Project Number: MB2-CM12

Introgressive hybridization between *Orconectes quinebaugensis* and *Orconectes virilis* in the Blackstone River Valley

A Major Qualifying Project Report

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Bachelor of Science

in

Biology and Biotechnology

and

Biochemistry

by

_____

Alyce Buchenan

_____

Hiral Dutia

_____

Emily Scott-Solomon

Advisor:

_____

Michael Buckholt

Advisor:

_____

Lauren Mathews

April 26, 2012

Abstract

In the Blackstone River Valley watershed, morphological data indicate that the endemic crayfish species, *Orconectes quinebaugensis,* is hybridizing with the more widely dispersed *Orconectes virilis*. The goal of our project was to identify the level of hybridization and quantify introgression between the two species using the molecular technique amplified fragment length polymorphism, which generates species-specific markers. This study has generated insights into the evolutionary processes occurring in the watershed as well as their potential implications on biodiversity and conservation efforts.

Acknowledgements

The authors would like to thank our advisors Professors Michael Buckholt and Lauren Mathews for their continued assistance and guidance throughout this endeavor. We would also like to thank Abbie White for her assistance with ordering inventory, maintaining the lab facilities, and shipping our completed samples. Additionally, we would like to acknowledge the contributions of those teams that collected crayfish and began studies into these two species in the years preceding our project, as well as the team who continued these crayfish studies alongside us (Jessica McKniff and Kortni Violette), with whom we coordinated much of our laboratory work and data analysis. Finally, we would like to recognize the contributions of David Rollé for his assistance in using the analysis software.

Table of Contents

Background and Reason for Study

Understanding the mechanisms leading to speciation remains one of the central goals of ecology and evolutionary biology. Hybrid zones provide a unique opportunity to study the role of geographic isolation on the emergence and maintenance of species (Boecklen & Howard, 1997). Hybrids arise from the mating between individuals of separate species that have only recently diverged on the evolutionary time scale. Identifying and studying these incidents of hybridization has become easier within the last few decades with the development of novel molecular techniques such as amplified fragment length polymorphism (AFLP), which generates species-specific markers and can be used to analyze parentage and the degree of hybridization within a community. Since hybrids can sometimes outcompete with parent species for resources as well as limit or interfere with the mating of pure individuals of the parent species, analysis of hybrid genotypes can provide insight into species extinction that might otherwise go undetected (Mooney & Cleland, 2001; Perry *et al*., 2001a; Perry *et al*., 2001b; Perry *et al*., 2002). Human interaction with ecological communities has increased the incidence of hybridization as previously isolated species are brought into contact, through the introduction of a species to a new locale or through modifications of the local geography, potentially threatening the indigenous species (Vähä & Primmer, 2006). Thus, identifying and studying hybrid zones could have important consequences for conservation efforts. In the New England area, morphological data indicate that the endemic crayfish species, *Orconectes quinebaugensis*, is hybridizing with the more widely dispersed *Orconectes virilis*. The analysis of the species-specific markers identified by AFLP could provide insight into how the two species are hybridizing, and in what collection sites and populations this hybridization is most notable.

A. Hybridization and Introgression

Hybridization occurs when two individuals from different species (typically recently diverged on the evolutionary timescale) mate and produce a hybrid, an individual that is a mixture of the two gene pools. Introgression is the process by which genes typically associated with one species enter the genome of another related species (Streit *et al*., 1994). The concepts of hybridization and introgression are often studied together, but the two terms are not interchangeable. This study examines these two evolutionary concepts in the context of *O. virilis* and *O. quinebaugensis*. Because *O. virilis* may be invasive to this region, the occurrence of hybridization and introgression between these two species could have important ecological impacts, in that non-native species are not always balanced in a new environment in terms of predator-prey interactions and mating rituals. However, hybridization and introgression do not only occur when an invasive species is present, nor does the presence of an invasive species always result in the occurrence of these evolutionary mechanisms. Finally, while invasiveness, hybridization, and introgression can all impact the local habitat and have ecological consequences, this study is not intended to show exactly what these consequences are or whether the impacts are positive or negative, but merely to point out that researchers should be aware of these impacts when conducting research with field implications.

Hybridization and introgression can occur naturally or through anthropological interference. Often studied as a mechanism of speciation, hybridization can reduce native gene pools, especially in smaller, more susceptible populations (Perry *et al*., 2001a; Perry *et al*., 2001b). A crayfish species endemic to the Ohio River, northern Kentucky, and southeastern Indiana, *Orconectes rusticus*, was introduced to northern Wisconsin and Michigan in the 1960s. *O. rusticus* has since extirpated native crayfish such as *O. virilis and O. propinqu*s (Olsen *et al*.,

1991). Furthermore, Perry *et al*. (2001) showed that in the Wisconsin and Michigan areas, *O. rusticus* and *O. propinquus* can interbreed, resulting in fertile hybrids. These hybrids can then backcross with *O. rusticus*, thereby further displacing the indigenous species (Perry *et al*., 2001a; Perry *et al*., 2001b). Sometimes, when two distinct lineages breed, the hybrid offspring phenotype results in reproductive isolation. The *Heliconius* butterfly, *H. heurippa*, has an intermediate morphology generable by mating *H. melpomene*, *H. cydno*, and F1 hybrids. Reproductive isolation between the parent species is due to predator-mediated selection and assortative mating based on wing color patterns. Similarly in hybrids, the intermediate wing color pattern was found necessary to stimulate courtship between hybrids while acting as a deterrent to members of the parent species, thus isolating the hybrids from the parental species and giving rise to a third species. Furthermore, other *Heliconius* members have similar hybrid patterns, indicating hybrid speciation has occurred more than once in this area (Mavárez *et al*., 2006). Nijman *et al.* (2003) suggest that hybridization may have had a significant role in the formation of current domestic breeds. Thus, studying hybrid zones could provide valuable information for studying the gain and loss of species.

Crayfish present apt targets for studies of hybridization. Studies have shown that *O. virilis* has been an object of introgression by an invasive species as in Olsen *et al*. (1991), but also that it has acted in the capacity of an invasive species introgressing into a native species. For example, Ahern *et al*. (2008) documented the presence of *O. virilis* in a concrete pond in the United Kingdom, leftover from a personal collection of exotic animals. Genetic analysis suggested a dispersal rate of two kilometers per year, with introgression with native species occurring along the way. *O. virilis* has similarly been introduced into non-native parts of North and Central America, including US coastal watersheds and parts of Mexico (Hamr, 2002).

Although these cases show hybridization introgression occurring when an invasive species is present, the presence of an invasive species does not necessarily indicate the occurrence of these mechanisms.

1. Hybridization using Morphological Data, Mitochondrial DNA, and Nuclear DNA

While important for the initial identification of hybrids, phenotypic studies of morphological and behavioral data can only provide limited information on the genotype of the individuals; therefore, evaluating the parentage of hybrids or the backcrossing frequencies within populations is difficult with just these data. The development of different molecular techniques offers researchers tools to identify otherwise invisible variation between parent species to evaluate hybridization frequencies. These techniques often generate large numbers of loci, some of which can be used as markers for the parental species. The frequencies of these markers within the genome can identify hybrids, parentage, backcrossing frequency, and the level of introgression (Rhymer & Simberloff, 1996).

Nuclear and mitochondrial DNA provides different datasets that are crucial in generating comprehensive phylogenetic investigations. One reason nuclear DNA is required for hybridization studies is because uniparentally inherited organelles, such as mitochondria or chloroplasts, have DNA markers from a single parent (Meudt & Clarke, 2007). Mitochondrial DNA is inherited as a single linkage group; therefore, it only represents one independent marker for genetic comparison. Hybridization studies require the comparison of two or more independent genetic markers in order to identify membership coefficients and thereby determine to what extent hybridization is occurring (Moore, 1995). Furthermore, DNA from uniparentally inherited organelles is believed to undergo little to no recombination. Thus, one cannot assess interbreeding and hybridization in organisms as easily, due to the more limited genetic

information from the DNA of uniparental organelles which are inherited as a single locus, rather than offering multiple loci of study (Ahern *et al*., 2009; Bussell *et al*., 2005). Additionally, hybridization and phylogenetic studies based solely on mtDNA can be misleading due to introgression of inherited organelle DNA (Kai *et al*., 2002; Bachtrog *et al*., 2006; Coyne & Orr, 2004). For example, mtDNA cannot distinguish between morphotypes of a species as well as nuclear DNA (Bachtrog *et al*., 2006). In a study of three different morphotypes of the black rockfish, *Sebastes inermis*, genetic variation among the three could not be distinguished by mtDNA. However, AFLP analysis of the nuclear genome identified five to six diagnostic loci that distinguished the three morphotypes from one another (Kai *et al*., 2002).

Because mitochondrial DNA and nuclear DNA have different methods of inheritance and recombination, they can provide different phylogeographical inferences that can complement each other. These complementary inferences can be useful for understanding the geographic distribution of hybridization and introgression. When both nuclear and mtDNA are studied, a more complete analysis of the gene flow between divergent mtDNA lineages can be gathered (Ahern *et al*., 2009). Because mtDNA is assumed to be strictly maternally inherited, mitochondrial markers can be probed to study interspecies gene flow (Carr *et al*., 1986; Lamb & Avise, 1986). For example, studies of mtDNA can indicate mating biases in hybrid zones. In a study of tree frogs, hybridization occurred in shared breeding sites of two species, which generally exhibited habitat isolation (Lamb & Avise, 1986). Mitochondrial DNA data revealed that F1 hybrids shared the same mitochondrial genotype, indicating directional hybridization with male green tree frogs mating with female barking tree frogs (Lamb & Avise, 1986). Furthermore, parentage studies of hybrids can explain phenomena such as range displacement of closely related species. The range of white-tailed deer in New Mexico and Texas began to

expand in the mid-1900s, displacing mule deer over the proceeding decades. A study using mtDNA showed hybridization occurring between mule deer bucks and white-tailed deer does, with the mule deer gene pool absorbing the hybrids. Thus, hybridization may have contributed to the mule deer decline in contact zones (Carr *et al*., 1986). By quantifying mtDNA genotype occurrence within F1 hybrids, mtDNA distribution in backcrosses and later-generation hybrids can also be predicted (Lamb & Avise, 1986). However, it should be noted that because species that hybridize have similar genetic markers, in some cases data drawn from mtDNA and nuclear DNA may not lead to the same conclusions about differentiation and hybridization (Vallender *et al*., 2007).

2. Multi-Locus Genotyping and AFLPs

Molecular genetics, specifically multi-locus genotyping, allows for the study of hybridization with more accuracy (Vallender *et al*., 2007). For multi-locus genotyping, four types of genetic markers are most commonly used to detect polymorphisms in DNA. Microsatellites are short tandem repeat sequences of one to six base pairs frequently used in genetics studies. Random Amplified Polymorphic DNA (RAPD) is a simple and low-cost molecular technique in which random segments of DNA are amplified using PCR to study phylogenies of plant or animal species. Restriction Fragment Length Polymorphisms (RFLP) is a molecular technique in which restriction enzymes digest DNA and the resulting fragments are separated by size via gel electrophoresis. Amplified Fragment Length Polymorphism (AFLP) generates species-specific markers by restriction enzyme digestion of DNA and selective PCR amplification (Meudt & Clarke, 2007).

Microsatellites were first developed for human genetics studies, but became more popular among plant and animal researchers in the early-mid 1990s. Although microsatellites have

several advantages such as being highly polymorphic, highly variable, and behaving in a co-dominant fashion, disadvantages include costly and time-consuming isolation of markers (Bensch & Akesson, 2005). AFLP has a number of advantages over other molecular marker techniques. AFLP combines the techniques of RAPD and RFLP. Like RFLP, genomic DNA is digested with restriction enzymes and adaptors are ligated to sticky ends. Similar to RAPD, AFLP employs a two PCR step-protocol which selectively reduces the number of fragments amplified by 1/256 (Bensch & Akesson, 2005). Below, we describe both marker types in more detail, and compare the advantages and disadvantages of both.

While microsatellites provide important information for genetic studies, they have long start-up times and extremely high costs, which in most cases restricts the number of markers used in a study to fewer than fifty loci. In contrast, AFLPs have a short start-up time in most species, and thus many loci can be studied at low to moderate costs (Bensch & Akesson, 2005). AFLP is an ideal technique for studying the genomics of closely related species and non-model organisms because large numbers of fragments spanning the entire genome can be generated in a short amount of time, making the technique well-suited for hybridization studies. The fragments can show rare variations that would otherwise not be detected, resolving differences between closely related populations and species. Furthermore, AFLP requires no prior sequence information, so it can be used effectively to study non-model organisms (Vos *et al.*, 1995; Meudt & Clarke, 2007). Additionally, AFLP results have high reproducibility. AFLP is subject to one major drawback in that DNA quality can affect generated fingerprints (Vallender *et al*., 2007). Higher purity DNA needs to be used to minimize incomplete digestion, which can cause confounding partial fragments (Blears *et al*., 1998). However, the advantages of AFLP in hybridization studies far outweigh the need for higher purity DNA.

Currently, AFLPs are used for parentage analysis and assessment of individual genetic similarity, linkage mapping, investigation of population genetic structure and phylogenetic reconstructions, species phylogenies, genetic diversity of species or populations, and hybridization studies (Meudt & Clarke, 2007). AFLPs have mostly been used with studies of plants, fungi, and bacteria, primarily to determine genetics of economically important traits such as productivity and disease resistance, but recent research suggests they have much undiscovered potential in molecular ecology (Bensch & Akesson, 2005).

