for capillary electrophoresis on an Applied Biosystems 3500xl Genetic Analyzer (Thermo Fisher Scientific, USA) at the Rutgers Plant Biology DNA Barcoding Core Facility. 2 µl of PCR product was diluted in 40 µl of sterile dH_2O. GeneScan 1200 LIZ (Thermo Fisher Scientific, cat. #4379950) was used as a size standard. Fragments were identified and binned using the

two-surrounding-peaks sizing method in the microsatellite plug-in from Geneious software (version 10). Fragments with peaks above 200 RFU were kept for downstream analysis. M13-labeled fragments amplified from 9509 were Sanger sequenced to confirm predicted sequence product. PCR amplifications that yielded more than one band were gel extracted and 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`.

3.2.5 SNP Sequencing

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].

3.2.6 Amplification from *S. intermedia* species

PCR reactions for 10 *S. intermedia* clones (listed in Fig. 3.5) 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.

3.2.7 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. 3.6 and 3.7, clones in clusters formed at 25% dissimilarity or less are similarly colored.

3.2.8 Code and data availability

The computer code for the length and SNP polymorphism analyses, along with the rankings for each analysis, are available at `https://github.com/glenwilson/variant_analysis`.

3.3 Results

3.3.1 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, we identified 53 NB-ARC-related genes through our Pfam-based HMM search that were used for downstream analysis. Those genic regions were analyzed across our nine Illumina sequenced genomes to identify conserved regions for primer design with PCR products ranging from 200–900 bp in length. 6,576 potential PCR regions were ranked according to the bottleneck and max-min methods to maximize length polymorphism across clones. After filtering out low confidence INDELs, potential primer sets were re-analyzed using the bottleneck and max-min methods.

Four windows were chosen based on their rankings, the quality of their primer sets, and their ability to collectively distinguish all nine sequenced clones under consideration (Figure 3.2A). Three of these regions were based on our criteria for markers that allowed 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 1104684–1105262 (Sp13) between exons 7-8. Sp17G01730 was annotated as "Similar to RPS2: Putative disease resistance protein RPS2 (*Arabidopsis thaliana*)" in the 9509 reference genome, Sp02G05200's annotation was "Similar to RGA4: Putative disease resistance protein RGA4 (*Solanum bulbocastanum*)" and Sp12G01220 had an annotation of "Similar to CDC48C: Cell division control protein 47 homolog C (*Arabidopsis thaliana*)" [Michael et al., 2017]. These windows were chosen primarily based on their ranking with the max-min method of ranking windows (Table 3.1). Although the rankings of the window in Table 3.1 may not seem optimal, many overlapping windows were ranked similarly and were effectively identical in which clones they were predicted to distinguish, the most significant difference being the quality of their primer sets. For example, window Sp17 appears in a group of 13 overlapping windows ranked 1-13, hence, it is effectively the best window identified by our ranking. Sp13 arises in the third best cluster of overlapping windows, and Sp02b arises in the sixth best cluster of overlapping windows. Furthermore, of the 6,576 windows analyzed with the q-max-min method, only the first 841 windows were able to distinguish any pairs of windows with our threshold of a 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. When our analysis pipeline analyzed windows of length 200–1200 bp, Sp02a is ranked 69th out of 8,037 windows using the max-min method without filtering out low quality INDELs.

	Sp17	Sp13	Sp02b
np	2	149	141
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

Table 3.1: The rankings of windows selected for length polymorphism analysis. 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 with quality less than 800 removed.

In a preliminary analysis of the windows, which did not filter out variants with quality scores less than 800, many of the windows identified and tested with PCR were found not to harbor the INDELs that were predicted from our analysis and thus did not demonstrate the expected length polymorphisms when aliquots of the PCR products were assayed on agarose gel electrophoresis. Instead, many of these fragments appeared to be single PCR products. These amplicons were submitted for Sanger sequencing, aligned, and found to have multiple diagnostic SNPs. From this analysis, three regions were selected for experimentation based on the presence of several SNPs observed from the Sanger sequencing (Fig. 3.2B): Sp17G03410 chromosome positions 2322464–2323187 (Sp17-SNP), Sp13G03150 positions 3078156–3078820 (Sp13-SNP), and Sp02G05200 positions 3718402–3718995 (Sp02-SNP). These three genes were previously 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 6,576 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 2,563, but still expected to distinguish 27 out of 45 pairs from the training data. We note that 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. Sp02-SNP is at a different location on Sp02G05200, while Sp17-SNP and Sp13-SNP are NB-ARC-related genes on the same chromosomal regions as Sp17 and Sp13, respectively. This indicates that these three loci in the *S. polyrhiza* genome could be hotspots for polymorphisms. The design of all chosen primer sets is diagrammed in Table 3.2.

3.3.2 Fingerprinting of *S. polyrhiza* clones

The four NB-ARC-related windows Sp17, Sp02a, Sp02b and Sp13 were tested using PCR and capillary electrophoresis. We observed two possible PCR fragments for the Sp17 window resulting in four fragment combinations, eight possible fragments in Sp02a for 10 combinations, four possible fragments for window Sp02b with four fragment combinations,

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

Table 3.2: PCR Primer Sequences. The first group of primer sets were used for length polymorphism analysis, while the second group (-SNP) were used for SNP polymorphism analysis. Forward primers have M13(-21) sequence at the 5' end.

and four fragments for Sp13 with four combinations (Table 3.4). Bins 7–10 in Sp02a were added as new bins with the inclusion of the unsequenced clones to the data set. 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. 3.3, 3.4, and 3.5). Amplicons larger than 1200 bp were not detected by fragment analysis using our automated sequencer.

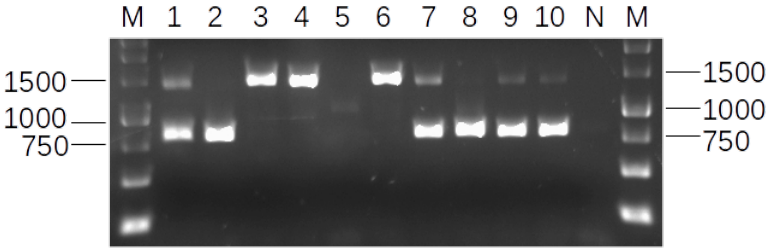

Figure 3.3: Fragment comparison between 10 clones of *S. polyrhiza* using primer set Sp17. PCR amplification from an intron within Sp17G01730 with a 66 °C annealing temperature. The lane headings correspond to assigned numbers for each sequenced clone in Table 3.3. M: 1 kb ladder as size marker (GoldBio), numbers to each side correspond to the size of DNA in base pairs. N: negative control without template (water added).

PCR products of the 9509 clone from each primer set were sequenced to confirm the

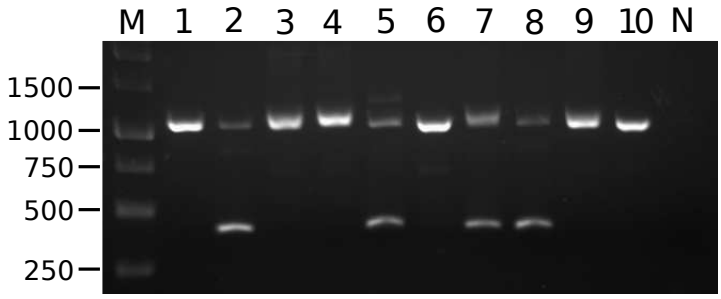

Figure 3.4: Fragment length polymorphism between 10 clones of *S. polyrhiza* using Sp02a. Fragments amplified from Sp02G05200 window using annealing temperature at 67 °C. Lane headings correspond to same clones used in Fig. 3.3. M = 1 kb ladder (GoldBio), N = negative control (water).

predicted amplicon. All 9509 amplicons, except those from Sp17, amplified the expected sequence. 9509 gDNA PCR product from Sp17 contained an additional fragment that matched to Sp17G03530 by BLAST (Fig. 3.3), 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`). In the case where multiple bands were amplified in 9509, those sequences were cloned and sequenced. We assume that if the PCR products from the other clones share the same fragment size as 9509 that they likely correspond to the same loci as that amplified from the 9509 template since these genomes are highly conserved [Michael et al., 2017].

Out of the nine *S. polyrhiza* clones that served as our training dataset, each clone has a unique fingerprint, consistent with our pipeline's accuracy in performing its intended function. With the addition of 7498 and 13 additional unsequenced clones, the three Indian clones 9506, 7379, 9503 could not be distinguished from one another based on fragment length polymorphism markers, and the clones 9622 and 9514 could not be distinguished from one another either. However, 18 out of the 23 clones examined had unique fingerprint patterns using our four primer combinations (Table 3.3).

3.3.3 Genotyping with SNP Primers

Primer sets for Sp17-SNP, Sp02-SNP, and Sp13-SNP were designed and tested on the same 23 *S. polyrhiza* clones. The length of the 9509 amplicon (minus the forward and

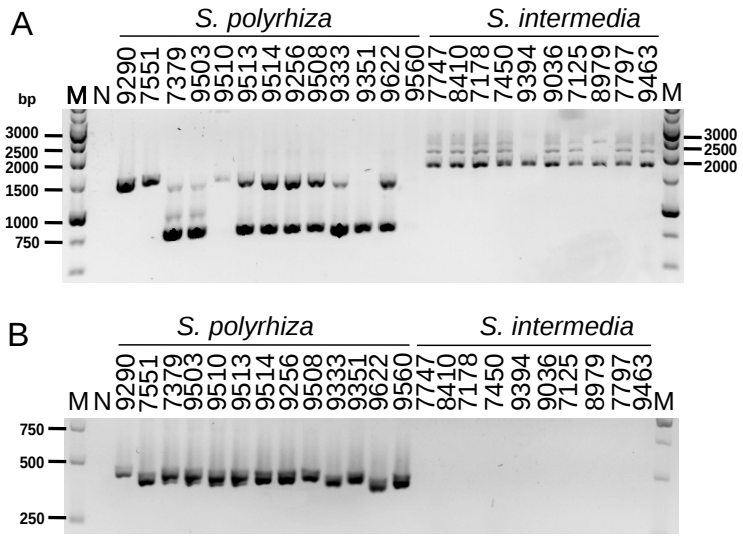

Figure 3.5: Use of NB-ARC primer set for interspecific barcoding application. (A) The fragment length polymorphism primer set Sp17 (from Sp17G1730) is used for PCR amplification with gDNA of non-reference *S. polyrhiza* and *S. intermedia* with annealing at 66 °C. A 0.8% agarose gel was run at 50V for 2 hours. (B) SNP primer set Sp13 (Sp13G01220) with gDNA of non-reference *S. polyrhiza* and *S. intermedia* with annealing at 67 °C. A 2% agarose gel was run at 50V for 2 hours. M = 1 kb ladder (GoldBio), N = negative control (water), lane headings indicate the clone numbers in the Landolt collection.

reverse primer sequences) from Sp17-SNP was 588 bp, 546 bp for Sp02-SNP, and 493 bp for Sp13-SNP. Some of the chromatograms from the Sp17-SNP genomic DNA amplicons contained double peaks, corresponding to a 9 bp heterozygous INDEL in the sequence. It was necessary to either sequence in both directions or use the Poly Peak Parser program to determine the full sequence. 9242 and 9290 contained an additional one bp T INDEL in one of the alleles that occurred upstream of the 9 bp INDEL. Also, many of the Sp02-SNP genomic DNA chromatograms from the unsequenced *S. polyrhiza* had complicated traces, possibly due to increase in copy number of this NB-ARC-related locus, making it necessary to clone the PCR products and sequence individual clones instead of directly from the amplicon DNA to get clean reads. We defer to the Sanger-sequenced colony results when a discrepancy arises amongst the Illumina, genomic DNA or Sanger sequencing results. Multiple sequence alignments of the sequenced PCR products revealed SNPs within each marker with heterozygous positions which are illustrated in Table 3.5. We found that the three Indian clones 9503, 9506 and 7379 share the same SNP combinations using these

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

Table 3.3: NB-ARC-related primer capillary results. See table 3.4 for the allele combination abbreviations. CZ is an abbreviation of Czech Republic and MZ is an abbreviation of Mozambique. All clones listed in lanes 1–10 above the line have been sequenced.

three primer sets and therefore remain indistinguishable.