While AFLP is a valuable resource for genomics, a number of factors have to be considered when evaluating AFLP-generated data. AFLPs are a dominant marker system and can result in hundreds to thousands of genome-wide di-allelic loci; these loci are not informative individually, but can provide relevant information when studied in large groups (Bensch & Akesson, 2005). AFLP markers are treated as dominant markers; that is, polymorphisms are identified as either present (plus) or absent (null), designated as 1 and 0, respectively, and heterozygotes cannot typically be identified (Blears *et al*., 1998; Meudt & Clarke, 2007). This is a drawback of AFLPs when compared to co-dominant systems such as microsatellites, where more than ten alleles per locus can be observed, and heterozygotes can be identified (Bensch & Akesson, 2005). However, co-dominant systems such as microsatellites often have only a few loci identified for study, thus the vast amount of fragments generated by AFLP causes the technique to outperform other systems for identifying hybrids and differentiating between taxa and populations (Bensch & Akesson, 2005; Meudt & Clarke, 2007).

Although AFLPs are typically treated as dominant markers, they can display various degrees of co-dominance which have the potential to interfere with evaluating population parameters. AFLP markers generally represent non-coding regions due to the fact that these regions are

typically where non-detrimental mutations are able to accumulate without negatively impacting the fitness of the individual. As a result, AFLP fragments can contain repetitive sequences such as microsatellites (Meudt & Clarke, 2007; Wong *et al*., 2001). Because microsatellite insertions and deletions can occur independently of mutations in AFLP restriction sites, they often register as co-dominant (Wong *et al*., 2001). Damselflies have been used to evaluate the dominant nature of AFLP markers. Of the markers studied, fewer than 25% were evaluated as a strictly dominant marker with the majority registering various degrees of co-dominance. Of the eleven distinct AFLP fragments, two were found to contain repetitive sequences (Wong *et al*., 2001).

While statistical methods are currently available to aid in evaluating genetic population parameters that compensate for the dominant nature of AFLPs, it can still be difficult to separate dominance between homozygous and heterozygous genotypes without influencing the dataset with researcher biases (Bensch & Akesson, 2005; Darling *et al*., 2004). Some studies show that co-dominant markers can be extracted from AFLPs by identifying obvious differences in intensity of alleles (Meudt & Clarke, 2007; Wong *et al*., 2001). By assuming strong bands are homozygous (1/1) and weaker bands (if showing at approximately 50% the strength of homozygous bands) are heterozygous (1/0), heterozygosity can be determined from AFLPs. This method, however, relies heavily on researcher input, reducing the impartiality that is achieved by the use of computer programs for band-scoring (Bensch & Akesson, 2005).

In addition to the limitations imparted by their dominant nature, a number of other assumptions can limit the applications of AFLPs. One such assumption is that fragments that migrated the same distance on a gel are homologous. This assumption can be violated if the marker pool is subjected to high rates of homoplasy, where fragments representing different loci have the same size and are scored as being a shared character (Meudt & Clarke, 2007; Bensch &

Akesson, 2005). Another potential cause of a false reading can result from a substitution within a restriction site, triggering an absence of an allele at one locus but a presence at another locus (Meudt & Clarke, 2007). Additionally, size variation caused by a microsatellite within the AFLP fragment can cause two homologous bands to be scored as two different loci (Bensch & Akesson, 2005; Wong *et al.*, 2001). However, the sheer number of fragments generated by AFLPs can compensate for the disadvantages of its dominant nature, as the quantity of markers limits the impacts of a flawed assumption in any individual markers on the dataset as a whole (Bensch & Akesson, 2005; Crawford *et al*., 2011; Meudt & Clark, 2007). AFLP markers provide information for hybridization studies with minimal start-up costs and rapid results as compared to other competitive methods such as microsatellites, RFLPs, and RAPDs.

B. Crayfish Phylogeny and Characteristics

There are more than 540 species of crayfish in the world, belonging to the three families, *Astacidae*, *Cambaridae*, and *Parastacidae*, all of which are decapod crustaceans that form a monophyletic relationship with the marine lobsters (Porter *et al*., 2005). The family *Cambaridae*, to which the two crayfish species in this investigation belong, contains approximately twelve genera and 390 species of freshwater crayfish, making it the most diverse of the crayfish families (The Global Invasive Species Database, 2005). Roughly 99% of the species in the family *Cambaridae* are native to North and Central America (Hobbs, 1989). Of the cambarid genera, the genus *Orconectes* is the third most diverse. It contains ten sub-genera, eighty-one species, and thirteen subspecies, most of which are endemic to the eastern part of the continent (Hobbs, 1989; Integrated Taxonomic Information System, 2011).

Orconectes quinebaugensis is a recently described species that is a closely-related sister taxon to *O. virilis* (Mathews & Warren, 2008; Mathews *et al.*, 2008). In the Blackstone River

Valley, the morphology of collected specimens suggests the possibility that the two evolutionary lineages may be hybridizing (Mathews, personal communication). *O. quinebaugensis* was discovered in southern New England, and named for the Quinebaug River, its type locality, the location from which the "type" specimens were collected in order to first describe the species (Mathews & Warren, 2008). This river flows from eastern Connecticut through Central Massachusetts, with a Connecticut River Basin of 60.97 square miles (United States Geographical Society, 2006). *O. quinebaugensis* is typically found in small, rocky streams or shallow ponds with muddy substrates. It lives in conjunction with other crayfish species, notably *O. virilis* (Allard, 2010; Becker *et al*., 2009; Mathews & Warren, 2008; McMurrough & Saltzman, 2009).

Orconectes virilis, commonly known as the "virile crayfish", has a wide range throughout North America. The species has been identified in areas ranging from eastern Canada to the midwestern United States to as far south as Texas. Despite this broad range, debate remains as to whether *O. virilis* is native to all of these sites. The type locality of *O. virilis* is in Michigan, suggesting a natural range centered in northern central Unites States and southern Canada (The Global Invasive Species Database, 2005; Mathews, personal communication). Human intervention has spread its population to habitats as distant as Europe, which has been bolstered by its use as food and bait. As such, it is considered invasive in many regions (Ahern *et al*., 2008). The New England population is potentially invasive, though no empirical information exists pertaining to its introduction. For this reason, *O. virilis* could provide a key example of an invasive species that may be a subject of introgression into closely-related native species. As previously discussed, while invasiveness does not necessitate introgression and introgression is not only a product of invasion, when a species invades a new habitat, it can have significant

impacts on the current state of the habitat. Invasive species are not necessarily an appropriate member of the predator-prey interactions of the region, and can therefore impact a variety of populations throughout the food chain. Additionally, when an invasive species is closely-related genetically and morphologically, it may mate with native species, resulting in hybridization and introgression, which can have similarly wide-spread impacts (Smith & Smith, 2005). In their northern habitats, such as New England, *O. virilis* populations migrate into deeper waters in the winter months in order to survive the harsher conditions, thereby depending on permanent bodies of water that are deep enough to not freeze during the winter. Additionally, they require shelter from predators, using rocks, logs, or vegetation for this purpose (Mathews *et al*., 2008).

With common habitat preferences and similar morphology, *O. virilis* and *O. quinebaugensis* present suitable subjects for a study of introgression between closely-related evolutionary species. While they are considered two separate species, *O. virilis* and *O. quinebaugensis* may not have undergone complete reproductive isolation. This condition could be the result of two different scenarios. First, *O. virilis* is not native to the region and was introduced by humans, which means that it did not evolve in the same location and therefore did not require reproductive barriers in order to undergo speciation from *O. quinebaugensis*. Second, both species evolved in the same area from a common ancestor and many years ago were separated by physical barriers; thereby, speciation occurred without the need for reproductive isolation (Coyne & Orr, 2004). Reproductive isolation is the product of "barriers" against mating between individuals of separate species or subspecies; these can occur in the form of different mating seasons, practices, or pheromones; physical separation by distances; or an inability to mate due to morphology. This last option is of particular note when studying crayfish, as it has been hypothesized that reproductive isolation may occur in crayfish as a results of mechanical

isolation, or the inability to physically mate due to differences in reproductive morphology (Mathews, personal communication; Becker *et al*., 2009; McMurrough & Saltzman, 2009; Smith & Smith, 2005). On the other hand, if such isolating factors are not present between the species, interbreeding would be possible and could result in hybridization.

The current research is a continuation of a series of projects conducted by Becker *et al*. (2009), McMurrough & Saltzman (2009), and Allard (2010) with the purpose of determining the relationship of *O. virilis* and *O. quinebaugensis* with regards to population genetics, introgression, and hybridization. Becker *et al*. (2009) utilized genetic and behavioral techniques to investigate reproductive isolation between the species, considering specimens at only two sites: site M2 in the Quinebaug River in Sturbridge, MA, and site R1 in the Blackstone Gorge in North Smithfield (Figure 1). Site M2 is the type locality for *O. quinebaugensis*, and previous genetic and morphological data indicate that this population shows little or no sign of hybridization with *O. virilis* (unpublished data). Also based on previous morphological and genetic data, site R1 in Rhode Island site represents a local population of *O. virilis* that shows little or no sign of hybridization with *O. quinebaugensis* (unpublished data). In a study conducted by McMurrough & Saltzman (2009), a total of 505 individuals, identified by morphology as members of the sister clade formed by *O. virilis* and *O. quinebaugensis*, were collected from the M2 and R1 sites as well as twenty-three additional sites in Massachusetts, as shown in Figure 1. These individuals were compared using the mitochondrial gene cytochrome oxidase I (COI) as well as the microsatellite-containing locus Ov54. Within this sampling, the mitochondrial COI gene had three haplotypes: an Ov haplotype group with two haplotypes differing by a single nucleotide polymorphism (SNP), and a third Oq haplotype characteristic of individuals identified as *O. quinebaugensis*. The Ov and Oq mtDNA haplotypes varied by about

fifteen SNPs. The microsatellite locus showed two alleles in the total collection of crayfish, with all three genotypes (denoted AA, BB, and the heterozygous AB) represented.

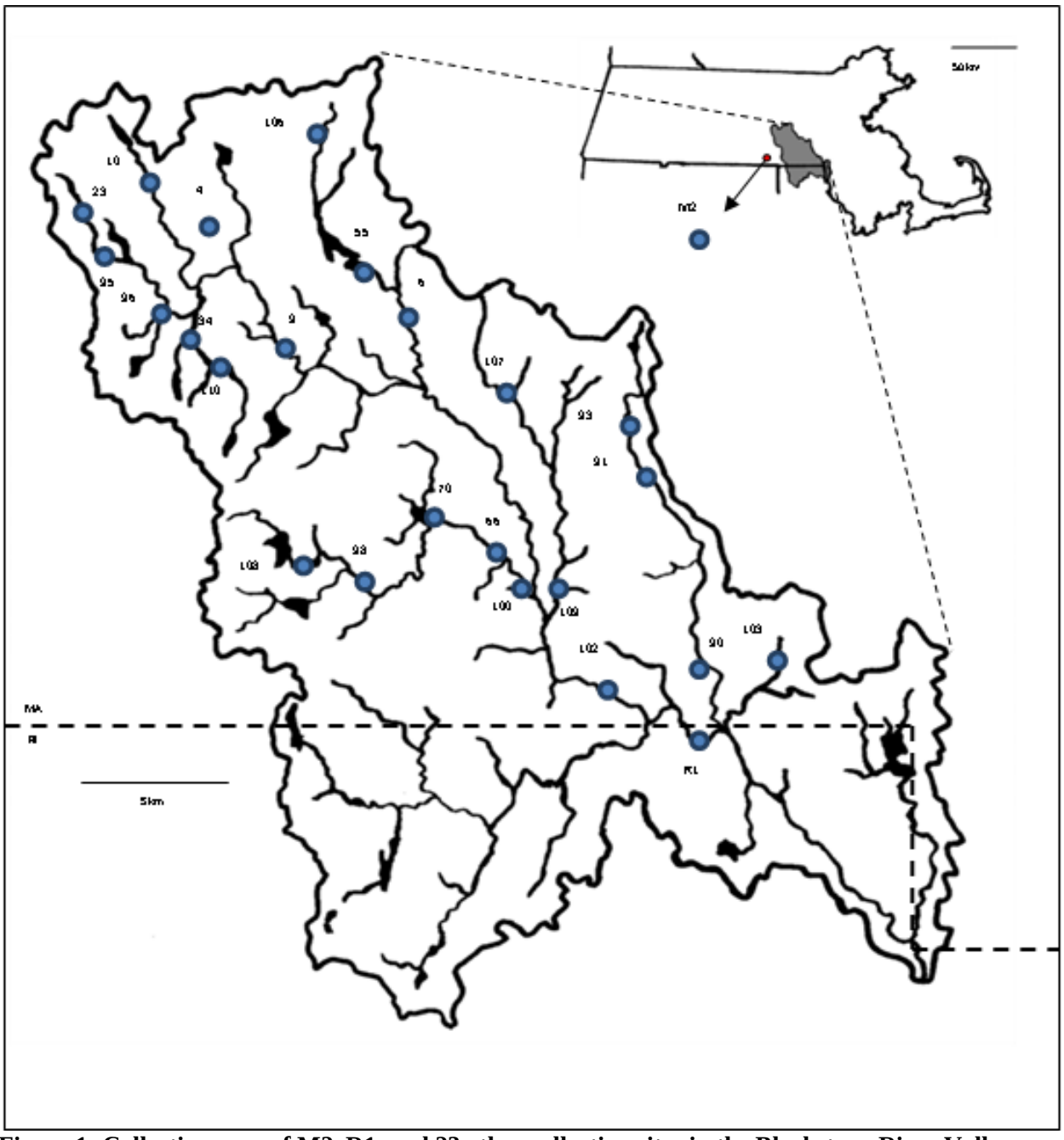

Figure 1: Collection map of M2, R1, and 23 other collection sites in the Blackstone River Valley watershed.

Of the eighty-eight specimens collected for Becker *et al*. (2009) from R1, all exhibited the Ov mtDNA haplotype, with 83% exhibiting microsatellite allele A. Of the eighty-seven

specimen from M2, 85% exhibited an Oq mtDNA haplotype, with 93% exhibiting microsatellite allele B. Thus, each population showed strong linkage disequilibrium between the mtDNA haplotype and microsatellite genotype, indicating that mismatches between the two loci may represent hybrids. This correlation provided an important tool for subsequent studies of introgression (Becker *et al*., 2009; McMurrough & Saltzman, 2009). Figure 2 shows the high degree of assortment of the A microsatellite allele with the mtDNA haplotype Ov and the B microsatellite allele with the mtDNA haplotype Oq. This data analysis highlights the linkage disequilibrium in the twenty-three study sites, when compared with the "pure" sites of M2 and R1, which suggests that hybridization is occurring more frequently in these sites.

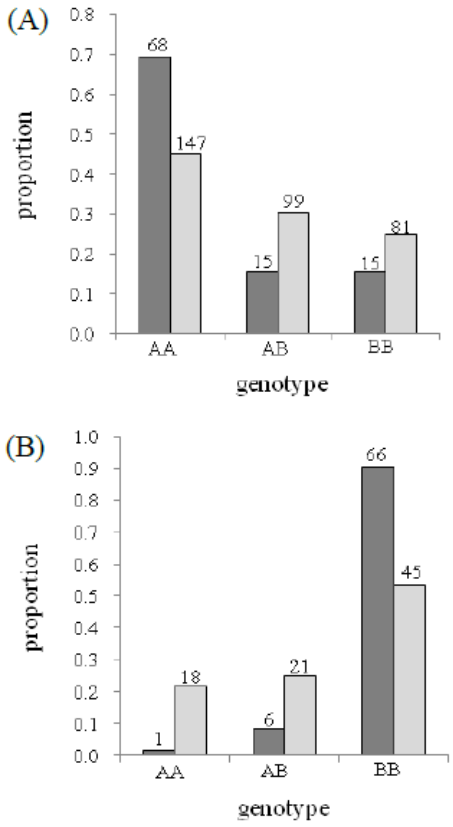

Figure 2: Graph representation of the proportion of individuals. A) Haplotypes associated with *O. virilis* **B)** Haplotypes associated with *O. quinebaugensis*. Dark bars represent data from "pure" populations M2 and R1, whereas light bars represent data from the other twenty-three collection sites in question. Numbers above the bars indicate sample sizes.

These findings were further supported by the discovery of intermediate morphology in sites of suspected hybridization. There was a statistically significant correlation between the top projection pleopod angle and the different COI haplotypes and the Ov54 microsatellite genotypes, such that genotype A was significantly associated with a large pleopod angle. Additionally, the top projection pleopod angle and Ov54 genotype A were found to be significant ($p= 0.0501$ and $p=0.054$, respectively) predictors of COI haplotype, such that they were associated with COI haplotype Ov (McMurrough & Saltzman, 2009). Figure 3 shows the manner in which pleopod measurements were taken in each study.

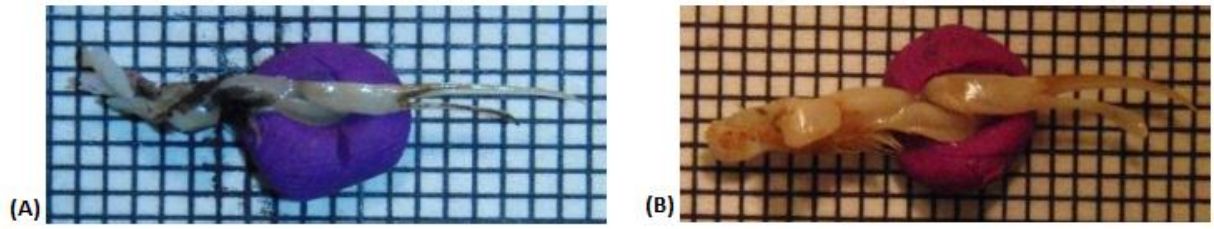

Figure 3: Pleopods of (A) O. virilis and (B) O. quinebaugensis, attached to grid paper for measurements (McMurrough & Saltzman, 2009).

Furthermore, Becker *et al*. (2009) conducted mating trials in order to explore reproductive capabilities and isolation between the two pure populations. Form I, or sexually mature males, were paired with females for a period of three hours, during which time observers recorded the number of mating attempts as well as the number of times a mating position was successfully attained. Twenty-four trials were conducted for each mating type (M2-M2, R1-R1, M2-R1, and R1-M2). Figure 4 graphically depicts the results of these mating trials in terms of attempts and successes, as determined by Becker *et al*. (2009). Intrapopulation mating was found to be significantly ($p<0.001$) more successful than interpopulation mating, suggesting that these populations were subject to partial reproductive isolation.

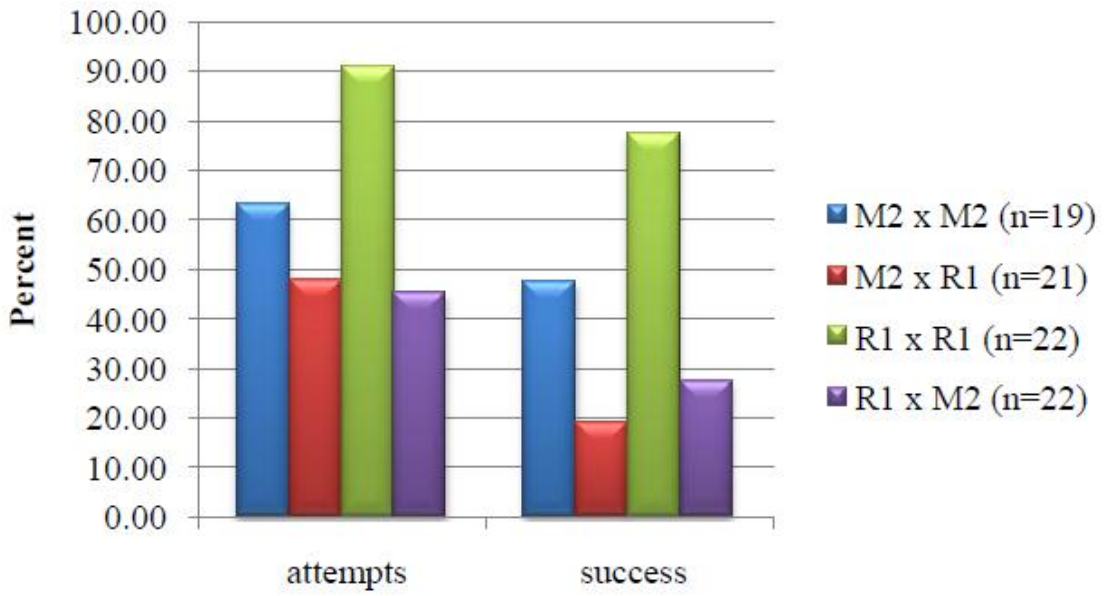

Figure 4: Results of mating trials by site pairing. N represents the number of trials in each treatment group. Intrapopulation pairings were significantly more successful ($p<0.001$) than interpopulation pairings (Becker *et al*., 2009).

In a related study conducted the following year, Allard (2010) searched for further genetic differences between *O. virilis* and *O. quinebaugensis* using AFLPs, and focusing on crayfish from three localities. The goal of Allard's (2010) study was to lay the groundwork for using AFLP as an analysis tool for the hybridization of *O. virilis* with *O. quinebaugensis* by first establishing R1 as a local site of most genetically pure *O. virilis* specimens.

The same collection of crayfish were utilized in this investigation as described in McMurrough & Saltzman (2009) and Becker *et al*. (2009) to complete the preliminary AFLP analysis begun by Allard (2010). These data, in combination with the microsatellites and mitochondrial DNA, provides a more complete genetic sampling for analysis than any of the previous studies could alone. In both McMurrough & Saltzman (2009) and Becker *et al*. (2009), the focus was on microsatellite and mitochondrial DNA data. As discussed in the previous sections, studies focused solely on mtDNA or utilizing microsatellites can provide misleading

readings due to few available loci for study. For this reason, it was necessary to use AFLPs in order to increase the number of independent loci available for analysis. Increasing the number of loci for study greatly increases the accuracy of the results (Vähä & Primmer, 2006; Boecklen & Howard, 1997). Allard (2010) laid the groundwork for the current investigation by developing a successful AFLP protocol. By utilizing AFLP as a technique to generate large numbers of informative markers, and the full available sample set of 467 genomic DNA samples from the Blackstone River Valley, this study will have access to a large number of species-specific markers that will provide insights into the number of gene pools present in sites of interest as well as indications as to the likelihood of membership of individuals to a particular population.

As previous data suggested that hybridization is occurring between *O. virilis* and *O. quinebaugensis*, the goal of this study was to identify AFLP markers that could be used to assess the degree of hybridization and introgression that might be occurring in sites where both *O. virilis* and *O. quinebaugensis* are present. In order to do this, the alleles of two "pure" genomes must be compared for differences in present and absent alleles in order to identify markers that may be species-specific. We used specimens from the site M2 in Sturbridge, MA, to represent a "pure" sample of *O. quinebaugensis*. Unfortunately, because the type locality for *O. virilis* is in Michigan, we were unable to obtain a large sample of individuals of *O. virilis* from the "native" range of that species. Instead, we used specimens from the site R1, identified by previous genetic and morphological data as a likely "pure" population of *O. virilis*, as well as a smaller sample of four individuals collected from a site in Michigan. Finally, we used our set of species-specific markers identified by comparison of specimens from those two sites to examine the occurrence and geographic distribution of hybridized and introgressed individuals among the other twenty-three sites in the Blackstone River Valley.

Methodology

A. Sample Collection & DNA extraction

The 467 genomic DNA samples used in this project are a subset of the 505 individuals identified morphologically as members of the *O. virilis* group that were collected from sites in the Blackstone River Valley watershed by McMurrough & Saltzman (2009). Researchers collected individuals from twenty-three sites located in Massachusetts and one in Rhode Island via hand collection, dip nets, seines, and crayfish traps baited with raw chicken or salmon. DNA from a detached leg of each collected specimen was extracted according to the Solid Tissue Protocol in Gentra's Puregene: Genomic DNA Purification Kit, with complete methods described in McMurrough & Saltzman (2009) and Becker *et al*. (2009). After genomic DNA at a concentration of ~10 ng/µL was obtained for each individual, all individuals were sacrificed and stored at -80˚C (McMurrough & Saltzman, 2009).

B. AFLP Protocol

Crayfish genomic DNA was analyzed using an AFLP protocol. The digestion-ligation reaction was conducted by adding 10 µL of mastermix to 2.5 µL of 10ng/µL genomic DNA. The mastermix (per reaction) consisted of:

- 1.25 µL 10x T4 ligase buffer
- 1 µL 0.5 M NaCl
- 0.5 µL 1 mg/mL BSA
- 0.5 µL 50 mM *Mse* I adaptor
- 0.5 µL 5 mM *Eco* RI adaptor
- 0.1 µL 10,000 U/mL *Mse* I enzyme
- 0.15 µL 20,000 U/mL *Eco* RI enzyme
- 0.05 µL 400,000 U/mL T4 DNA ligase
- 4.95 µL de-ionized water for a total volume of 10 µL

Samples were incubated at 37˚C for 2 hours. The samples were then held up to overnight at 4˚C until they could be transferred to the freezer or used in the following reaction. Each

digestion-ligation reaction was diluted 1:10 with deionized water in preparation for pre-selective PCRs. The pre-selective PCR (PS-PCR) was conducted by adding 12.5 µL of mastermix to 2.5 µL diluted digestion-ligation reaction. The mastermix (per reaction) consisted of:

- 1.5 µL 10x Thermopol I buffer
- 0.75 µL 2.5 mM dNTPs
- 0.4 µL 10 mM *Mse* primer (A or C)
- 0.4 µL 10 mM *Eco* primer (A or C)
- 0.075 µL 5,000 U/mL *Taq* polymerase
- 9.4 µL deionized water for a total volume of 12.5 µL

Four sets of primer combinations were used: (*Mse*-A, *Eco*-A), (*Mse*-A, *Eco*-C), (*Mse*-C, *Eco*-A), and (*Mse*-C, *Eco*-C), resulting in four sets of PS-PCRs labeled A, B, C, and D, respectively. The PCR reaction consisted of the following steps: 95˚C for 2 minutes; then 40 cycles of 95˚C for 30 seconds, 48˚C for 30 seconds, and 72˚C for 1 minute, followed by a final 72˚C extension for 10 minutes. The samples were then held up to overnight at 4˚C until they could be transferred to the freezer or used in the next reaction. Each pre-selective PCR reaction was diluted 1:10 with deionized water in preparation for selective PCRs. The selective PCR was conducted by adding 10 µL of mastermix to 2.5 µL of diluted pre-selective PCR reaction. The mastermix (per reaction) consisted of:

- 1.5 µL 10x Thermopol I buffer
- 0.75 µL 2.5 mM dNTPs
- 0.6 µL 10 mM *Mse* primer (ATC or CTC)
- 0.2 µL 10 mM 6FAM label *Eco*-ACG primer
- 0.2 µL 10 mM VIC label *Eco*-CAG primer
- 0.2 µL 10 mM NED label *Eco*-CTC primer
- 0.075 µL 5,000 U/mL *Taq* polymerase
- 9.0 µL deionized water for a total volume of 12.5 µL

Two sets of primer combinations were used: (*Mse*-ATC, *Eco*-ACG, *Eco*-CAG, *Eco*-CTC) and (*Mse*-CTC, *Eco*-ACG, *Eco*-CAG, *Eco*-CTC) labeled 1 and 2 respectively. We carried out

multiplexed selective PCR reactions by adding pre-selective PCRs A and B to mastermix with primer combination 1, and pre-selective PCRs C and D to mastermix with primer combination 2. The PCR reaction was run on the same program used for the pre-selective PCR. The samples were held up to overnight at 4˚C until they could be transferred to the freezer or prepared for analysis on an automated sequencer. To prepare for shipment to an automated fragment analysis facility, 1.5 µL of selective PCR reaction diluted 1:10 with deionized water was added to 9.5 µL HiDi formamide and 0.4 µl LIZ 600 standard. The biochemistry of AFLPs is explained in greater detail in Appendix A. All sequence information for adaptors and primers is listed in Appendix B.