By combining both fragment length and SNP-based markers, only the three Indian clones 9506, 7379, and 9503 are inseparable from one another out of the 23 *S. polyrhiza* clones examined. We found the average number of SNPs from all nine clones in the three NB-ARC-related loci to be 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].

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

Since the NB-ARC-related markers can apparently distinguish a majority of the intraspecific genotypes in *S. polyrhiza* tested, we would expect them to also provide interspecific genotyping capability since a greater degree of sequence divergence would be expected.

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
	2	582
	3	572, 582
	4	554, 564, 572, 582
Sp13	1	419
	2	446
	3	419, 446
	4	430, 446

Table 3.4: Key for table 3.3. The Fragment combination column lists the lengths of the PCR product for the given window as determined by capillary electrophoresis.

In the *Spirodela* genus, the other 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 conducted for the *S. polyrhiza* templates (Fig. 3.5). The amplification pattern from *S. polyrhiza* clones is clearly distinct from *S. intermedia* PCR products when using this primer set, with the amplified fragment from all 10 tested *S. intermedia* clones migrating at a larger apparent size than the fragments observed with *S. polyrhiza* clones. This clear and consistent difference in fragment pattern between the two *Spirodela* species thus provides a simple genotyping tool between them that can obviate any need for DNA sequencing such as those required for plastidic barcodes. Addition of one of the other fragment length primer set that only amplifies from *S. polyrhiza* should provide additional support for the species identification (Fig. 3.5B). In sum, the combined use of the Sp17 and Sp13 primer sets defined from

		Sp02-SNP										Sp13-SNP			Sp17-SNP									
Clone	Origin	99	184	216	250	268	274	288	299	380	498	111	116	191	46	54	203	343	446	468	509	527	545	INDEL
9509	Germany	A	A	G	A	A	A	C	C	T	G	T	G	A	A	C	Y	W	Y	R	R	Y	R	0,9
9316	India	R	R	R	M	R	R	Y	Y	C	A	W	R	W	G	T	C	T	T	A	A	C	A	0
9501	Albania	R	A	R	A	R	R	Y	Y	Y	R	W	G	W	A	C	Y	W	Y	R	R	Y	R	0,9
9502	Ireland	R	R	R	A	G	G	T	T	C	A	W	G	W	A	C	Y	W	Y	R	R	Y	R	0,9
9504	India	G	A	A	C	G	G	T	T	C	A	A	A	T	G	T	C	T	T	A	A	C	A	0
9511	Russia	A	A	G	A	A	A	C	C	T	G	W	G	W	A	C	Y	W	Y	R	R	Y	R	0,9
9512	Russia	R	R	R	A	G	G	T	T	C	A	A	G	T	A	C	T	A	C	G	G	C	G	9
9506	India	R	R	R	M	R	R	Y	Y	C	A	W	R	W	G	T	C	T	T	A	A	C	A	0
9242	Ecuador	G	A	R	M	R	G	Y	Y	Y	R	T	G	A	A	C	C	T	T	A	R	C	A	0
7498	USA	G	A	R	M	R	G	Y	Y	C	R	T	G	A	A	C	C	T	T	A	G	C	A	0
7379	India	R	R	R	M	R	R	Y	Y	C	A	W	R	W	G	T	C	T	T	A	A	C	A	0
9290	India	G	A	R	M	R	G	Y	Y	C	R	T	G	A	A	C	C	T	T	A	G	C	A	0
9503	India	R	R	R	M	R	R	Y	Y	C	A	W	R	W	G	T	C	T	T	A	A	C	A	0
7551	Australia	R	R	R	A	G	G	T	T	C	A	A	G	T	A	C	T	A	C	G	G	C	G	9
9333	China	G	A	R	A	R	G	Y	Y	Y	R	A	G	T	A	C	Y	W	Y	R	R	C	R	9
9351	Vietnam	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	A	R	T	R	C	C	T	T	A	A	C	A	0,9
9256	Finland	G	A	R	A	R	G	Y	Y	Y	R	A	G	T	A	C	Y	W	Y	R	R	Y	R	0,9
9508	Poland	R	A	G	A	A	R	C	C	T	G	T	G	A	A	C	Y	W	Y	R	R	Y	R	0,9
9510	MZ	R	A	R	A	R	R	Y	Y	Y	R	W	G	W	A	C	Y	W	Y	R	R	C	R	9
9513	CZ	A	R	G	A	R	R	Y	Y	Y	R	A	G	T	A	C	C	T	T	A	A	C	A	0
9514	Austria	R	A	R	A	R	R	Y	Y	Y	R	A	G	T	A	C	Y	W	Y	R	R	Y	R	0,9
9560	Hungary	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	A	A	T	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
9622	Germany	R	R	G	A	R	G	Y	Y	Y	R	T	G	A	A	C	Y	W	Y	R	R	C	R	0,9

Table 3.5: Representative SNP sequencing results for NB-ARC-related windows from Sp02G05200, Sp13G03150, and Sp17G03410. CZ is an abbreviation of Czech Republic and MZ is an abbreviation of Mozambique. Column headings above SNP calls correspond to positions in the 9509 reference genome full length PCR product for each primer set. The following IUPAC nucleotide codes are used for heterozygous loci: W = A,T; R = A,G; Y = C,T; M = A,C. NA indicates no amplification. Sp17-SNP also exhibited a 9 bp INDEL at position 343. Relative to the 9509 reference genome, the presence of this INDEL corresponds to an insertion of a 9 bp sequence. A 9 indicates the clone is homozygous for the insertion relative to the 9509 reference sequence, a 0 indicaties the clone is homozygous for the absence of the insertion relative to the 9509 reference sequence, and 0,9 indicates it is heterozygous for the INDEL.

this study can be deployed as a simple genotyping tool to positively distinguish between these two *Spirodela* species by a simple PCR assay. This will be far superior in ease and economy than previous barcoding or AFLP strategies.

3.3.5 Distance Analysis of *S. polyrhiza*

Using only length polymorphism data from four markers, 18 out of 23 *S. polyrhiza* clones displayed unique fingerprints by clustering analysis (Fig. 3.6). The German 9622 and Austrian 9514 clones share the same fingerprint, while the three Indian clones 9503, 9506 and 7379 were indistinguishable from one another as shown on the dendrogram. Clones within a clustering of $\leq 25\%$ dissimilarity differ from clones outside of the cluster by at

least two markers. The combined SNP and length polymorphism dendrogram further revealed unique fingerprints in 20 out of 23 clones under investigation (Fig. 3.7). Clones that are separated in clusters by $\geq 25\%$ dissimilarity differ from one another in at least six markers. Although the STY study [Bog et al., 2015] demonstrated that clones 9506, 9503, and 7379 have distinct phenotypes, we could not find differences for these three clones using the seven hyper-polymorphic markers.

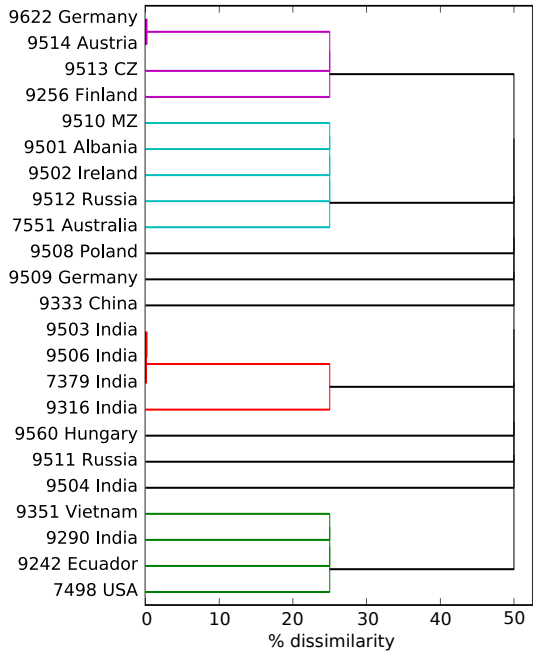

Figure 3.6: Dendrogram of 23 *S. polyrhiza* clones based on length polymorphism markers. X-axis represents distance between clusters. *S. polyrhiza* clones along the Y-axis. Clades formed at 25% or less dissimilarity are colored the same.