C. Determining Failure Rates and Analyzing Allele Data in GeneMapper

After completing the first AFLP run of all 467 available genomic samples, we used GeneMapper software to determine whether these samples produced results. First, we loaded all samples into the program and assessed whether the reactions were successful. Next, we changed the protocol to "AFLP", so as to ensure that the program would be utilizing the same general protocol in determining failure rates as would be used in the future analysis of the data. We changed the panel to "none", as panels are used in order to record size and dye information of the determined markers and the purpose of this analysis run was not to generate any information, but merely to determine that there were peaks present in the samples and that the size standard was of an acceptable quality to later be used to determine peaks. We also changed the size standard to "LIZ600," which contains thirty-six single-stranded markers, in size fragments ranging from 20-600 bp and scored via a proprietary fluorophore label (Applied Biosystems, 2005). An ideal peak profile for this size standard can be seen on the referenced Applied Biosystems webpage (Applied Biosystems, 2005). A preliminary analysis run under this simple AFLP protocol

highlighted the sample quality and sizing quality. If GeneMapper was unable to score all of the bands present in the size standard, we attempted to score the size standard manually. However, in all such cases, we were unable to resolve size standard errors with manual scoring. The samples for which there were no allele peaks were re-run through the AFLP protocol detailed above and sent to the fragment analysis facility for a second time.

For all samples that yielded few or no scored alleles in the first attempt, we repeated the entire AFLP procedure starting from genomic DNA. Having completed two full trials in order to maximize the available samples with successful AFLP data, we then created a folder of all available samples. Those files for samples that failed were deleted, and replaced with their retrials, and failed retrials were also deleted. Additionally, if any samples only yielded acceptable data for either primer set A or primer set B, both were removed in order to eliminate partial AFLP profiles. Once we had uploaded and carried out preliminary analyses of all samples that had yielded good data for both primer sets, we began peak analyses. Six runs were completed: A and B samples were run separately three times, once for each dye (6FAM, NED, and VIC). By running six separate analyses, the program was able to delete common alleles, thus minimizing extraneous information and limiting the dataset to those markers that were most informative of differences within and between populations. In addition to deleting common alleles, each analysis program was set to analyze with the range of 50 to 500 bp, as this was the range in which the size standard was available to determine the relative florescence units (rfus) that are used to mark the presence or absence of a loci based on the fluorescent dyes attached to the amplified DNA fragments. The analysis method was also programed to produce its own panel. Panels and bins are denoted for each loci in order to mark the presence based on fragment size and dye color, thereby permitting comparison of individuals based on these many loci.

However, without known genomic information, a panel must be determined via an algorithm instead of inputted based on known information about the species. In a number of samples, the size standard worked, but was cut off early. Reviewing the results for those samples that did have the complete size standard revealed that few, if any, alleles were present after 500 bp. In order to maintain consistency as well as maximize efficiency and the number of samples available for analysis, the analysis range was set to detect alleles in the 50 to 500 bp range (as opposed to re-running the cut-off samples with minimal genomic DNA available). Within this range, peaks under 200 relative fluorescence units (rfu) were dropped and those between 200-300 rfu were marked "check". This was done to limit the dataset to informative markers, as it was noted that below 200 rfu there was a certain basal level detection from the dyes that was not necessarily indicative of actual presence of a marker. Those peaks over 300 rfu were passed as significant for further study. The programming of this analysis method is shown for 6FAM in Figure 5. The resultant genotypes table was exported as a color-coded Microsoft® Excel file.

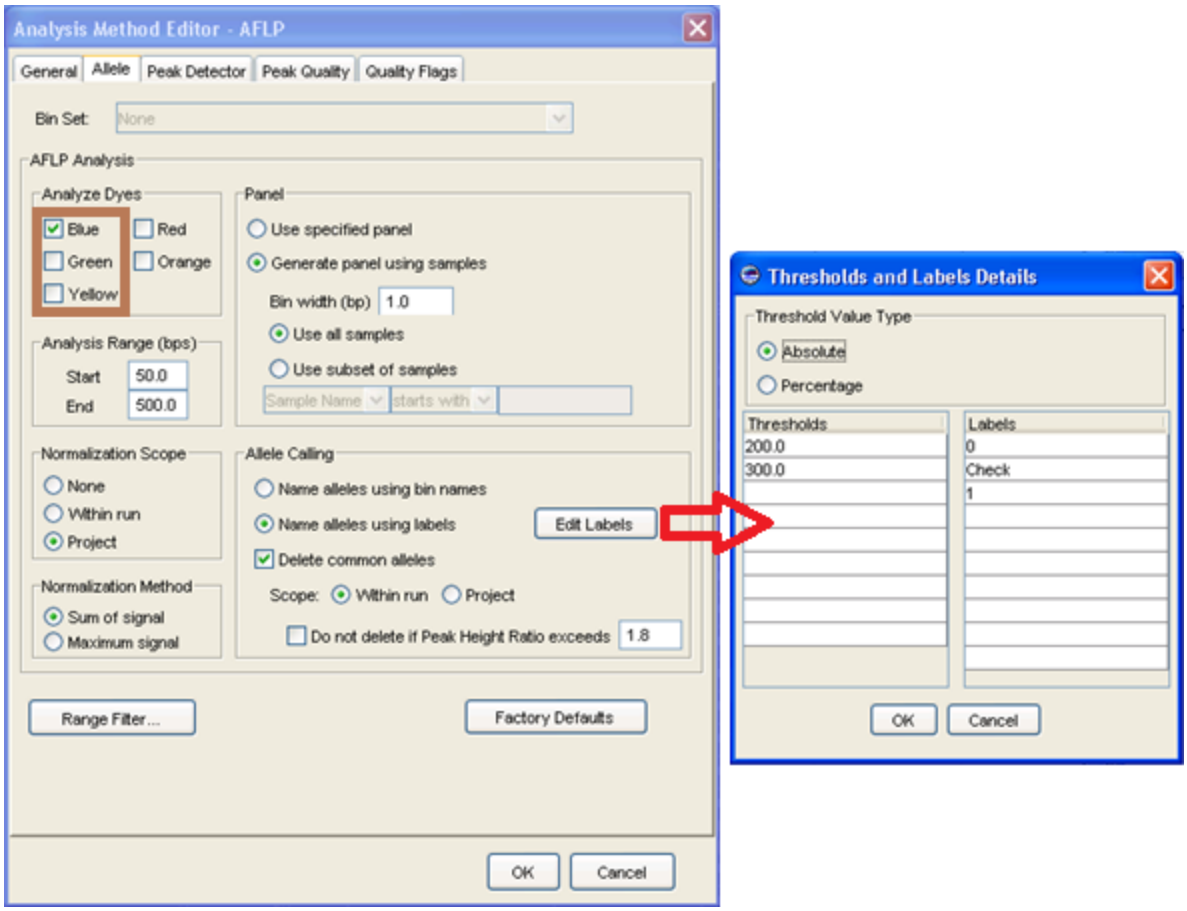

Figure 5: Programming details in GeneMapper shown for 6FAM. The red arrow shows how the thresholds were recorded, and the brown box highlights where each dye was noted.

In the outputted Excel files, we first narrowed the dataset based on the assumption that loci found present or absent in over 95% of the population were non-diagnostic and therefore uninformative. In the remaining loci, checks were re-evaluated and scored based on the following criteria: if more than 60% of the population had either presence or absence, checks were scored to conform with the majority; if there was a roughly even split between alleles, checks were scored based on peak height. Peaks greater than 250 rfus were scored present and those less than 250 rfus were scored absent.

D. Chi-squared Analysis to Determine Significance of Markers

Lineage-specific alleles were identified by comparing data from M2 and R1. Presumably because the species have only recent diverged evolutionary, there were no uniform markers (100% presence in one population with a 0% presence in the other). Markers did, however, have high variability in frequency between populations. To determine which markers showed significant associations with one or the other lineage, a Chi-squared test for independence was performed ($P < 0.05$) in Microsoft Excel. Markers were labeled as diagnostic for a given species if P had a value less than 0.05 and were assigned to the species with the higher frequency. Eighty markers had a significant association with one of the two populations over the other, and were labeled as species-specific markers.

E. Building Population Models in STRUCTURE and Determining Membership Coefficients in CLUMPP and DISTRUCT

After the dataset was narrowed to the relevant eighty species-specific loci, the Excel file was saved as a "comma-separated values" (.csv) file to be loaded into STRUCTURE version 2.3.3 (Pritchard *et al.*, 2010). A total of 420 samples and eighty loci were used in the analysis. STRUCTURE, which uses a Bayesian analysis algorithm, works best when used to analyze a co-dominant system. However, it can be adequately modified to work with AFLP when a large dataset is used. Presence markers are scored as 1/"missing data" (denoted as 1/-9), in order to avoid assumptions as to their nature as heterozygotes or homozygotes. The value of -9 was chosen because it did not appear elsewhere in the dataset. In STRUCTURE, population models were run with assumed K values of 1 through 25 populations. Each K value was run a total of twelve times, such that the repetitions provided an indication of statistical likelihood from the provided LnP(D) values. Likelihood was determined according to the following formulas:

$$K(K) = \ an\ average\ of\ 12\ values\ of\ LnP(D)$$

$$L\text{^}'(K) = L(K) \quad \langle\!\langle L(K) \rangle\!\rangle\_n - \quad \langle\!\langle L(K) \rangle\!\rangle\_(n-1)$$

$$L\text{^}''(K) = \quad \langle\!\langle L'(K) \rangle\!\rangle\_n - \quad \langle\!\langle L'(K) \rangle\!\rangle\_(n-1)$$

$$\Delta K = \quad (L\text{^}''(K)) \,/\, Stdev$$

M2 and R1 were evaluated separately from all other collection sites, using all 226 loci, to determine if they represented "pure" populations of the parental species. Ten independent runs of STRUCTURE were performed for K=1 through K=5. K=1 through K=3 had the highest likelihood values. A Wilcoxon signed-rank test was used to produce likelihood solutions as described by Rosenberg *et al.* (2001) in R software. The Wilcoxon signed-rank test is a paired difference test; it is used to compare repeated measurements on a single sample or matched samples. Each dataset was comprised of ten STRUCTURE runs per K value; likelihoods were assigned a rank based on the ordering of the values in each dataset. The following steps are used to determine the test statistic (Petruccelli, Nandram, & Chen, 1999):

1. Subtract the median under the null hypothesis (H_0) from each observation to center the data: $Y'_i=Y_i-\Theta$, i=1…n
2. Use the absolute values of the centered observations to compute the ranks. For example, $R_i=\text{rank}(|Y'_i|)$, i=1…n; if $|Y'_n|$ is the largest of all $|Y'_i|$, then $R_i=n$.
3. The test statistic is defined as *W*: $W = \sum_{(i:Y_i'>0)} R_i$

A p-value of 0.05 was used to determine significance. Samples were matched based on run number (run 1 for K=1 was matched with run 1 for K=2, for example). From this analysis, a K=2 was determined to have the optimal likelihood, supporting the assumption that two gene pools were present in M2 and R1.

The programs CLUMPP and DISTRUCT were used to analyze the raw data from STRUCTURE. Clustering programs, such as STRUCTURE, give a matrix in which each individual is assigned a membership coefficient (Q) for each cluster. Q represents the probability that an individual or a fraction of its genome has membership in a cluster, which can be used to infer ancestry of individuals (Jakobsson & Rosenberg, 2007a). Together, all these individual Q values form a Q matrix, which is used by CLUMPP and DISTRUCT to infer ancestry of populations. These programs can use a stochastic simulation when making inferences about population structure from multi-locus data, which can result in different outcomes for independent runs regardless of the same initial starting conditions. The two most common discrepancies that occur among multiple runs are "label switching" and "genuine multimodality" (Jakobsson & Rosenberg, 2007a). Markov chain Monte Carlo (MCMC) methods are used in Bayesian analysis to draw inferences from complex statistical models. MCMC output has the intrinsic problem of "label switching" in which replicated runs obtain the same membership coefficients, but in a different arrangement (Jasra, Holmes, & Stephens, 2005). "Genuine multimodality," a more difficult hurdle for analysis than "label switching," occurs when replicated cluster analyses arrive at distinct solutions that are not equivalent, which result from algorithms providing multiple equally viable solutions (Jakobsson & Rosenberg, 2007a).

CLUMPP is a permutation program that can be used to find the optimal alignment of replicated cluster analyses of the same data. It is used to account for "genuine multimodality" and "label switching." The program has three algorithms a user can choose from: FullSearch, Greedy, and LargeKGreedy. The output is a mean of Q of the permuted matrices across replicated datasets (Jakobsson & Rosenberg, 2007a). Analysis was performed on twenty-two runs of K=3 and K=7 using individual membership coefficients. Greedy was used without

weighting by the number of individuals, so populations would be considered equally regardless of the number of individuals that were sampled. One thousand random input orders were used for analysis (Jakobsson & Rosenberg, 2007b). The output from CLUMPP can be used in the graphical program DISTRUCT. A supplementary manual is available in Appendix C in addition to the one available from the authors (Jakobsson & Rosenberg, 2007b).