3.4 Discussion

Previous attempts at distinguishing *S. polyrhiza* with PCR-based genotyping markers have proved to be challenging. No difference was found in a prior study of 24 *S. polyrhiza* clones using the plastidic markers *rpl16* and *rps16*, while only 4–17% of clones can be differentiated using AFLP [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. However, with the

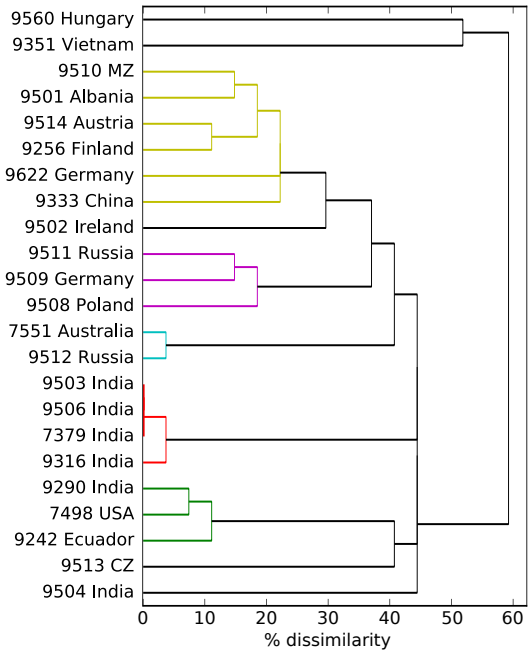

Figure 3.7: Dendrogram of 23 *S. polyrhiza* clones based on SNPs, INDELs, and length polymorphisms from all markers. Distance between clusters on the X-axis, *S. polyrhiza* clones represented along Y-axis. Clones in clusters of less than 25% dissimilarity are colored the same. These clusters consist of clones that are differentiated from one another by only a few markers.

recent availability of high quality genome data for this species [Michael et al., 2017], it became possible to leverage these new resources 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, RFLP or AFLP markers, is that background non-informative (hypo-polymorphic) targets can be minimized while the known target genes 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 in NB-ARC-related loci to generate markers that can discriminate among a set of nine *S. polyrhiza* clones selected from a previous STY study that have been re-sequenced [Michael et al., 2017, Kuehdorf et al., 2014]. In addition, by resequencing the five clones of *S. polyrhiza* that were not resolved by our fragment length markers, we can include these sequences into our informatics pipeline

together with the nine previously sequenced genomes from our training set. This "iterated training" of our database could potentially increase the resolving power of our approach by identifying additional discriminatory targets that can be added to our genotyping set for this species.

We selected and tested seven regions in NB-ARC-related genes on chromosomes 2, 12, 13 and 17. 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. Perhaps its polymorphic nature is related to more frequent occurence of micro-rearrangements, gene duplications and recombination events commonly seen in clustered NB-LRR genes [Meyers et al., 2003]. These windows were intended to amplify from one locus (mapped uniquely to the 9509 genome by BLASTN), but it is 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 may help resolve this issue. 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, the technology is 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 (Fig. 3.3, Fig. 3.5, Table 3.3). 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 originate 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, but 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. One possible explanation is that these three clones diverged from the other Indian clones in our analysis due to differences in pathogen pressures; however, this hypothesis would require further research. Alternatively, these clones may differ at the epigenetic level and thus their respective phenotype could potentially arise from changes independent of DNA sequence *per se.* Our NB-ARC derived markers were trained on nine clones which included 9506 but not 7379 nor 9503. 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. It is important to point out that our data set contained six clones collected from India, the highest concentration represented for a geographic region (26%), half of which were not included in the original training set.

For the European clones, clones 9509 and 9622 might have been expected to have

similar fingerprints since they are both from Germany. 9509 was originally collected from Lotschen, Stadtroda and 9622 originated in Baden-Wurttemberg, approximately 426km 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). These results would indicate that clones 9509 and 9622 may have spread into Germany via different ancestors that have passed through different countries in Europe.

Hungarian clone 9560 is peculiar, being distinct from all other clones and only successfully amplifying in four of the seven markers tested. To ensure that this clone wasn't incorrectly typed or a result of contamination, its barcode was verified using the *psbK* plastid barcoding marker (data not shown). Hungary is bordered on most sides by mountainous regions, so perhaps its unique genotype from other *S. polyrhiza* populations resulted from its geographic isolation. An adaptation of this clone to its local pathogen pressure could explain its unique NB-ARC genotype. Further testing of local clones from Hungary would provide more evidence to support this observation. In addition, the Vietnamese clone 9351 did not amplify at all using Sp02-SNP, but amplified with the other markers.

Using combined length polymorphism and SNP data improved our cluster analysis. The analysis based on combined length polymorphism and SNP data suggests that 9511, the Russian clone from Moscow, is more similar to European clones than to the other Russian clone originally collected from the Olkha river in Shelekhov, approximately 5207 km east of Moscow. 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 *Landolita 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.

3.5 Conclusion

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]. Our study here presents a novel approach to systematically identify and select targeted markers based on comparative analysis of polymorphism between multiple sequenced genomes for a species of interest. By targeting the highly polymorphic NB-ARC-related gene family, which is well conserved in plants, and maximizing for variants across nine *S. polyrhiza* genomes, we aim to increase the signal to noise in the amplification procedure where random loci in the genome with low polymorphism will not confound our analysis, as in the case of random primer-driven approaches. 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 could 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 loci [Sommer, 2005].

Although in the present study, we are only able to distinguish 20 out of the 23 clones studied, we believe the inclusion of a larger, and potentially more diverse genome dataset from additional clones to our training set for marker selection could lead to more powerful NB-ARC-related markers with greater resolution. This may allow us to probe into this interesting gene family and potentially associate with other demographic differences that would not be possible with markers that amplify random sequences such as AFLP, RFLP, and SSRs. 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 its divergence of NB-ARC-related gene evolution.

3.6 Funding

This work was supported by the National Science Foundation Integrative Graduate Research and Education Traineeship Renewable and Sustainable Fuel Solutions for the 21st Century [DGE-0903675]; and Hatch grant from the New Jersey Agricultural Experimental Station at Rutgers University [project #12116].

3.7 Acknowledgements

We would like to thank Todd Michael for bioinformatics analysis and assistance in accessing data files, Ryan Gutierrez for technical support, 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 exploring various degenerate primer sets against the conserved regions of the NB-LRR genes in duckweed. His early mentorship of P.C. in this endeavor is gratefully acknowledged.

Chapter 4

Applying polymorphism analysis pipeline to multiple *Arabidopsis thaliana* genomes

Authors: Philomena Chu, Glen M. Wilson, Eric Lam

4.1 Introduction

Our goal in this chapter is two-fold: (1) to determine if the bioinformatics pipeline devised in Chapter 3 can be applied to other plant genomes, (2) to evaluate how informative and meaningful our ranking system from the bioinformatics pipeline is. The bioinformatics pipeline and predictions worked fairly well for *Spirodela polyrhiza*, although we encountered some issues with false positives when predicting the presence of SNPs and INDELs. We wondered if our workflow would be successful on a species with better characterized genome datasets, such as *Arabidopsis thaliana*. The 1001 genomes project provides a rich and varied resource from which we could harvest polymorphisms such as SNPs and INDELs from a genetically diverse collection of naturally inbred *A. thaliana* lines `http://1001genomes.org` [Consortium, 2016].

4.2 Materials and Methods

4.2.1 *A. thaliana* genomes

We downloaded the VCF data set for all 1,135 *A. thaliana* accessions from the 1001 genomes project, `http://1001genomes.org`, that resulted from the intersection of the GMI-GATK and MPI-SHORE pipeline [Consortium, 2016] with a minimum quality value of 25. There appeared to be no heterozygous calls (SNPs or INDELs) in any of the VCF files that we examined. From this set, we selected 49 geographically diverse accessions

from each of the nine clusters that were previously identified in the 1001 Genomes Consortium paper [Consortium, 2016] to act as our training set: four accessions from each of the eight non-relict clusters: Western Europe (Seattle-0 #7332, For-2 #5741, Ag-0 #6897, IP-Pro-0 #9571); Germany (Rsch-4 #7322, Mn-0 #7255, Wu-0 #7415, Ca-0 #7062); Northern Sweden (Tamm-2 #6968, Eden-1 #6009, Bil-5 #6900, Ost-0 #9351); Southern Sweden (Spro3 #9452, T1080 #6098, Hau-0 #7164, Fja2-4 #6021); Central Asia (Kas-2 #8424, Valm #15560, Yeg-1 #10011, Ms-0 #6938); Italy/Balkans/Caucasus (Ivano-1 #9701, Kastee-1 #10006, Basen-1 #9647, Epidauros-1 #9725); Central Europe (Bik-1 #9761, Dobra-1 #10018, Kn-0 #7186, Bai-10 #9779); Iberian Peninsula (Mer-6 #9946, IP-Den-1 #9539, IP-Vaz-0 #9593, IP-Urd-1 #9901). We further selected relict groups (populations of ice age refugia near the Mediterranean sea that occupied post-glacial Eurasia first; #9606 Aitba-1, #7063 Can-0, #6911 Cvi-0), and admixed accessions with the primary origin country from each of the clusters (Bur-0 #7058, Tsu-0 #7373, Kz-13 #6830, Qar-8a #9764, Etna-2 #9762, Iasi-1 #9744, Amel-1 #6990, Ler-0 #7213, Ty-1 #5784, Oy-0 #7288, Strand-1 #10023, Hovdala #6039, IP-Mos-1 #9508, IP-Orb-10 #9565). At least one accession from each of the eight non-relict clusters had to be 100% comprised of the cluster that it belonged to. North American and British accessions were included (Seattle-0, For-2, Ty-1) because of their identification in recent long-range dispersal [Consortium, 2016].

4.2.2 Bioinformatics analysis

All analyses were run on the Epigenome server in the Lam lab.

NB-LRR identification and window selection

The NB-LRRs and NB-LRR-related protein-encoding genes used in this study were from previous curations, with three exceptions, AT1G51485, AT1G58842, AT3G25515, which were identified as having misannotations or errors [Tan et al., 2007, Meyers et al., 2003], and did not have entries in the *Arabidopsis thaliana* database (`https://www.arabidopsis.org`) as of April 2018. We searched for possible homologous genes in the

neighboring regions around these three genes and made the following changes to our NB-LRR list: AT1G51485 was replaced with AT1G51480, AT1G58848 replaced AT1G58842, and AT3G25515 was replaced with AT3G25510. Although we used protein-encoding genes that may deviate from the canonical nucleotide-binding leucine-rich repeat structure observed in functional proteins, we refer to our curated list as "NB-LRRs" for the sake of simplicity.

Identification of the most polymorphic NB-LRR windows was conducted in a similar workflow as described in sections 3.2.2–3.2.3. Briefly, conserved regions across all 49 *A. thaliana* genomes of the test set were first identified. These were designated as potential locations for primers. Then regions of length 200–900 bp flanked by the potential primer locations (conserved regions of at least 20 bp at the 5' and 3' ends) were identified. We refer to these regions as "windows." Windows were input into Primer3 (version 4.0.0) [Untergasser et al., 2012]. Primers with a penalty score of at most 2 were kept for downstream analysis.

The general workflow for SNP and length polymorphism analyses followed as such:

1. Identify and test hyperpolymorphic NB-LRR windows.

    (a) Use the bottleneck and max-min methods for INDEL polymorphism and the SNP method, summarized in figure 3.1, using the test set of 49 *A. thaliana* accessions described in section 4.2.1. The top-ranked windows were manually curated, where necessary, using the following criteria: if a window overlaps with any of the five windows either directly above or below in the rankings and has the same distinguishing power, then the smallest window is kept and the other overlapping window is filtered out.

    (b) Bioinformatically test the distinguishing power of the top-ranking windows from the previous step on the full set of 1,135 accessions (number of pairs of accessions on the full set of accessions that a window can distinguish). For the INDEL analyses, calculate the pairwise differences of fragment length of a given window using the VCF files for the 1,135 accessions. For a window produced from the SNP analysis, for every pair of accessions calculate the

number of distinct bases between the predicted fragments using the VCF files for the 1,135 accessions.