DISTRUCT produces a graphical representation of STRUCTURE output with each individual represented as a line segment. The line is portioned into colored components based on membership coefficients. The program requires not only the individual membership coefficients, but also the putative population membership coefficients (Rosenberg, 2007). Q for a putative population is the mean of the membership coefficients of the individuals comprising the population. For example, when K=3, there will be three Q values per putative population. A supplementary manual is available in Appendix D in addition to the one available from the author (Rosenberg, 2007).

Results

Six AFLP primer combinations generated 226 loci among the 420 individuals sampled. Using a Chi-squared test on R1 and M2 samples, eighty loci were identified as species-specific, in that variation in those loci was distributed significantly differently from the null expectation of 50:50 between the two populations. However, it is worth noting that as these two species have recently diverged from a common ancestor, when comparing the two lineages, loci can be fixed for the presence of an allele, fixed for the absence of an allele, or variable in both lineages as a result of incomplete lineage sorting. Thus, species-specific loci are not necessarily fixed and can remain variable; the variability can be the result of an inherited characteristic or a novel mutation. The other 141 loci were not distributed significantly differently from null expectations between the two populations, thus not assorting in such a manner as to denote species specificity. Interestingly, eight loci were not present in the "pure" populations represented in M2 and R1, but were present in a statistically significant number of individuals when all twenty-four populations were compared by Chi-squared analysis (largest $p < 0.04$). Locus YB12 was of particular interest ($p < 2.25*10^{-43}$). This locus was present in the majority of the M91, M93, and M95 samples (75.0%, 84.6%, and 81.0% respectively), as well as in a few individuals of the M96, M103, and M107 (<15% for each) populations. A number of potential explanations exist for this phenomenon. A polymorphic insertion or deletion could be present in only a small number of populations, either as a novel mutation or inherited from a common ancestor. Additionally, a third unidentified species could be hybridizing with either of the two targets of this study.

The population structure of M2 and R1 was explored using STRUCTURE to verify that the two populations represented "pure" populations of *O. quinebaugensis* and *O. virilis*, respectively. Using the protocol described by Pritchard *et al.* (2000) for STRUCTURE, all 226 loci were used

to detect the optimal K value for M2 and R1. A total of ten independent runs for each of K=1 to K=5 were performed. The clustering solutions with the highest likelihoods were K values of 1, 2, and 3.

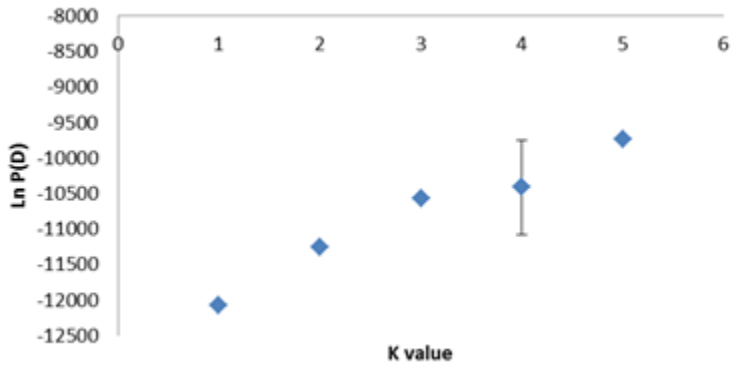

Figure 6: Graphical likelihood of M2 and R1 K values. LnP(D) represents the likelihood value for each K value. K=1 through K=3 are the average of ten values, and K=4 through K=5 are the average of five values. Greater variance at K=4 and the leveling off of likelihood after K=3 indicated the optimal K value was 1, 2, or 3. Standard deviation is shown for all data values; however, all except K=4 are too small to be seen at this scale.

A Wilcoxon two-sample rank test was performed as described by Rosenberg *et al.* (2001). Figure 6 shows the graphical representation of the likelihood of M2 and R1 K values. The graph begins to level off after the optimal K value; however, there may still be small increases in the LnP(D) values after the optimal K value. In addition, there are generally large variances in likelihoods, as can be seen in the value of K=4, after the optimal K. These values were compared in a pair-wise fashion with the following results. Both K=2 vs. K=1 and K=3 vs. K=1 had significant differences in likelihoods (p=0.001953 and p=0.005889, respectively). When K=3 was compared to K=2, there was also a significant difference in likelihoods (p=0.0001806). Because the likelihood of K=2 when compared to K=1 (p=0.001953) was less similar than the likelihood of K=3 when compared to K=1 (p=0.005889), K=2 was the most likely cluster value. This coincides with the two putative populations of M2 and R1. Another four runs were performed in STRUCTURE for K=2 for use in further analysis. CLUMPP and DISTRUCT were

used to visualize the population coefficients (Q) from fourteen STRUCTURE runs at K=2 (Figure 7). M2 and R1 were grouped into distinct clusters; however, a few individuals in both populations appeared to have mixed ancestry (Figure 8). The STRUCTURE analysis indicates that M2 and R1 represent mostly pure populations of *O. quinebaugensis* and *O. virilis*, respectively (Figure 7).

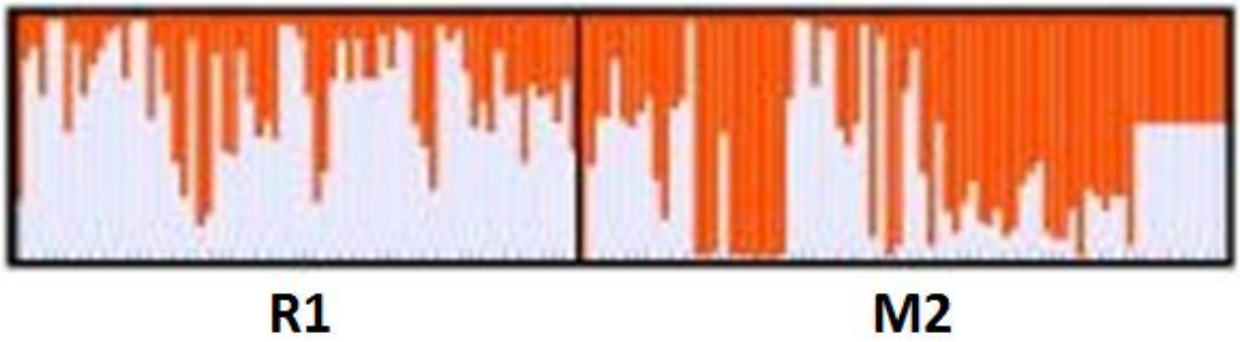

Figure 7: M2 and R1 population structure as determined by all 80 loci. Fourteen STRUCTURE runs were performed on 146 individuals putatively representing *O. virilis* (R1) and *O. quinebaugensis* (M2) to determine whether the sites represented pure populations. DISTRUCT output from the STRUCTURE analysis was used to visualize the Q matrix from when K=2. Gray represents cluster 1 and orange represents cluster 2. The populations are separated by a black line.

Figure 8 shows a more quantitative analysis of the ancestry of individuals in populations M2 and R1: the number of individuals with a putative *O. virilis* ancestry of more than 50% is higher in R1 than in M2. M2 shows more individuals with less than 50% putative *O. virilis* ancestry. A graphical representative of the putative *O. quinebaugensis* ancestry would show the reciprocal of this information. Additionally, Tables 1 and 2 show the percent putative *O. virilis* and *O. quinebaugensis* ancestry for collection site. For each site, the number of individuals in a particular bin are listed for percentages 0-20%, 20-40%, 40-60%, 60-80%, and 80-100% likely membership to *O. virilis* or *O. quinebaugensis* ancestry. This information is later shown visually in DISTRUCT output in Figure 9A.

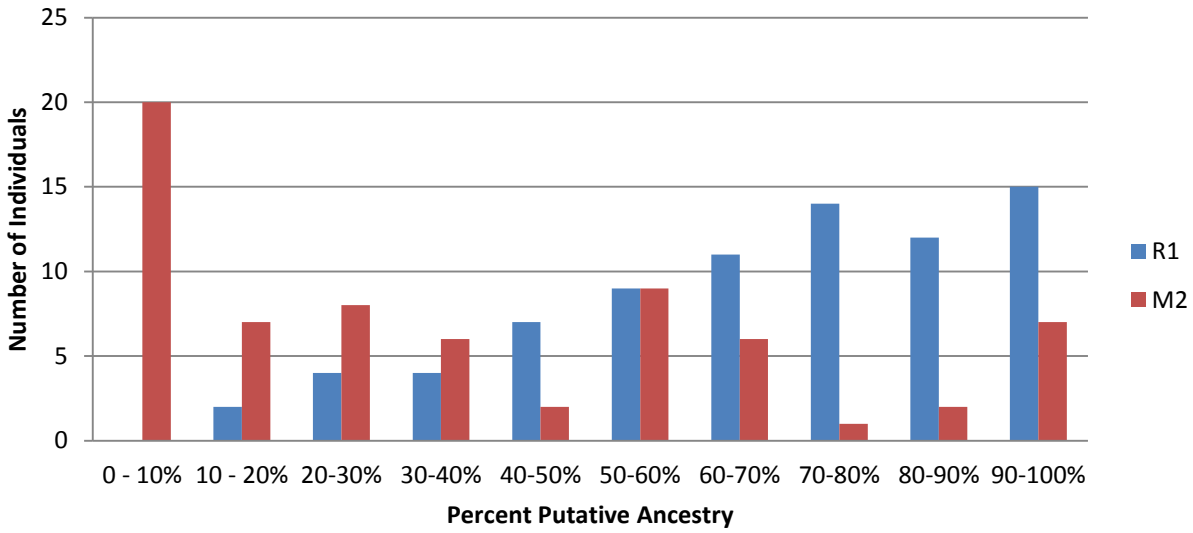

Figure 8: Putative *O. virilis* ancestry in pure populations of M2 and R1. More individuals in R1 population cluster in the >50% bins, whereas more individuals in M2 cluster in the <50% bins.

Table 1: Putative O. virilis ancestry by collection site for K=3. Individuals are binned within each percentage of belonging to the O. virilis ancestry.

Population	0-20%	20-40%	40-60%	60-80%	80%-100%	Total Samples
R1	8	6	4	14	46	78
M2	38	10	4	4	12	68
M4	3	0	0	0	0	3
M6	0	1	4	3	9	17
M9	0	0	0	3	9	12
M10	1	0	0	1	7	9
M23	2	0	1	1	0	4
M34	10	0	0	0	0	10
M55	12	0	0	0	0	12
M66	0	0	1	3	10	14
M70	11	0	0	0	0	11
M90	0	0	1	3	5	9
M91	12	0	1	0	3	16
M93	13	0	0	0	0	13
M95	21	0	0	0	0	21
M96	11	0	0	0	0	11
M98	11	0	2	1	1	15
M102	14	0	0	0	0	14
M103	3	1	2	1	0	7
M106	17	0	0	0	1	18
M107	9	5	0	1	1	16
M108	7	3	0	0	0	10
M109	11	3	0	0	1	15
M110	5	3	1	2	6	17

Table 2: Putative *O. quinebaugensis* ancestry by collection site for K=3. Individuals are binned within each percentage of belonging to the *O. quinebaugensis* ancestry.

Population	0-20%	20-40%	40-60%	60-80%	80%-100%	Total Samples
R1	59	5	2	5	7	78
M2	14	3	6	14	31	68
M4	0	0	0	0	3	3
M6	12	2	2	1	0	17
M9	9	3	0	0	0	12
M10	8	0	0	0	1	9
M23	0	1	1	0	2	4
M34	0	0	0	0	10	10
M55	0	0	0	0	12	12
M66	12	1	1	0	0	14
M70	0	0	0	0	11	11
M90	5	3	1	0	0	9
M91	10	1	4	1	0	16
M93	8	2	3	0	0	13
M95	21	0	0	0	0	21
M96	10	1	0	0	0	11
M98	4	0	0	0	11	15
M102	12	1	0	1	0	14
M103	6	0	1	0	0	7
M106	15	1	0	1	1	18
M107	14	1	1	0	0	16
M108	10	0	0	0	0	10
M109	11	0	1	2	1	15
M110	8	2	0	3	4	17

Since M2 and R1 represented distinct populations of the two parent species, the eighty "diagnostic" markers could be used for ancestral analysis of putative hybrid populations. Twelve independent runs of STRUCTURE with K values ranging from K=1 to K=25 were performed. We assumed that the true number of populations was unlikely to exceed the total number of sampling sites, twenty-four. The optimal K value was determined using the *ad hoc* method by calculating the second order rate of change of the likelihood (ΔK) (Evanno *et al*., 2005). Two K values had optimal likelihoods, K=3 and K=7. An additional ten simulations were run for each of these K values. In simulation studies of STRUCTURE with mixed ancestry populations carried out by Pritchard *et al.* (2000), a similar event occurred where the authors also obtained two optimal K values. Since there are two putative evolutionary lineages in our dataset (*O.*

quinebaugensis and *O. virilis*), K=3 seems like the more biologically realistic case; in addition, the *ad hoc* guide gave a larger value for K=3 than K=7. We suggest that the occurrence of the second K value is due to the difference in grand-parentage for each individual from the two original species (Pritchard *et al.*, 2000). As before, CLUMPP and DISTRUCT were used to visualize the population coefficients (Figure 9). When K=3, M2 and R1 were assigned to different clusters, as would be expected if those two sites represent pure (or mostly pure) populations of *O. virilis* and *O. quinebaugensis*, respectively (Figure 9A). STRUCTURE could not completely resolve membership for individuals in a number of putative hybrid populations. For example, STRUCTURE assigned a 60-70% probability of belonging to cluster one to three individuals collected from M9. Similarly, of the nine individuals from M90, three had a 60-70% probability of belonging to cluster one; furthermore, one individual had a 50% probability of belonging to cluster one. This could indicate hybridization is occurring at these sites. In addition, assuming an F_2 hybrid generation would have a 75% probability of belonging to one of the two clusters, individuals with a membership assignment of 60-70% could indicate backcrossing.

Interestingly, when the population coefficients from K=7 were used to assign individuals to clusters, M4, M34, M55, and M70 clustered separately from *O. quinebaugensis*, even though they had clustered with the parental species when K=3 (Figure 9B). Additionally, M91, M93, M95, M96, M102, M103, M107, M107, M108, and M109 did not cluster with either parental species for either K value, though some of them still showed potential ancestry from one of the parental species (Figure 9). When population structure was analyzed using a value of K=7, these populations formed two distinct clusters (Figure 9B). It should be noted M91, M93, and M95, all of which had locus YB12 in a large number of their population, were clustered together. The

clustering of these populations separately from parental populations could have occurred for two reasons. First, speciation could be occurring with preferential mating with other hybrids. The second hypothesis is that a third, unidentified species, is hybridizing with *O. virilis* and *O. quinebaugensis*.

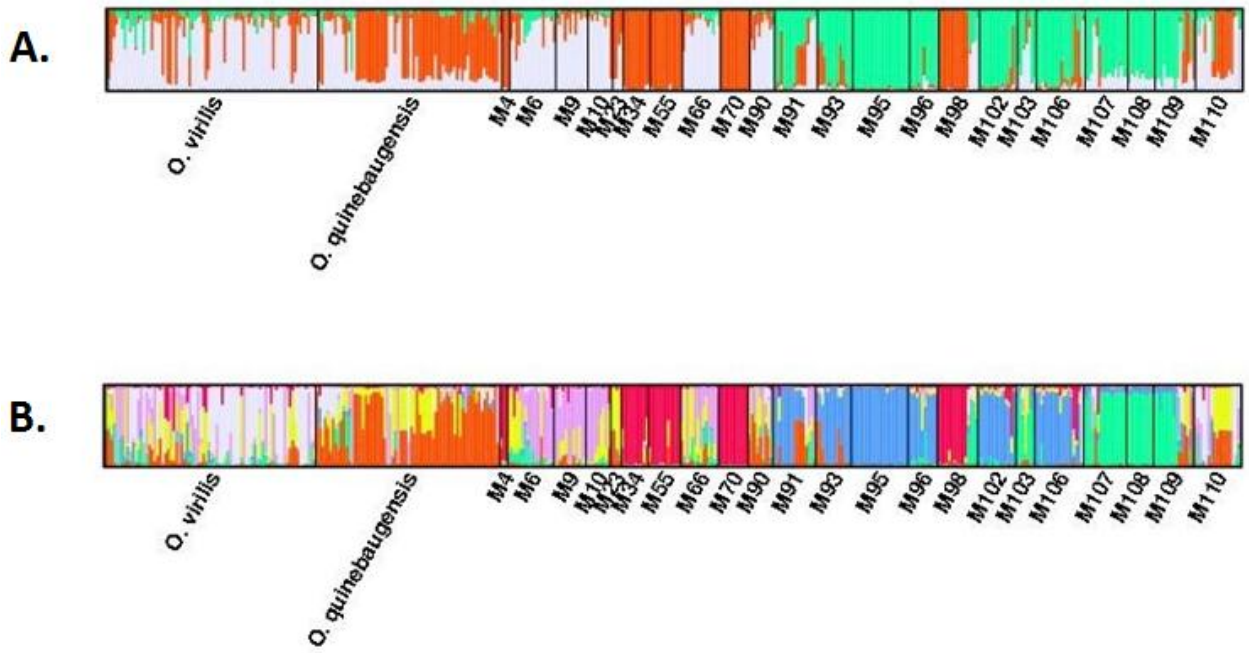

Figure 9: STRUCTURE analysis of 420 individuals from 24 collection sites. Populations are separated by black lines. Each vertical line represents an individual. **A)** DISTRUCT output from STRUCTURE analysis where K=3. Gray is cluster 1, orange is cluster 2, and green is cluster 3. **B)** DISTRUCT output from STRUCTURE analysis where K=7. Orange represents cluster 1, green represents cluster 2, blue represents cluster 3, yellow represents cluster 4, dark pink represents cluster 5, gray represents cluster 6, and red represents cluster 7. Twenty-two independent runs were performed for each K value.

Populations which clustered together in STRUCTURE analysis (Figure 9) did not necessarily cluster geographically (Figure 10). For example, a few sites such as M107, M108, and M109, which clustered together, were geographically distant from one another. M4, M34, M55, M70, and M98, which also were assigned the same cluster in the K=7 analyses, were also geographically distant. However, when compared with the mitochondrial data from McMurrough & Saltzman (2009) and Becker *et al*. (2009), all sites except M34 have an *O. virilis*

mitochondrial haplotype. About 33.3% of the M34 population has the *O. virilis* haplotype, and 13.3% of M2 has the *O. virilis* haplotype (Figure 11). This could indicate either that hybrids are present in that population or that there are genetically pure individuals of *O. virilis* in the supposedly "pure" population of *O. quinebaugensis*. Future studies in this area would examine the individual genotype and mitochondrial haplotype (McMurrough & Saltzman, 2009) data in order to address this issue. For the hybrid populations that did not strongly associate with either of the parental species (when K=3), there does not appear to be a bias toward a certain haplotype. For example, M91, M93, M95, and M96 were clustered; however, the majority of M91 and M93 have an *O. quinebaguensis* haplotype while M95 and M96 have an *O. virilis* haplotype.

Previous studies found that while there was only one haplotype for *O. quinebaugensis*, there were two haplotypes for *O. virilis* (Ov A and Ov B); these haplotypes differed by only one single nucleotide polymorphism (Mathews, personal communication). The genotype map (Figure 10) was compared with a map of the Ov haplotypes (Figure 12) to determine if a specific haplotype correlated with populations that had a high probability of ancestry from the third unidentified genotype. Sites M103 and M108 were almost 100% Ov A haplotype while only a minority of M107 and M109 were Ov A (Figures 9B and Figure 12). M102 and M96 had almost 100% Ov B haplotype; in addition, M91, M95, and M106 had a predominately Ov B haplotype, but M93 had almost 100% Ov A haplotype (Figures 9B and Figure 12). Comparison of the frequency of Ov haplotypes with populations having the unidentified gene pool did not appear to have a definitive correlation. However, analysis at an individual level may reveal that one of the Ov haplotypes originated in a third species.

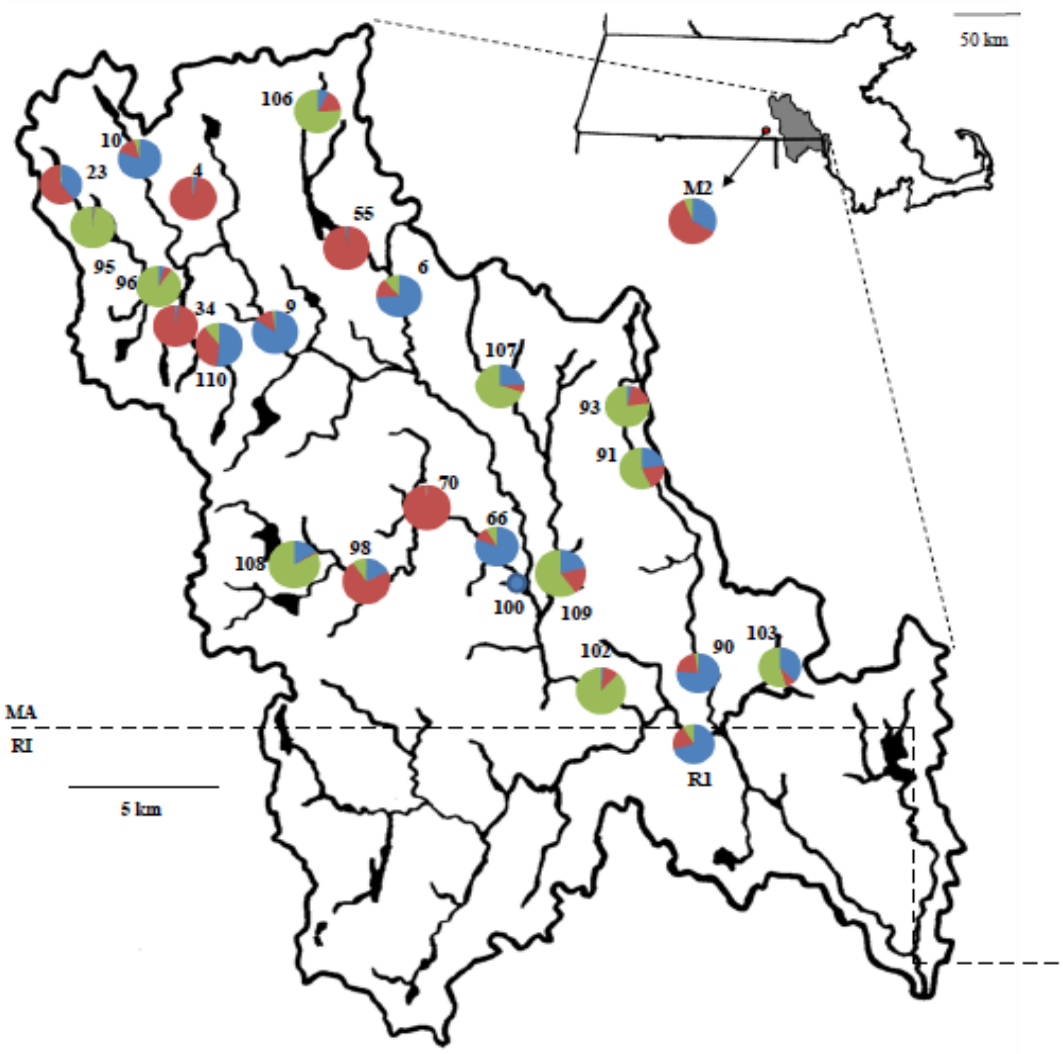

Figure 10: Population ancestry geographic distribution. The population membership coefficients (Q) were determined from the individual values of Q determined by STRUCTURE. This represents the putative ancestry of the population. Blue represents the cluster that was most common at R1, the site that contains a population primarily considered to be *O. virilis*. Red represents the cluster that was most common at M2, the site that contains a population considered primarily to be *O. quinebaugensis*, and green represents cluster 3, whose relationship with the other two lineages is unknown. Pie charts are located over each collection site.

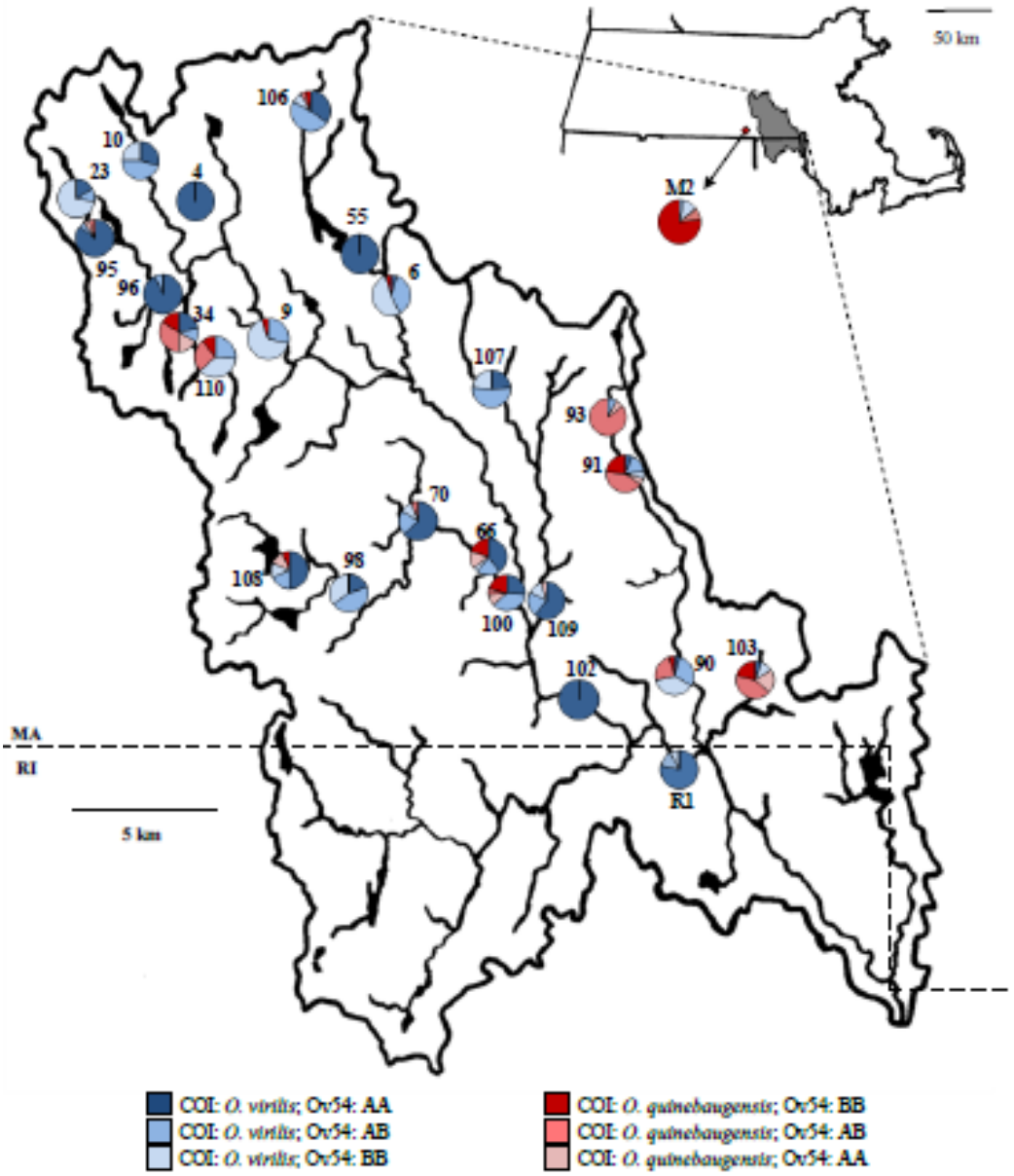

Figure 11: Mitochondrial haplotype & microsatellite genotype distribution. Distribution of crayfish Ov and Oq haplotypes and Orco54 genotype in the Blackstone River Valley (unpublished data).