2. Test efficacy of the ranking algorithms. Determine whether or not the metrics used for ranking windows on the test set (49 accessions) can be extrapolated to the full set (1,135 accessions). To accomplish this goal, we calculate the Pearson correlation coefficient (r) for the average number of SNPs or INDELs for a window on the training data against the full data set. The Pearson correlation coefficient will also be calculated for the distinguishing power of windows on the training set versus the full data set. 95% confidence intervals are produced using bootstrapping with replacement on either 10,000 (for SNP analysis) or 50,000 replicates (for INDEL analysis).

3. Build families of windows that distinguish as many accessions as possible for future wet lab testing. Determine which windows can be used together for maximal distinguishing power on all pairs of accessions for each of the top 10 ranked windows in SNP and INDEL methods.

SNP analysis

Pairwise SNP counts for NB-LRR windows were tallied, as described in section 3.2.3. For a NB-LRR window, a minimum of at least two SNPs for a pairwise comparison between accessions constitutes a distinguished pair.

Length polymorphism: Bottleneck and Max-min

Bottleneck and max-min methods were carried out and the outputs were ranked according to number of distinguished pairs and average INDEL metrics. The average INDEL metric in the max-min method is taken to be the average distance of an accession's window length compared to the reference window length. We define "average indel" in the bottleneck method as the average pairwise bottleneck distance. Further details are described in detail in section 3.2.2.

4.3 Results

4.3.1 SNPs

Predicting hyperpolymorphic NB-LRR windows

The SNP analysis begins with the 12,860 windows with sizes ranging from 200–900 bp and a maximum Primer3 penalty of 2 identified from the list of 174 NB-LRRs. These windows were then ranked with the SNP methodology applied to the test set of 49 accessions listed in section 4.2.1. The top 1.5% windows from this analysis are predicted to distinguish at least 99.5% pairs (1170 out of 1196 pairs) of the 49 accessions. The raw data output from the rankings revealed some overlap in neighboring top-ranked windows. To correct for this, we analyzed the top-ranked windows that may have significant overlap with the five windows immediately ranked 5 places above and 5 places below. If a window overlapped more than 30% of its total length with another window in this cluster and it had the same distinguishing power as the other window, then the smallest window with the largest percentage of distinguished pairs was kept and the other associated windows were filtered out. From this analysis, we found that the top 10 hyperpolymorphic windows originate from five different NB-LRR genes (Table 4.1). Each of the top three windows, AT1G61190 from chromosome coordinates 22550328–22560226 and AT3G44630 from 16198718–16199602 and from 16199571–16200335, could make distinctions between all 1,176 possible pairs of accessions, with an average of approximately 10–16 SNPs between pairs of accessions. The next seven windows can distinguish 99.7% of all possible pairs of accessions from the test set.

Validation of predicted hyperpolymorphic windows in *A. thaliana*

We predicted that the best ranking windows from the SNP analysis on the test set of accessions would be able to distinguish most, if not all, of the accessions from the full data set of 1,135 accessions. To demonstrate this, we tested the top 10 ranked windows on the 1,135 accessions using the available VCF files. These windows performed reasonably well, according to the the number of distinguished pairs and associated percentages (Table

NB-LRR	start	stop	window size	SNPs	DP	%
AT1G61190	22559328	22560226	899	15.6	1176	100
AT3G44630	16198718	16199602	885	11.2	1176	100
AT3G44630	16199571	16200335	765	9.6	1176	100
AT1G31540	11289147	11290044	898	20.0	1175	99.9
AT1G31540	11290024	11290523	500	18.2	1175	99.9
AT5G41750	16695351	16696098	748	17.9	1175	99.9
AT5G48620	19719803	19720508	706	11.2	1175	99.9
AT5G48620	19719897	19720744	848	16.3	1174	99.8
AT1G61190	22559121	22559963	843	13.9	1174	99.8
AT5G48620	19720041	19720744	704	13.8	1173	99.7

Table 4.1: The rankings of top 10 windows identified from test set selected for SNP analysis. Column headings: start and stop coordinates of NB-LRR windows, "SNPs" is the average number of SNPs, "DP" is the number of distinguished pairs, "%" stands for percentage of pairs that are distinguished, i.e., DP/1176 .

4.2). The top three high-ranking windows that could bioinformatically distinguish 100% of the pairs from the test set could successfully distinguish 98.5–99.9% of the 643,545 possible pairs from the full set. A high percentage of distinguishing power is maintained over the next seven windows ranging from 98.7% to 99.1%.

NB-LRR	SNPs	DP	%
AT1G61190	18.3	642638	99.9
AT3G44630	11.5	633655	98.5
AT3G44630	9.76	635857	98.8
AT1G31540	20.3	637730	99.1
AT1G31540	19.0	636633	98.9
AT5G41750	16.9	635573	98.8
AT5G48620	15.7	641740	99.7
AT5G48620	17.6	639960	99.4
AT1G61190	13.6	639821	99.4
AT5G48620	15.0	635236	98.7

Table 4.2: The rankings of top 10 windows from test set selected for SNP analysis applied to 1,135 accessions (full set). These windows correspond to the windows listed in Table 4.1. Column "SNPs" stands for average number of SNPs, column "DP" stands for number of distinguished pairs, and % is the percentage of distinguished pairs, i.e., DP/643545.

Demonstrating the efficacy of the ranking algorithms

We now demonstrate the efficacy of the SNP ranking methodology. We randomly sampled windows and compared their distinguishing power on the test set of accessions to their distinguishing power on the full set of accessions to verify that our method of ranking windows is effective at locating hyperpolymorphic windows. Specifically, we will show that the values of the metrics "number of distinguished pairs" and "average number of SNPs" when evaluated on the test set of accessions are linearly correlated to their values on the full set of accessions. This will be shown with the help of the Pearson correlation coefficient. Since the distribution of our data is unknown, we perform bootstrapping with replacement (10,000 repetitions) on 200 randomly selected windows to obtain a 95% confidence interval for the Pearson correlation coefficient. We find that the Pearson correlation coefficient for average number of SNPs and number of distinguished pairs is close to 1 with highly significant p-values and 95% confidence intervals of 0.96–0.99 and 0.96–0.98, respectively (Table 4.3, Fig. 4.1). Thus, a top-ranking window from the SNP analysis on the test set will be much more likely to have greater distinguishing power when tested on the full data set than a lower-ranking one.

SNPs (r)	p	CI	DP (r)	p	CI
0.98	2.5×10^{-137}	0.97–0.99	0.97	4.8×10^{-120}	0.96–0.98

Table 4.3: Pearson correlation calculations for SNP analysis on test set extrapolated to full set. Two hundred randomly selected windows were chosen from the set of 12,860 windows. Column headings: "SNPs (r)" means the Pearson correlation for average number of SNPs for the 200 randomly selected windows, "p" lists the p-value for the Pearson correlation calculation, "CI" shows the 95% confidence interval obtained from a bootstrap with replacement analysis with 10,000 repetitions, and "DP (r)" is the Pearson correlation for number of distinguished pairs calculated for the 200 randomly selected windows.

Hyperpolymorphic families of windows from the SNP analysis

In our previous SNP analysis, we consider a difference of two SNPs to be the minimum amount of SNPS to be a distinguished pair, which we'll call "resolution 2." The window in AT1G61190 alone can bioinformatically distinguish 99.9% of all pairs of accessions from the full set (Table 4.2), but we wanted to investigate the possibility of achieving 100%

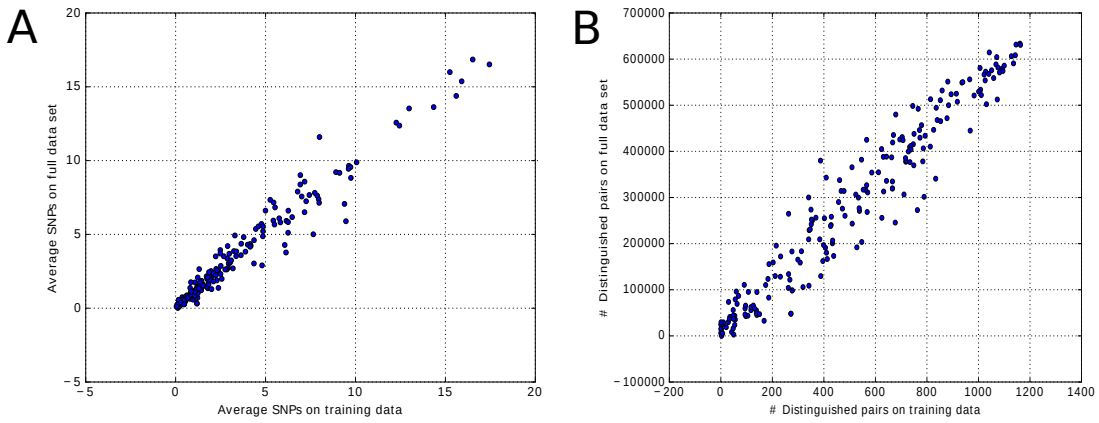

Figure 4.1: Metrics from SNP analysis on training data set (49 accessions) vs. full data set (1,135 accessions) on 200 randomly selected windows are linearly correlation. (A) Average number of SNPs from training set (x-axis) plotted against average number of SNPs on full data set (y-axis). (B) Number of distinguished pairs on training data (x-axis) vs. number of distinguished pairs on full data set (y-axis).

distinction of all pairs of accessions from the full set. In other words, we wanted to know which primers could be used in conjunction with the goal of achieving 100% distinction. We'll refer to these combinations as "families." Families are constructed following a greedy algorithm strategy. Starting with a top ranked window, each additional window that is added to the family should maximally increase the number of distinguished pairs that the windows in the family can distinguish. For practical reasons, we set the maximum allowable number of windows belonging to a family to be 10 and only the top-ranked 200 windows were included in this part of the analysis.

If we take the top ranked window, AT1G61190, and investigate the best complimentary window (out of the top 200 windows) with the goal of distinguishing as many accessions as possible, and if we increase the requirement of the number of SNPs it would take to separate the two accessions, it is not surprising that the number of indistinguishable accessions increases quickly (Table 4.4). The AT1G61190 and AT4G16860 (the 171th highest ranking window) windows together can separate all but two pairs of accessions from the full set with a requirement of at least one SNP difference. Both sets of undistinguished pairs include American accessions only. The best pairs of windows with resolution 2 have 28 indistinguishable pairs. These pairs group into eight bins that include two groups of

indistinguishable American accessions, a pair of UK accessions, three southern Sweden bins with 2–3 accessions, a southern Sweden and Spanish accession, and a mixed bin of accessions (3 from UK, 1 USA, 1 Italy, 1 Germany, 1 southern Sweden) (data not shown). With a resolution of 3, the number of indistinguishable pairs goes up to 97, grouped into eight bins: a pair each of US and France, two pairs from southern Sweden, a cluster of five from the UK, a group of three, another of nine from the US, and a mixed group (data not shown). The most complimentary window at resolutions 2 and 3 come from the same NB-LRR AT5G43470, ranked 56th and 55th respectively, with an overlap of 765 bp.The requirement of a minimum of at least three SNPs to separate shifted the window.

Resol	NB-LRR	start	stop	Length	Ranking	U	%
	AT1G61190	22559328	22560226	899	1		
1	AT4G16860	9490614	9491481	868	171	2	3.11×10^{-6}
2	AT5G43470	17463307	17464165	859	57	28	4.35×10^{-5}
3	AT5G43470	17463175	17464072	898	56	97	1.51×10^{-4}

Table 4.4: Best pair of windows in the SNP analysis for the top ranked window from AT1G61190 at distinguishing accessions from the full set at different resolutions. Top ranked AT1G61190, listed in the first data row, is paired with the window listed in row with resolution 1–3 and their joint distinguishing power is calculated. "Resol" stands for the minimum number of SNPs that are required to distinguish a pair of accessions. NB-LRR start and stop coordinates are listed, as well as their ranking position in our method. The column labeled "U" lists number of indistinguishable pairs for the pair of windows, and the column % shows the percentage of indistinguishable pairs for the relevant pair of windows, i.e., U/643545.

Next, we examined which windows could be used in conjunction in order to best distinguish all pairs from the 1,135 accessions at different resolutions without restrictions of the number of primer sets used. Using a minimum of one SNP as a requirement to distinguish between two accessions, the #2 ranked window in AT3G44630 can be used with the previously identified windows in AT1G61190 and AT4G16860 to bioinformatically differentiate between all 1,135 accessions (Table 4.5). Differentiating pairwise accessions at resolutions 2 and 3 is more challenging, thus requiring four and six primer sets, respectively.

Resol	NB-LRR	start	stop	Length	Ranking	U	%
	AT1G61190	22559328	22560226	899	1		
1	AT4G16860	9490614	9491481	868	171		
	AT3G44630	16198718	16199602	885	2	0	100
2	AT5G43470	17463307	17464165	859	57		
	AT4G11170	6811334	6812020	687	111		
	AT3G44630	16198718	16199602	885	2	0	100
3	AT5G43470	17463175	17464072	898	56		
	AT5G48620	19719616	19720508	893	9		
	AT3G44630	16198718	16199602	885	2		
	AT3G46730	17214453	18215349	897	114		
	AT1G31540	11289147	11290044	898	4	0	100

Table 4.5: Best family of windows in the SNP analysis for top ranked AT1G61190 at distinguishing full set of accessions at different resolutions. Combinations of windows with the top ranked AT1G61190 window are built to maximize distinguishing power on the full set of accessions. The column "Resol" lists the minimum number of SNPs that are needed to distinguish a pair of accessions. NB-LRR start and stop coordinates are listed, as well as the window's ranking in our method. "U" lists the number of indistinguishable pairs from the full set of accessions using the family of windows in that section of the table with the top ranking window AT1G61190. % is the percentage of indistinguishable pairs for the family of windows, i.e., U/643545.

4.3.2 INDELs

Predicting hyperpolymorphic NB-LRR windows

The INDEL analyses begin with the 12,860 windows with sizes ranging from 200–900 bp and a maximum Primer3 penalty of 2 identified from the list of 174 NB-LRRs, as with the SNP analysis. These windows are then ranked in two different ways, following the bottleneck method and the max-min method of section 3.2.2. After filtering out significantly overlapping windows with the same distinguishing power, the top 10 ranking windows from the INDEL bottleneck and max-min analyses on the test set of 49 accessions reveal some interesting patterns (Table 4.6). We observe several of the same NB-LRRs with more than one occurrence on each list: AT2G17050, AT5G45200, AT3G25510, AT4G08450, AT2G51480. Furthermore, there are NB-LRRs that occur on both lists: AT2G17050, AT5G45200, and AT3G25510. Examination of the rest of the rankings may reveal other NB-LRRs that are hyperpolymorphic. In the bottleneck analysis, we quantify the INDEL variation of a window by considering the average pairwise

bottleneck distance over all pairs of accessions; for brevity, we refer to this quantity as average INDELs. For the max-min method, we define "average INDELs" as the average distance of a window from the accession to the reference window's length.

Amongst the top 10 ranked windows from the bottleneck method, the average pairwise bottleneck distance ranged from 0.69 to 2.65, with approximately 45–63% of the 1176 pairs in the test set being differentiated (Table 4.6). The highest ranking window, located in AT2G17050, could bioinformatically distinguish approximately 63% of the test pairs with an average pairwise bottleneck distance of 1.18, which is interestingly not the highest (2.65).

In the max-min analysis, the average INDEL distance for the top 10 ranked windows ranged from 1.88–3.88, and the window with the highest average INDEL score (ranked #10), AT3G25510, could distinguish approximately 53% of test pairs compared to the approximately 73% of the highest ranked window, AT5G45200.

Validation of predicted hyperpolymorphic windows in *A. thaliana*

Next, we tested the performance of the top ranking windows from Table 4.6 on the full set of 1,135 accessions. These windows perform similarly well on the full set of accessions as compared to the test set, differing by no more than approximately 10% (Table 4.6, 4.7). In the bottleneck method, most windows had slightly worse percentages of distinguished pairs of accessions on the full set than on the test set, however, AT3G25510 windows and the 6th ranked AT5G45200 had a higher percentage of distinguishable pairs on the full set than on the test set of accessions. Several windows in the max-min method had a greater percentage of distinguished pairs on the full set than on the test set.

Demonstrating the efficacy of the ranking algorithms

To test if our ranking algorithms for the INDEL analyses could be extrapolated from the test set to the full set of accessions, we randomly selected 500 windows and assessed their performance on the full set of 1,135 accessions. The Pearson correlation (r) was calculated for the average INDELs metric and the number of distinguished pairs for the random sample with both methods. A 95% confidence interval for the Pearson correlation

Ranking	NB-LRR	start	stop	length	AvgINDELs	DP	%
1	AT2G17050	7414649	7415202	554	1.18	742	63.1
2	AT5G45200	18288813	18289542	730	1.61	683	58.1
3	AT3G25510	9267949	9268401	453	2.64	629	53.5
4	AT3G25510	9267504	9268249	746	2.65	626	53.2
5	AT2G17050	7414965	7415568	604	0.76	588	50.0
6	AT5G45200	18288337	18289165	829	1.26	555	47.2
7	AT4G08450	5367304	5367910	607	0.87	554	47.1
8	AT4G08450	5367785	5368606	822	0.70	538	45.7
9	AT4G08450	5367746	5368524	779	0.70	536	45.6
10	AT4G08450	5367552	5367910	359	0.69	534	45.4
1	AT5G45200	18288813	18289542	730	2.53	857	72.9
2	AT3G25510	9263351	9264179	829	2.45	792	67.3
3	AT3G25510	9263351	9263992	642	2.33	765	65.1
4	AT3G25510	9263351	9263729	379	2.29	749	63.7
5	AT2G17050	7414649	7415202	554	2.08	742	63.1
6	AT1G51480	19093369	19094052	684	2.29	730	62.1
7	AT1G31540	11289147	11290044	898	1.51	663	56.4
8	AT1G51480	19093541	19094143	603	2.02	659	56.0
9	AT5G45200	18288337	18289165	829	1.88	633	53.8
10	AT3G25510	9267504	9268249	746	3.88	630	53.6

Table 4.6: The rankings of the top 10 windows for INDELs-bottleneck (first 10 entries) and max-min analysis (last 10 entries) on the test set of accessions. The column "length" shows the length of the window in Col-0 TAIR10 reference assembly, "AvgINDELs" lists the average number of INDELs, column "DP" is the number of distinguished pairs, while % represents the percentage of pairs of accessions distinguished by the given window, i.e. DP/1176.

coefficient was determined for both methods by a bootstrap analysis with replacement with 50,000 repetitions. The Pearson correlation coefficients for average INDELs and number of distinguished pairs were close to 1 (Table 4.8, Figure 4.2, and Figure 4.3), establishing a linear relationship among the metrics average INDELs and number of distinguished pairs from the test set to the full set of accessions. Both r values are strongly significant, and have confidence intervals of 0.88–0.97 and 0.93–0.97 for bottleneck and max-min methods, respectively.

Building families of windows to distinguish accessions

Following section 4.3.1, we sought to build families of windows with the goal of distinguishing all of the accessions in the full set of 1,135 accessions. Since we had already

Ranking	NB-LRR	AvgINDELs	DP	%
1	AT2G17050	1.01	345738	53.7
2	AT5G45200	1.42	355219	55.2
3	AT3G25510	3.04	408129	63.4
4	AT3G25510	3.04	407773	63.3
5	AT2G17050	0.63	269844	41.9
6	AT5G45200	1.31	322535	50.0
7	AT4G08450	0.70	278639	43.3
8	AT4G08450	0.68	275563	42.8
9	AT4G08450	0.69	279107	43.3
10	AT4G08450	0.69	282048	43.8
1	AT5G45200	2.36	457854	71.1
2	AT3G25510	3.3	477342	74.2
3	AT3G25510	2.92	435474	67.7
4	AT3G25510	2.77	419319	65.2
5	AT2G17050	1.50	345738	53.7
6	AT1G51480	2.13	397965	61.8
7	AT1G31540	1.64	364478	56.6
8	AT1G51480	1.90	377486	58.7
9	AT5G45200	1.88	390126	60.6
10	AT3G25510	5.47	411415	63.9

Table 4.7: The performance of the top 10 windows from the INDEL bottleneck method (first 10 entries) and the max-min analysis (last 10 entries) applied to 1,135 accessions. Column "INDELs" lists average number of INDELs, column "DP" shows number of distinguished pairs, and % lists the percentage of distinguished pairs, i.e., DP/643545.

identified the top ranking INDEL windows on the test set (Table 4.6), our task was to determine the minimum combination of windows with maximal distinguishing power. We wanted to test how the combinations of windows with the greatest distinguishing power on the test set would fare on the full set.

At resolution 1, all constructed families for the top 200 windows could distinguish all but one pair of accessions in the test set (Table 4.9). At resolutions 2 and 3, six pairs of accessions out of the 1176 pairs were indistinguishable from one another. Requiring at least four INDELs as a minimum to be considered distinct increased the number of indistinguishable pairs to 18–25, with those pairs being grouped into 1–5 bins. The number of windows that were required to distinguish as many accessions as possible (with a maximum of 10 windows) also increased from resolution 1 through 4.