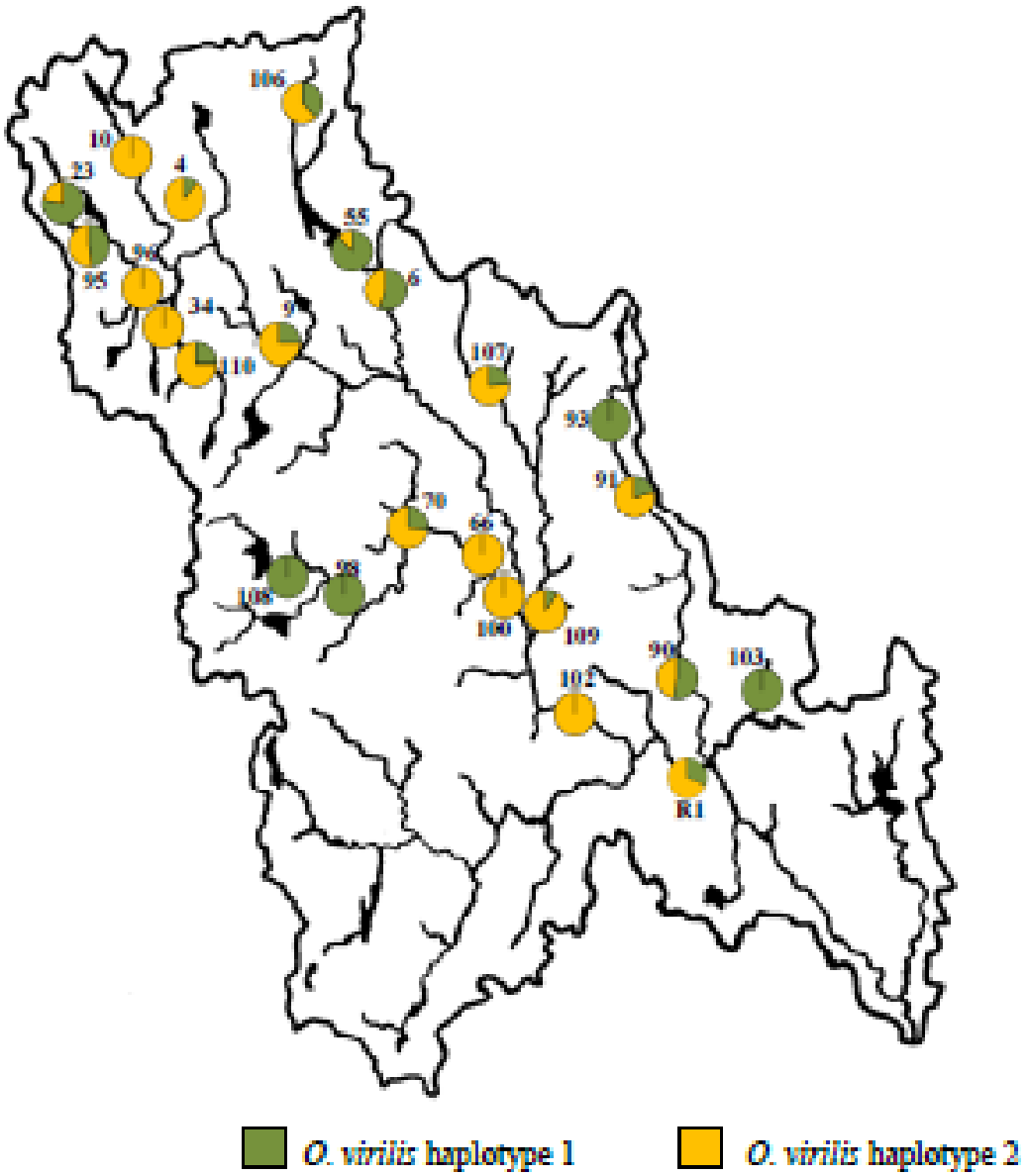

Figure 12: Ov haplotype geographic distribution. Distribution of Ov haplotypes 1 (A) and 2 (B) in the Blackstone River Valley (unpublished data).

Discussion

Diagnostic markers were generated using AFLP and identified by statistical analysis. The eighty identified markers were used to analyze the ancestry of twenty-two putative admixture populations to evaluate the level of hybridization occurring between *O. quinebaugensis* and *O. virilis.* The results of this study supported the hypothesis that hybridization is occurring between *O. quinebaugensis* and *O. virilis*. In addition, the AFLP data revealed the presence of a third gene pool whose relationship to *O. quinebaugensis* and *O. virilis* remains unknown.

The sites M2 and R1 were used as pure samples of *O. quinebaugensis* and *O. virilis*, respectively. STRUCTURE analysis with 226 loci successfully separated the populations into two clusters, indicating the presence of only two gene pools. M2 was predominately assigned to cluster 2 while R1 was predominately assigned to cluster 1. However, some individuals appeared to have mixed ancestry in both populations. One explanation for this could be that hybridization is occurring in these populations since both species are present. Another explanation pertains to the ancestry of *O. virilis* and *O. quinebaugensis*. The two crayfish in this study are incompletely separated lineages; therefore, the markers used in this study were not all completely unique to one species. Markers were selected based on which had a significantly greater frequency in one population (M2 or R1) than the other. The common ancestor for these two species most likely had variability in the occurrence of some of these markers, which may or may not have undergone complete sorting by genetic drift since the separation of the lineages. Thus, extant populations may still show some variation in these markers that was retained from the ancestral condition, and would not reflect hybridization. Comparison of individuals with apparent "mixed" genotypes with their haplotypes could provide further information on ancestry and distinguish between these two possibilities. Additionally, if AFLP profiles appear more

"mixed" in other populations compared to the pure populations, this may indicate hybridization in the other populations. Overall, individuals in R1 and M2 were entirely or predominately assigned to their respective clusters; so, it is reasonable to assume these sites represent pure populations of the two parent species in question.

When STRUCTURE was run to determine the number of gene pools within the 420 individuals, two optimal K values were calculated, K=3 and K=7. K=3 was determined to represent the "true" optimal K value for a number of reasons, including morphological data, the putative distinctive populations having a value of K=2, and a higher peak value during analysis with *ad hoc*. The second value, K=7, most likely occurred because the program assigned individuals to new clusters according to the number of grandparents they had from each parent species (Pritchard *et al.*, 2000). Natural variation also exists between populations; therefore, some variation in marker expression is expected. If there is a low migration rate between the populations, this variation could be more pronounced, affecting inferences in K. Because the sites that clustered together in K=7 were not all geographically close, low migration rates probably did not significantly influence the appearance of a second K value. In addition, Pritchard *et al.* (2000) found two potential K values in a simulation study of an "ideal" admixture population and concluded the most likely biological basis was ancestry. As a result, differences in individuals grand-parentage is the most likely cause of the K=7. This is supported by the fact clusters in K=7 tended to correspond with one of the three gene pools. For example, clusters indicated by blue and green in Figure 9B only corresponded to the unidentified gene pool in Figure 9A.

Although definitive conclusions cannot be drawn for K=7, the value could hint at backcrossing (the mating of hybrids with parental species). This hypothesis is supported by the

fact individuals in some populations had about 70% of their genotype corresponding to the cluster for one ancestral species. Backcrossed individuals would ideally be expected to have 75% of their genotype corresponding to one parent species. Comparison of membership coefficients from K=3 with mitochondrial haplotypes for "hybrids" could provide further support for backcrossing. For example, an individual that mapped primarily with the cluster corresponding to *O. quinebaugensis* that also has an Ov haplotype would support the hypothesis backcrossing is occurring.

Graphical representation of the individual membership coefficients supported the hypothesis hybridization was occurring in putative hybrid populations. When the genotype data obtained from AFLP markers in this study was combined with haplotype data from previous studies (Allard, 2010; Becker *et al*., 2009; McMurrough & Saltzman, 2009), some sites showed more complicated relationships between genetic material from both *O. virilis* and *O. quinebaugensis* present in individuals. These sites were of most interest because they indicated not only hybridization, but also the potential for speciation. Site M55, for example, shows an Ov haplotype but its membership coefficient indicates that a significant percentage of the genetic material present derives from *O. quinebaugensis*. Future studies which compare the haplotypes with the genotypes on an individual level could provide further information on the prevalence of hybridization within populations. For example, these studies could demonstrate if an individual actually has mixed ancestry or is a "pure" individual within the population. Such comparisons not only serve to further confirm that hybridization is taking place, but also to suggest some asymmetry in introgression such as Ov mothers and Oq fathers.

However, STRUCTURE analysis identified three gene pools in contrast to the expected two. Two possible explanations can account for this unidentified third gene pool: first, that

hybridization is causing speciation; or second, that a third species is hybridizing with the two species of study. A map of the collection sites was used to evaluate the geographic dispersal of population ancestry to determine if the third gene pool was confined to a particular area. The third gene pool was dispersed throughout the collection sites which would support the presence of a third species. In a separate analysis of STRUCTURE which used the "non-species-specific" markers, an optimal K value of K=2 was determined. Because the "species-specific" markers were removed, the expected K value would be one (McKniff & Violette, 2012). This second gene pool observed in the "non-species-specific" dataset corresponded with the third gene pool observed in the "species-specific" dataset, which supports the presence of a third species. In addition, the second gene pool observed in the non-species-specific dataset had a strong presence in some of the putative hybrid populations that clustered with the parent species when K=3, but separately when K=7. These populations also had two *O. virilis* haplotypes (unpublished data). This could indicate a third species is hybridizing with the two species of study and would account for the two observed Ov haplotypes. As the YB12 locus was prevalent in the populations that mapped to the third gene pool, it could be a marker for the unidentified species.

In conclusion, this study suggests that the Blackstone River Valley is a site of hybridization between *O. virilis* and *O. quinebaugensis*. Additionally, a third unidentified gene pool is present in the Blackstone River Valley and may represent a third species hybridizing with the two species of this study. The prevalence of cryptic evolutionary lineages has attracted increasing research interest in recent years, such that Pfenninger & Schwenk (2007) reported 2,207 cryptic species incidences from a total of 771,931 studies that were appropriate for detecting such cryptic lineages. This increased interest may stem from the increased availability of molecular and sequencing techniques, allowing researchers to compare closely related species with more

than just morphological comparisons (Bickford *et al.*, 2006). Some evidence indicates that *O. virilis* may represent a sibling species complex with an unknown number of members. Mathews *et al.* (2008) reported evidence for a number of such cryptic lineages even when sampling within only a small portion of the described geographic range of *O. virilis* sensu lato. Further molecular analysis with more primer combinations could determine if more markers are present that distinguish the third gene pool identified in our marker collection. This could provide evidence that the third gene pool is a cryptic species that either was present or is still present within the Blackstone River Valley. The presence of this third gene pool should be a focus of further research in the Blackstone River Valley watershed, as the findings of this study, in combination with those preceding it, could have important impacts on the study of hidden biodiversity within the region. The sources of these gene pools as well as their interactions with current and native species to the region should be a part of on-going research into the evolutionary mechanisms and their possible ecological consequences in the watershed. The presence of additional gene pools and their interactions within these habitats have important implications for the evolutionary dynamics of this taxon, and investigating this particular taxon should also yield insights applicable to other freshwater species.

Works Cited

Ahern, D., England, J., & Ellis, A. (2008) The virile crayfish, *Orconectes virilis* (Hagen, 1870) (Crustacea: Decapoda: Cambiridae), identified in the UK. *Aquatic Invasions*, **3**(1): 102-104.

Ahern, R.G., Hawthorne, D.J., & Raupp, M.J. (2009) Phylogeography of a specialist insect, *Adelges cooleyi*: historical and contemporary processes shape the distribution of population genetic variation. *Molecular Ecology*, **18**: 343-356.

Allard, M.M. (2010) Investigation of Introgression in the Blackstone River Valley: Locating AFLP Markers. Worcester Polytechnic Institute: 55p.

Applied Biosystmes. (2005) GeneScan™ – 600 Liz® Size Standard. Available from http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_082407.pdf

Bachtrog, D., Thornton, K., Clark, A., & Andolfatto, P. (2006) Extensive Introgression of Mitochondrial DNA Relative to Nuclear Genes in Drosophila yakuba Species Group. *Evolution*, **60**(2): 292-302.

Becker, M., Mulhern, T., & Perry, M. (2009) Introgressive Hybridization between Native and Invasive Crayfish: A Test of Reproductive Isolation. Worcester Polytechnic Institute: 40p.