We then took the families from the top 10 windows and applied them to the full set of

Windows	AVGindel (r)	p	CI
Bottleneck 500 random	0.94	4.52×10^{-238}	0.88–0.97
max-min 500random	0.96	9.68×10^{-276}	0.93–0.97

Windows	DP (r)	p	CI
Bottleneck 500 random	0.93	6.98×10^{-221}	0.89–0.96
max-min 500random	0.95	1.14×10^{-260}	0.92–0.97

Table 4.8: Pearson correlation calculations for INDEL bottleneck and max-min analyses. 500 windows were randomly selected from the 12,860 identified windows. Column headings: "AVGindel (r)" lists the Pearson correlation for the average INDEL metrics (average pairwise bottleneck distance in the bottleneck analysis and the average distance of ecotype windows length to the reference window length in the max-min analysis), "p" is the p-value of the Pearson correlation calculation, "CI" is the 95% confidence interval obtained from bootstrap analysis with replacement at 50,000 repetitions, "DP (r)" lists the Pearson correlation for number of distinguished pairs.

1,135 accessions to test their distinguishing power. Less than 1% of pairwise accessions are indistinguishable up to resolution 3, at which point 1.18% of accession pairs are inseparable (Table 4.9). Minimum requirements at resolution 4 are more challenging, as reflected in the larger percentages of indistinguishable pairs (1.62–1.83%) and smaller number of bins that these pairs are binned into, compared with the number of indistinguishable pairs and bins at resolution 1–3.

Resol	#windows	Ut	Bins	Uf	Bins
1	4–5	1	1	3201–3487	161–172
2	8–9	6	1	4473–4513	138
3	8–10	6	1	7604	117
4	10	18–25	1–5	10430–11846	73–104

Table 4.9: Distinguishing power of families from top 200 windows from the bottleneck analysis on the test set, and the top 20 windows on full set of *A. thaliana* accessions at different resolutions. "Resol" stands for the minimum number of INDELs that are needed to separate a pair of accessions. #windows is the number of windows needed to distinguish accessions established during testing on test set. "Ut" stands for the range of indistinguishable pairs on test set, and "Uf" stands for the range of indistinguishable pairs of full set. "Bins" lists the range of groups of indistinguishable accessions.

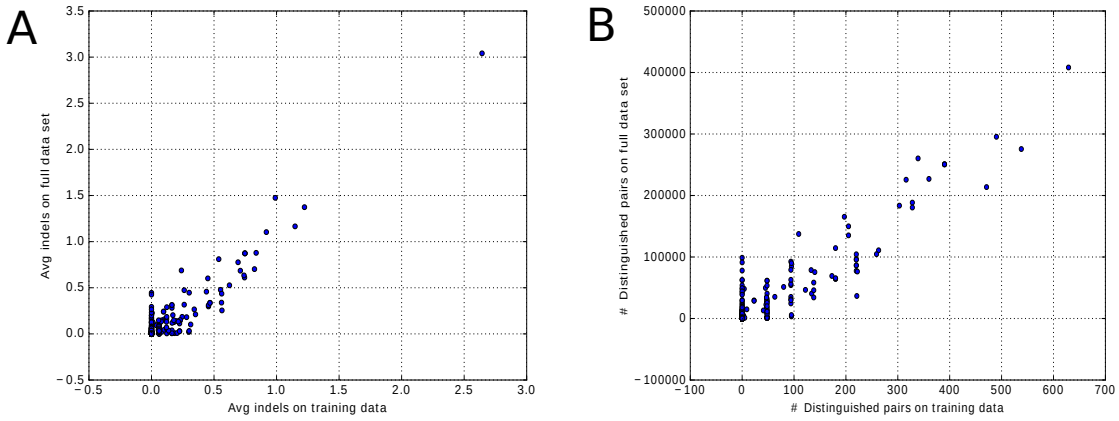

Figure 4.2: Results from INDEL bottleneck analysis on 500 randomly chosen windows. (A) Average number of INDELs (B) Number of distinguished pairs.

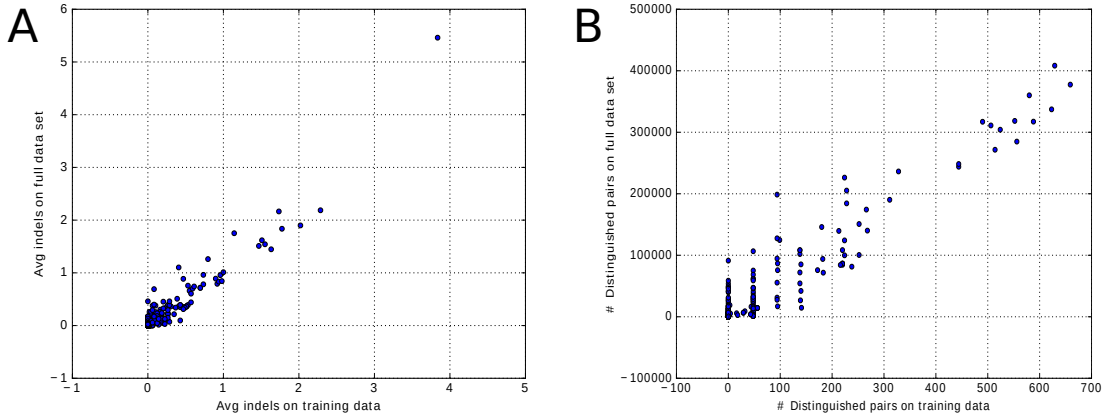

Figure 4.3: Results from INDEL maxmin analysis on 500 randomly chosen windows. (A) Average number of INDELs (B) Number of distinguished pairs.

Using the same requirements to construct best families of windows for the INDEL bottleneck analysis, we conducted a similar test with INDEL maxmin. Our INDEL maxmin analysis only used resolution 1 because pairwise distances seemed to be either 0 or 1. First, we built families of windows (maximum of 10 windows per family) for the top 10 ranked windows using combinations of the best 200 ranked windows. Each family contained 8–9 windows and left one pair of accessions undistinguished (Table 4.10). When we use these families on the full set of accessions, between 0.16% to 0.33% of pairwise accessions cannot be distinguished.

If we allow a maximum of 10 windows to distinguish the full set in the bottleneck

#windows	Ut	Bins	Uf	Bins
8–9	1	1	1018–2094	55–138

Table 4.10: Distinguishing power of families from top 200 windows from the max-min analysis on test set and top 20 windows on full set of *A. thaliana* accessions. #windows stands for number of windows needed to distinguish accessions established during testing on test set. The column "Ut" lists the range of indistinguishable pairs on test set, while "Uf" lists the range of indistinguishable pairs on the full set. "Bins" lists therange of groups of indistinguishable accessions.

analysis, then the families for the top 50 windows have a minimum of 343–380 pairs that cannot be distinguished, with the undistinguished pairs being grouped into 22–28 different bins at resolution 1 (Table 4.11). We then repeat this analysis for resolutions 2–4. The higher resolutions require a greater pairwise bottleneck distance to consider a pair of accessions distinct, hence the higher number of indistinguishable pairs compared to resolutions 1–3.

Resol	U	Bins
1	343–380	22–28
2	2168–2397	61–75
3	3447–3843	97–118
4	8724–9727	93–112

Table 4.11: Each row represents the properties of the families of windows for the top-ranked 50 windows from the INDEL bottleneck method. "Resol" stands for the minimum number of INDELs that are needed to separate a pair of accessions, the column labelled "U" lists the range of indistinguishable pairs observed . The column "Bins" shows the range of groups of indistinguishable accessions.

4.4 Discussion

Our analyses provide validation that our approach in predicting hyperpolymorphic NB-ARC genes in multiple *S. polyrhiza* genomes from Chapter 3 can be applicable to other plant species. This study also further demonstrates that the ranking algorithms from Chapter 3 do identify hyperpolymorphic windows that can have a broader applicability for distinguishing accessions or clones of a plant species beyond only those accessions or clones found in the training set.

Using SNP and INDEL sequencing data for a geographically diverse set of 49 *A.*

thaliana accessions, we identified hyperpolymorphic NB-LRR windows which could bioinformatically differentiate either a majority or all of the 1,135 *A. thaliana* accessions for which whole genome sequence data exists, depending on the technique of choice.

SNP analysis and resulting windows had greater distinguishing power than windows analyzed from either of the INDEL-based methods. The top 10 windows from our SNP ranking analysis from the test set of *A. thaliana* accessions came from only five different NB-LRR genes, while the INDEL analysis windows came from six NB-LRRs. This is somewhat striking, suggesting that these NB-LRR genes could be hotspots for SNP polymorphisms.

As we pointed out in the results section, several NB-LRRs were found on both INDEL-based bottleneck and max-min top 10 most hyperpolymorphic lists. Only one NB-LRR locus, AT1G31540, appeared on the top 10 SNP windows and top 10 INDELs-max-min windows list. When building families for the INDEL windows, perhaps if we allowed for primer sets with ranking greater than 200 to be allowed in the INDEL analysis, we could have distinguished 100% of the accessions bioinformatically.

There were similarities and differences during the procedure of building families of windows in SNPs vs. INDELs analyses. The procedure to build a family for both analyses was first worked out on the test set of accessions, then the distinguishing power of the family was checked on the full set. We had previously observed that top ranking SNP windows had greater distinguishing power than INDEL windows, making it necessary to use more windows to build each INDEL family. Our analysis could bioinformatically distinguish all pairwise accessions of the full set with SNP families, but less accessions could be distinguished with INDEL families. We took one further step in the INDELs bottleneck analysis by considering the full set of accessions as the training set with which to distinguish as many pairs of accessions as possible.

As we illustrated in Tables 4.4 and 4.5, at most two handfuls of SNP windows were successful at bioinformatically separating the full set of accessions, with certain combinations of primers matched with the #1 ranked AT1G61190 from coordinates 22559328–22560226. We therefore recommend experimenting with the primer sets in the windows listed in Table 4.12 for initial genotyping efforts by PCR followed by Sanger sequencing.

NB-LRR	Functional annotation	Ranking	Forward / Reverse
AT1G61190	RPP39	1	CAGAGCAGATATCGGGGCTG
			CGTCCTCCCATTCAAGCTCA
AT3G44630	RPP1-like	2	TCCGGAGTTTCTCGTCGAAC
			TGTCAAGAGCATGCGGGAAT
AT5G43470	RPP8	56	CCAGTTTCTCCTTCCACTCCC
			GGAGGGCTCATCCACTTGAG
AT5G43470	RPP8	57	GACATGGCATCGACCCTTCT
			ACATTGCTCAGGGTGTTGGA
AT4G11170	RMG1	111	TCAAGCTGGTGTTTGGACGA
			AATGTGGGACAATTTTAACATAAAACA
AT4G16860	RPP4	171	AACGAGGCCCTGAGGTAGAT
			TGGGAAAGACTCTCCACCTGA

Table 4.12: Candidate SNP primers for genotyping. Windows in Table 4.4 and the windows up to resolution 3 in Table 4.5 are included. Functional annotations for the NB-LRRs that these windows are found in are included. Forward and reverse primer sequences for the windows can be used for PCR.

The majority of our suggested SNP windows in Table 4.12 are annotated as *RPP (Resistance to Peronospora parasitica)* loci. The oomycete pathogen, *Peronospora parasitica*, later reclassified as *Hyaloperonospora parasitica*, is the causal agent of downy mildew.

For instance, AT5G43470 encodes *RPP8* in the Landsberg *erecta* ecotype [McDowell et al., 1998] and five closely related homologs in the Col-0 ecotype, found on five loci on two chromosomes [Initiative, 2000]. Sequence exchanges between *RPP8* homologs from the same locus occur more frequently than homologs from different loci, resulting in one type of disease resistance genes that are rapidly evolving [Kuang et al., 2008], perhaps contributing to the highly ranked windows in AT5G43470. Many potential mechanisms may be playing a part to generate the observed sequence diversity amongst the *RPP8* homologs and other NB-LRRs: point substitutions, segmental duplication, unequal crossing-over events, recombination, gene conversion and transposable elements [McDowell et al., 1998, Meyers et al., 2003, Kuang et al., 2008]. Additionally, the *RPP4* and *RPP1* gene encode TIR-NB-LRR proteins [van der Biezen et al., 2002, Botella et al., 1998], RPP39 is a CC-NB-LRR disease resistance protein that recognizes the oomycete's effector ATR39-1 [Goritschnig et al., 2012], and AT4G11170 encodes Resistance Methylated Gene 1 (RMG1), a TIR-NB-LRR gene that has been demonstrated to be a primary

target of RNA-directed DNA methylation [Yu et al., 2013].

The results from the analysis conducted in this chapter reveal some interesting observations when compared to the similar analysis carried out in *Spirodela polyrhiza* in Chapter 3. One difference between the two data sets was that the *A. thaliana* VCF files appeared to be entirely homozygous. This simplified the analyses, especially the indels-maxmin and bottleneck methods. Another observation is how representative the test sets of either *S. polyrhiza* or *A. thaliana* were to the full set of clones or accessions in their respective studies. Using available population analyses on the 1,135 *A. thaliana* accessions [Consortium, 2016], we curated a large, geographically diverse test set of *A. thaliana* accessions which formed the basis for our downstream analysis. The sequenced *S. polyrhiza* clones selected were a geographically diverse set from Europe, Asia and South America, but probably did not encompass the maximal geographic diversity seen in this globally distributed plant. It is likely that the *S. polyrhiza* clones that were selected for sequencing weren't an appropriately representative sample of the unsequenced clones that were tested or rather of the entire geographic diversity of this species, although they were chosen for their specific turion yield, which is a climatic adaptation and could be a measure of geographic diversity. Another potential issue concerns the genome assemblies. Problems with *S. polyrhiza* genome assembly and variant calling could also have had an effect on our results. Additionally, authors of the 1,135 *A. thaliana* genomes paper [Consortium, 2016] caution that their short-read-created pseudogenomes may still have issues with contiguity.

The bottleneck and max-min methods differed slightly from their application to the *S. polyrhiza* versus *A. thaliana* genomes. Mainly, this was due to the homozygosity observed in *A. thaliana* VCF files and *S. polyrhiza* heterozygous polymorphism calls. We had no haplotype phasing information for *S. polyrhiza*, so we had to factor in a degree of uncertainty with the VCF calls.

We capitalize on the volume of *A. thaliana* sequencing data by cherry-picking windows that demonstrated hyperpolymorphism in order to distinguish the full set of 1,135 accessions (Table 4.11), breaking with our previous methodology. However, it is reasonable to expect that we would want to use all the data available to us.

Because we didn't check if the primers for the windows from the test set contained sequence variations in those homologous regions in the 1,135 full set, it is possible that there would be deviation from our expected results if these primers were tested by PCR on accessions in wet lab experiments. These primers also assume amplification of single targets in the Col-0 genome, so actual number of PCR products may differ from the expected amplicon in Col-0 and in other accessions. Wet lab testing to confirm our predictions is underway.

Chapter 5

Conclusion

Duckweed is a promising sustainable feedstock that is currently used in a variety of applications. The large genetic diversity of duckweed plants, spanning 37 species, confers a wide range of physiological and biochemical properties and allows researchers to potentially find a duckweed plant that is well suited for a specific application. However, maintaining a genetically distinct population of duckweed, often in an open environment, necessitates a reliable method to identify and distinguish duckweed, preferably in a cost-effective and rapid manner. Previous efforts to genotype duckweeds at the species and sub-species level have been suboptimal. Selection of an inadequate genotyping method or marker can result in unsatisfactory taxonomic resolution. We improved existing assays and developed novel approaches to type duckweed plants.

In the first chapter of this thesis, a more complete set of barcodes was created by sequencing *atpF-atpH* and *psbK-psbI* regions for the six duckweed species that were not previously analyzed, and sequencing additional clones for underrepresented species. These new sequences and those from Wang et al. (2010) were then used to create a duckweed-specific database using the *atpF-atpH* and *psbK-psbI* barcodes. Thirty of the 37 duckweed species can be resolved with a BLAST-based protocol using the duckweed-only database. The remaining seven species were unable to be resolved by these plastidic barcodes, arising from previous misidentification errors or ambiguous assignments. This method was then used to identify the species of duckweed samples from our collection that were not previously typed, which demonstrated the usefulness of positive species identification of future untyped duckweed samples. These barcodes, however, are insufficient intraspecies markers, as demonstrated by *Spirodela polyrhiza* clones. Thus, a different approach is needed to genotype clones from the same duckweed species.

Because plastid markers have poor resolving power at the sub-species level, the nuclear genome was probed for suitable markers. The NB-LRR class of plant disease resistance protein-encoding genes was an attractive target, as it was previously demonstrated to have very high genetic diversity in other plants. We developed a pipeline that leverages sequence polymorphisms across multiple *S. polyrhiza* genomes to target hyperpolymorphic NB-LRR-related windows. These windows were ranked according to the number of distinguished pairs of clones. Twenty of the 23 *S. polyrhiza* clones tested were successfully identified through this method. Additionally, some of these markers can be used for interspecies distinction between the two closely related *Spirodela* species, *polyrhiza* and *intermedia*. The results from these first two chapters can facilitate the continued rise of duckweed-based research and applications.

Lastly, the pipeline to identify hyperpolymorphic NB-LRR-related windows was tested on *A. thaliana* accessions. A geographically diverse training set of 49 *A. thaliana* accessions was selected to approximate the diversity of the 1,135 genomes set. Hyperpolymorphic NB-LRR-related windows were identified from the test set of 49 accessions. The distinguishing power of these windows on the full set of accessions was demonstrated to be similarly powerful compared to the test set. Calculations from the Pearson correlation coefficient illustrated that the discriminatory power of the hyperpolymorphic windows from the SNP and INDEL ranking methods can be extrapolated from the test set to the full set of accessions. Then families of windows were produced to maximize distinguishing pairs of accessions on the full set of accessions in order to aid wet lab testing. These methods and results should provide researchers with a useful approach that can be extended to other organisms with sufficient genomic resources.

References

[Appenroth et al., 2013] Appenroth, K., Borisjuk, N., and Lam, E. (2013). Telling duckweed apart: Genotyping technologies for the Lemnaceae. *Chin. J. Appl. Environ. Biol.*, 19(1):1–10.

[Appenroth et al., 2015] Appenroth, K., Sree, K., Fakhoorian, T., and Lam, E. (2015). Resurgence of duckweed research and applications: report from the 3rd international duckweed conference. *Plant Mol. Biol.*, 89(6):647–654.

[Barnard, 2001] Barnard, R. (2001). Automotive engineering development. In Happian-Smith, J., editor, *An Introduction to Modern Vehicle Design*, pages 2–3. Reed Educational and Professional Publishing Ltd., Oxford, England.

[Bog et al., 2010] Bog, M., Baumbach, H., Schween, U., Hellwig, F., Landolt, E., and Appenroth, K. (2010). Genetic structure of the genus *Lemna* l. (Lemnaceae) as revealed by amplified fragment length polymorphism. *Planta*, 232:609–619.

[Bog et al., 2015] Bog, M., Lautenschlager, U., Landrock, M., Landolt, E., Fuchs, J., Sree, K., Oberprieler, C., and Appenroth, K. (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.

[Bog et al., 2013] Bog, M., Schneider, P., Hellwig, F., Sachse, S., Kochieva, E., Martyrosian, E., Landolt, E., and 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.

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

[Botella et al., 1998] Botella, M., Parker, J., Frost, L., Bittner-Eddy, P., Beynon, J., Daniels, M., Holub, E., and Jones, J. (1998). Three genes of the *Arabidopsis* RPP1 complex resistance locus recognize distinct *Peronospora parasitica* avirulence determinants. *Plant Cell*, 10(11):1847–1860.

[Browning and Browning, 2007] Browning, S. and Browning, B. (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.

[Chase et al., 2005] Chase, M., Salamin, N., Wilkinson, M., Dunwell, J., Kesanakurthi, R., Haidar, N., and Savolainen, V. (2005). Land plants and DNA barcodes: short-term

and long-term goals. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, 360:1889–1895.

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

[Consortium, 2016] Consortium, T. . G. (2016). 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. *Cell*, 166:481–491.

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

[Doyle and Doyle, 1987] Doyle, J. and Doyle, J. (1987). A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochemistry Bulletin*, 19:11–15.

[Eddy, 1998] Eddy, S. (1998). Profile hidden Markov models. *Bioinformatics*, 14:755–763.

[Fazekas et al., 2008] Fazekas, A., Burgess, K., Kesanakurti, P., Graham, S., Newmaster, S., Husband, B., Percy, D., Hajibabaei, M., and Barrett, S. (2008). Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. *PLoS ONE*, 3(e2802).

[Feng et al., 2017] Feng, B., Fang, Y., Xu, Z., Xiang, C., Zhou, C., Jiang, F., Wang, T., and Zhao, H. (2017). Development of a new marker system for identification of *Spirodela polyrhiza* and *Landoltia punctata*. *Int. J. Genomics*.

[Finn et al., 2016] Finn, R., Coggill, P., Eberhardt, R., Eddy, S., Mistry, J., Mitchell, A., Potter, S., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G., Tate, J., and Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.*, 44:D279-D285.

[Gan et al., 2011] Gan, X., Stegle, O., Behr, J., Steffen, J., Drewe, P., Hildebrand, K., Lyngsoe, R., Schultheiss, S., Osborne, E., Sreedharan, K., and *et al.* (2011). Multiple reference genomes and transcriptomes for *Arabidopsis thaliana*. *Nature*, 477:419–423.

[Goritschnig et al., 2012] Goritschnig, S., Krasileva, K. V., Dahlbeck, D., and Staskawicz, B. J. (2012). Computational prediction and molecular characterization of an oomycete effector and the cognate *Arabidopsis* resistance gene. *PLOS Genetics*, 8(2).

[Group, 2009] Group, C. P. W. (2009). A DNA barcode for land plants. *Proceedings of the National Academy of Sciences of the United States of America*, 106:12794–12797.

[Hammond-Kosack and Jones, 1997] Hammond-Kosack, K. and Jones, J. (1997). Plant disease resistance genes. *Annu. Rev. Plant Physiol. Plant Mol. Biol.*, 48:575–607.

[Hill et al., 2014] Hill, J., Demarest, B., Bisgrove, B., Su, Y., Smith, M., and Yost, H. (2014). Poly Peak Parser: Method and software for identification of unknown indels using Sanger sequencing of PCR products. *Dev. Dyn.*

[Initiative, 2000] Initiative, A. G. (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature*, 408:796–814.

[Jones et al., 2001] Jones, E., Oliphant, T., and Peterson, P. (2001). Scipy: Open source scientific tools for Python. `http://www.scipy.org/`.

[Klimyuk et al., 1993] Klimyuk, V., Carroll, B., Thomas, C., and Jones, J. (1993). Alkali treatment for rapid preparation of plant material for reliable PCR analysis. *The Plant Journal*, 3:493–494.

[Kluge and Farris, 1969] Kluge, A. and Farris, J. (1969). Quantitative phyletics and the evolution of anurans. *Systematic Zoology*, 18:1–32.

[Korf et al., 2003] Korf, I., Yandell, M., and Bedell, J. (2003). *BLAST: an essential guide to the basic local alignment search tool*. O'Reilly, Sebastopol, CA, USA.

[Kress et al., 2005] Kress, W., Wurdack, K., Zimmer, E., Weigt, L., and Janzen, D. (2005). Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the United States of America*, 102:8369–8374.

[Kuang et al., 2008] Kuang, H., Caldwell, K. S., Meyers, B. C., and Michelmore, R. W. (2008). Frequent sequence exchanges between homologs of RPP8 in *Arabidopsis* are not necessarily associated with genomic proximity. *Plant J.*, 54:69–80.

[Kuehdorf et al., 2014] Kuehdorf, K., Jetschke, G., Ballani, L., and Appenroth, K. (2014). The clonal dependence of turion formation in the duckweed *Spirodela polyrhiza*– an ecogeographical approach. *Physiol. Plant.*, 150(1):46–54.

[Lahaye et al., 2008a] Lahaye, R., Savolainen, V., Duthoit, S., Maurin, O., and van der Bank, M. (2008a). A test of *psbK-psbI* and *atpF-atpH* as potential plant DNA barcodes using the flora of the Kruger National Park (South Africa) as a model system. *Nature Precedings*.

[Lahaye et al., 2008b] Lahaye, R., van der Bank, M., Bogarin, D., Warner, J., Pupulin, F., Gigot, G., Maurin, O., Duthoit, S., Barraclough, T., and Savolainen, V. (2008b). DNA barcoding the floras of biodiversity hotspots. *Proceedings of the National Academy of Sciences of the United States of America*, 105:2923–2928.

[Lam et al., 2014] Lam, E., Appenroth, K., Michael, T., Mori, K., and 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.

[Landolt, 1986] Landolt, E. (1986). Biosystematic investigations in the family of duckweeds (lemnaceae). In *The family of Lemnaceae—a monographic study*, volume 1. Veröffentlichungen des Geobotanischen Instituts der ETH, Stiftung Rübel, Zürich, Switzerland.

[Les et al., 2002] Les, D., Crawford, D., Landolt, E., Gabel, J., and Kimball, R. (2002). Phylogeny and systematics of Lemnaceae, the duckweed family. *Syst. Bot.*, 27(2):221–240.

[Les et al., 1997] Les, D., Landolt, E., and Crawford, D. (1997). Systematic of the Lemnaceae (Duckweeds): inferences from mi- cromolecular and morphological data. *Plant Systematics and Evolution*, 204:161–177.

[Les and Sheridan, 1990] Les, D. and Sheridan, D. (1990). Biochemical heterophylly and flavonoid evolution in North American Potamogeton (Potamogetonaceae). *American Journal of Botany*, 77:453–465.

[McAloon et al., 2000] McAloon, A., Taylor, F., Yee, W., Ibsen, K., and Wooley, R. (2000). Determining the cost of producing ethanol from corn starch and lignocellulosic feedstocks. `http://www.nrel.gov/docs/fy01osti/28893.pdf`.

[McClure and Alston, 1966] McClure, J. and Alston, R. (1966). A chemotaxonomic study of Lemnaceae. *American Journal of Botany*, 53:840–860.

[McDowell et al., 1998] McDowell, J. M., Dhandaydham, M., Long, T. A., Aarts, M. G., Goff, S., Holub, E. B., and Dangl, J. L. (1998). Intragenic recombination and diversifying selection contribute to the evolution of downy mildew resistance at the RPP8 locus of *Arabidopsis*. *Plant Cell*, 10:1961–1874.

[Meyers et al., 2003] Meyers, B., Kozik, A., Griego, A., Kuang, H., and Michelmore, R. (2003). Genome-wide analysis of NBS-LRR-encoding genes in *Arabidopsis*. *The Plant Cell*, 15:809–834.

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

[Morello and Breviario, 2008] Morello, L. and Breviario, D. (2008). Plant spliceosomal introns: Not only cut and paste. *Curr. Genomics*, 9:227–238.

[Murray and Thompson, 1980] Murray, M. and Thompson, W. (1980). Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Research*, 8:4321–4325.

[Phuong and Schubert, 2017] Phuong, T. and Schubert, I. (2017). Reconstruction of chromosome rearrangements between the two most ancestral duckweed species *spirodela polyrhiza* and *s. intermedia*. *Chromosoma*, 126(6):729–739.

[Posada and Crandall, 1998] Posada, D. and Crandall, K. (1998). Modeltest: testing the model of DNA substitution. *Bioinformatics*, 14:817–818.

[Rismiller and Tyner, 2009] Rismiller, C. and Tyner, W. (2009). Cellulosic biofuels analysis: Economic analysis of alternative technologies. http://purl.umn.edu/53583.

[Ronquist and Huelsenbeck, 2003] Ronquist, F. and Huelsenbeck, J. (2003). MrBayes 3: bayesian phylogenetic inference under mixed models. *Bioinformatics*, 19:1572–1574.

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

[Sommer, 2005] Sommer, S. (2005). The importance of immune gene variability (MHC) in evolutionary ecology and conservation. *Front. Zool.*, 2.

[Sree and Appenroth, 2014] Sree, K. and Appenroth, K. (2014). Rediscovery of *Wolffia microscopica* (Griff.). *Kurz. ISCDRA Newsletter*, 3:2–4.

[Swofford, 2002] Swofford, D. (2002). *PAUP*: phylogenetic analysis using parsimony (* and other methods) Version 4*. Sinauer Associates, Sunderland, MA, USA.

[Tamura et al., 2011] Tamura, K., Peterson, D., Peterson, N., Stecher, G., M., N., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution*, 28:2731–2739.

[Tan et al., 2007] Tan, X., Meyers, B., Kozik, A., West, M., Morgante, M., St Clair, D., Bent, A., and Michelmore, R. (2007). Global expression analysis of nucleotide binding site-leucine rich repeat-encoding and related genes in *Arabidopsis*. *BMC Plant Biology*, 7:56.

[Thompson et al., 1997] Thompson, J., Gibson, T., Plewniak, F., Jeanmougin, F., and Higgins, D. (1997). The ClustalX Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research*, 25:4876–4882.

[Untergasser et al., 2012] Untergasser, A., Cutcutache, I., Koressaar, T., Ye, J., Faircloth, B., Remm, M., and Rozen, S. (2012). Primer3 - new capabilities and interfaces. *Nucleic Acids Res.*, 40(15).

[US Energy Information Administration, 2016a] US Energy Information Administration (2016a). Frequently asked questions: How much gasoline does the united states consume? april 4, 2016. `https://www.eia.gov/tools/faqs/faq.php?id=23&t=10`.

[US Energy Information Administration, 2016b] US Energy Information Administration (2016b). International energy statistics on renewables, biofuels production. `http://www.eia.gov/cfapps/ipdbproject/IEDIndex3.cfm?tid=79&pid=79&aid=1`. Table of Total Biofuels Consumption February 18, 2016.

[van der Biezen et al., 2002] van der Biezen, E., Freddie, C., Kahn, K., Parker, J., and Jones, J. (2002). *Arabidopsis* rpp4 is a member of the rpp5 multigene family of tir-nb-lrr genes and confers downy mildew resistance through multiple signalling components. *Plant J.*, 29(4):439–51.

[Vaughan and Baker, 1994] Vaughan, D. and Baker, R. (1994). Influence of nutrients on the development of gibbosity in fronds of the duckweed /textitLemna gibba. *J. Exp. Bot.*, 45:129–133.

[Wang et al., 2014] Wang, W., Haberer, G., Gundlach, H., Glaesser, C., Nussbaumer, T., Luo, M., Lomsadze, A., Borodovsky, M., Kerstetter, R., Shanklin, J., and *et al.* (2014). The *Spirodela polyrhiza* genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. *Nat. Commun.*, 5.

[Wang and Messing, 2012] Wang, W. and 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).

[Wang et al., 2010] Wang, W., Wu, Y., Ermakova, M., Kerstetter, R., and Messing, J. (2010). DNA barcoding of the *Lemnaceae*, a family of aquatic monocots. *BMC Plant Biol.*, 10(205).

[Xu et al., 2012] Xu, J., Cheng, J., and Stomp, A.-M. (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.

[Xu et al., 2011a] Xu, J., Cui, W., Cheng, J., and Stomp, A.-M. (2011a). Production of high-starch duckweed and its conversion to bioethanol. *Biosystems Engineering*, 110:67–72.

[Xu et al., 2011b] Xu, J., Cui, W., J.J., C., and Stomp, A.-M. (2011b). Production of highstarch duckweed and its conversion to bioethanol. *Biosystems Engineering*, 110:67–72.

[Yamaga et al., 2010] Yamaga, F., Washio, K., and Morikawa, M. (2010). Sustainable biodegradation of phenol by *Acinetobacter calcoaceticus* p23 isolated from the rhizosphere of duckweed *Lemna aoukikusa*. *Environmental Science and Technology*, 44:6470–6474.

[Yu et al., 2013] Yu, A., Lepere, G., Jay, F., Wang, J., Bapaume, L., Wang, Y., Abraham, A.-L., Penterman, J., Fischer, R., Voinnet, O., and Navarro, L. (2013). Dynamics and biological relevance of DNA demethylation in *Arabidopsis* antibacterial defense. *Proc. Natl. Acad. Sci. U.S.A.*, 110(6):2389–2394.

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