Bensch, S., & Akesson, M. (2005) Ten years of AFLP in ecology and evolution: why so few animals? *Molecular Ecology*, **14**: 2899-2914.

Bickford, D., Lohman, D.J., Sodhi, N.S., Ng, P.K.L., Meier, R., Winker, K., Ingram, K.K., & Das, I. (2006) Cryptic species as a window on diversity and conservation. *Trends in Ecology and Evolution*, **22**(3): 148-155.

Blears, M.J., De Grandis, S.A., Lee, H., & Trevors, J.T. (1998) Amplified fragment length polymorphism (AFLP): a review of the procedure and its applications. *Journal of Industrial Microbiology & Biotechnology*, **21**: 99-114.

Boecklen, W. J., & Howard, D.J. (1997) Genetic analysis of hybrid zones: numbers of markers and power of resolution. *Ecology*, **78**(8): 2611-2616.

Bussell, J.D., Waycott, M., & Chappill, J.A. (2005) Arbitrarily amplified DNA markers as characters for phylogenetic inference. *Perspectives in Plant Ecology, Evolution and Systematics*, **7**: 3-26.

Carr, S., Ballinger, S., Derr, J., Blankenship, L., & Bickham, J. (1986) Mitochondrial DNA analysis of hybridization between sympatric white-tailed deer and mule deer in west

Texas. *Proceeding of the National Academy of Sciences of the United States of America*, **83**: 9576-9580.

Coyne, J.A. & Orr, H.A. (2004) *Speciation*. Sutherland, MA: Sinauer Associates, Inc.

Crawford, L.A., Desjardins, S., & Keyghobadi, N. (2011) Fine-scale genetic structure of an endangered population of the Mormon metalmark butterfly (Apodemia mormo) revealed using AFLPs. *Conservation Genetics*, **12**: 991-1001.

Darling, J.A., Reitzel, A.M., & Finnerty, J.R. (2004) Regional population structure of a widely introduced estuarine invertebrate: Nematostella vectensis Stephenson in New England. *Molecular Ecology*, **13**: 2969-2981.

Evanno, G., Regnaut, S., & Goudet, J. (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology,* **14**(8): 2611-2620.

Hamr, P. (2002) Orconectes. In: Holdich, D.M. (Ed.), Biology of Freshwater Crayfish. Blackwell Science, Oxford, pp. 585–608.

Hobbs, Jr., H.H. (1989) An Illustrated Checklist of the American Crayfishes (Decapoda: Astacidae, Cambaridae, and Parastacidae). *Smithsonian Contributions to Zoology*, **480**: iii-236.

Integrated Taxonomic Information System. Accessed 23 Dec 2011 from: http://www.itis.gov/index.html.

Jakobsson, M. & Rosenberg, N.A. (2007a) CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**(14): 1801-1806.

Jakobsson, M., & Rosenberg, N.A. (2007b) CLUMPP: CLUster Matching and Permutation Program version 1.1.2. *Rosenberg lab at Stanford University*. Retrieved April 15 2012, from http://www.stanford.edu/group/rosenberglab/clumpp.html.

Jasra, A., Holmes, C.C., & Stephens, D.A. (2005) Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science*, **20**(1): 50-67.

Kai, Y., Nakayama, K., & Nakabo, T. (2002) Genetic difference among three colour morphotypes of the black rockfish, Sebastes inermis, inferred from mtDNA and AFLP analyses. *Molecular Ecology*, **11**: 2591-2598.

Lamb, T., & Avise, J. (1986) Directional introgression of mitochondrial DNA in a hybrid population of tree frogs: The influence of mating behavior. *Proceedings of the National Academy of Science of the United States of America*, **83**: 2526-2530.

Lindner, K.R, Seeb, J.E., Habicht, C., Knudsen, K.L., Kretschmer, E., Reedy, D.J., Spruell, P., & Allendorf, F.W. (2000) Gene-centromere mapping of 312 loci in pink salmon by half-tetrad analysis. *Genome*, **43**: 538-549.

Mathews, L., Adams, A., Anderson, E., Basile, M., Gottardi, E., & Buckholt, M. (2008) Genetic and morphological evidence for substantial hidden biodiversity in freshwater crayfish species complex. *Molecular Phylogenetics and Evolution*, **48**: 126-135.

Mathews, L. & Warren, A. (2008) A new crayfish of the genus Orconectes Cope, 1872 from southern New England (Crustacea: Decapoda: Cambariidae). *Proceedings of the Biological Society of Washington*, **121**: 374-381.

Mavárez, J., Salazar, C.A., Bermingham, E., Salcedo, C., Jiggins, C.D., & Linares, M. (2006) Speciation by hybridization in *Heliconius* butterflies. *Nature*, **441**: 868-871.

McKniff, J.T. & Violette, K.J. (2012) Investigation of the population genetics of crayfish (*Orconectes virilis*) using AFLP markers. Worcester Polytechnic Institute: 44p.

McMurrough, M. & Saltzman, L. (2009) Introgressive Hybridization between Native and Invasive Crayfish: A Study in the Blackstone River Valley. Worcester Polytechnic Institute: 56p.

Meudt, H.M. & Clarke, A.C. (2007) Almost Forgotten or Latest Practice? AFLP applications, analyses and advances. *TRENDS in Plant Sciences*, **12**(3): 1360-1385.

Mooney, H. & Cleland, E. (2001) The evolutionary impact of invasive species. *Proceedings of the National Academy of Science of the United States of America*, **98**: 5446-5451.

Moore, W.S. (1995) Inferring Phylogenies from mtDNA Variation: Mitochondrial-Gene Trees versus Nuclear-Gene Trees. *Evolution*, **49**(4): 718-726.

Nijman, I.J., Otsen, M., Verkaar, E.L.C., deRuijter, C., Hanekamp, E., Ochieng, J.W., Shamshad, S., Rege, J.E.O., Hanotte, O., Barwegen, M.W., Sulawati, T., & Lenstra, J.A. (2003) Hybridization of banteng (*Bos javanicus*) and zebu (*Bos indicus*) revealed by mitochondrial DNA, satellite DNA, AFLP and microsatellites. *Heredity*, **90**: 10-16.

Olsen, T.M., Lodge, D.M., Capelli, G.M., & Houlihan, R.J. (1991) Mechanisms of impact of an introduced crayfish (*Orconectes rusticus*) on littoral congeners, snails, and macrophytes. *Canadian Journal of Fisheries and Aquatic Sciences* **48**: 1853-1861.

Perry, W.L., Feder, J.L., & Lodge, D.M. (2001a) Hybrid zone dynamics and species replacement between Orconectes crayfishes in a northern Wisconsin lake. *Evolution*, **55**: 1153-1166.

Perry, W.L., Feder, J.L., & Lodge, D.M. (2001b) Implications of hybridization between introduced and resident Orconectes crayfishes. *Conservation Biology*, **15**: 1656-1666.

Perry, W.L., Lodge, D.M., & Feder, J.L. (2002) Importance of Hybridization Between Indigenous and Nonindigenous Freshwater Species: An Overlooked Threat to North American Biodiversity. *Systematic Biology*, **51**: 255-275.

Petruccelli, J.D., Nandram, B., & Chen, M. (1999) *Applied Statistics for Engineers and Scientists*. Upper Saddle River, NJ: Prentice Hall, Inc. pp. 408-419, 647-656.

Pfenninger, M. & Schwenk, K. (2007) Cryptic animal species are homogeneously distributed among taxa and biogeographical regions. *BMC Evolutionary Biology*, **7**: 121.

Porter, M.L., Perez-Losada, M., & Crandall, K.A. (2005) Model based multi-locus estimation of decapod phylogeny and divergence times. *Molecular Phylogenetic Evolution*, **37**: 355-369.

Pritchard, J.K., Stephens, M., & Donnelly, P. (2000) Inference of population structure using multilocus genotype data. *Genetics*, **155**: 945-959.

Pritchard J.K., Wen, X., & Falush, D. (2010) Documentation for structure software: Version 2. Available from http://pritch.bsd.uchicago.edu.

Rhymer, J.M., & Simberloff, D. (1996) Extinction by hybridization and introgression. *Annual Review of Ecology and Systematics* **27**: 83-109.

Rosenberg, N.A. (2007) DISTRUCT: a program for the graphical display of population structure. *Rosenberg lab at Stanford University*. Retrieved April 15, 2012 from http://www.stanford.edu/group/rosenberglab/distruct.html.

Rosenberg, N.A., Burke, T., Elo, K., Feldman, M.W., Freidlin, P.J., Groenen, M.A.M., Hillel, J., Mäki-Tanila, A., Tixier-Boichard, M., Vignal, A., Wimmers, K., & Weigend, S. (2001) Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics Society of America*, **159**: 699-713.

Smith T.N., & Smith, R.L. (2005) *Elements of Ecology*, 6th. San Francisco, CA: Benjamin Cummings.

Streit, B., Städler, T., Schwenk, K., Ender, A., Kuhn, K., & Schierwater, B. (1994) Natural hybridization in freshwater animals: ecological implications and molecular approaches. *Naturwissenschaften*, **81**(2): 65-73.

The Global Invasive Species Database. (2005) Accessed 21 Nov 2011 from: http://www.issg.org/database/.

United States Geographical Society. (2006) United States Geological Survey. Accessed 21 Nov 2011 from: http://pubs.usgs.gov/sir/2006/5154/pdf/sir20065154.pdf/.

Vähä, J.K., & Primmer, C.R. (2006) Efficiency of model-based Bayesian methods for detecting hybrid individuals under different hybridization scenarios and with different numbers of loci. *Molecular Ecology*, **15**: 63-72.

Vallender, R., Robertson, R.J., Friesen, V.L., & Lovette, I.J. (2007) Complex hybridization dynamics between golden-winged and blue-winged warblers (*Vermivora chrysoptera* and *Vermivora pinus*) revealed by AFLP, microsatellite, intron and mtDNA markers. *Molecular Ecology*, **16**: 2017-2029.

Vos., P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M., & Zabeau, M. (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Research*, **23**(1): 4407-4414.

Wong, A., Forbes, M.R., & Smith, M.L. (2001) Characterization of AFLP markers in damselflies: prevalence of codominant markers and implications for population genetic applications. *Genome*, **44**: 677-684.

Young, W.P., Schupp, J.M., & Keim, P. (1999) DNA methylation and AFLP marker distribution in soybean genome. *Theoretical and Applied Genetics*, **99**: 785-790.

Appendix A: Biochemistry of AFLPs

The development of the AFLP technique has provided a means to resolve rare variations in the genome of closely related populations and species. The molecular markers generated by AFLP have implications for conservation biology. For example, the markers can identify and track introduced populations, and can also be used to compare the genetic structures between native and non-native species that are closely-related evolutionarily (Darling *et al*., 2004). Markers are generated by digesting total genomic DNA with two restriction enzymes and ligating adaptors to the fragments. Primers containing selective base pairs amplify subsets of the generated fragments in pre-selective and selective PCR, which can then be resolved to identify polymorphisms. The fragments are run on a denaturing gel and bands are scored as either plus (present) or null (absent) (Vos *et al*., 1995). The protocol design allows for slight variations that attribute to the advantages of this molecular technique; however, there are several considerations and assumptions that must be accounted for when modifying the procedure for different genomic studies.

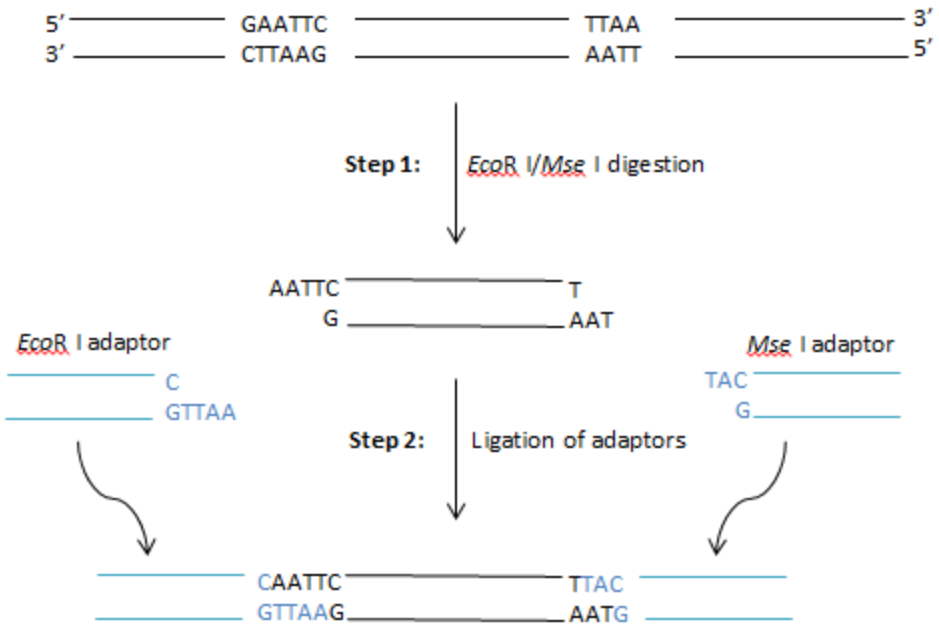

Figure A1. Schematic representation of the generation of fragments in AFLP. Nuclear DNA is digested by a rare cutter (*Eco* RI) and frequent cutter (*Mse* I) enzyme. This generates fragments with *Mse-Mse*, *Eco-Mse*, and *Eco-Eco* ends. Adaptors are then ligated to corresponding fragments.

The simultaneous digestion with two enzymes and ligation of adaptors to generated fragments contributes to the reproducibility and specificity of AFLP markers. Fragments are generated by digestion of the nuclear genome with a rare cutter (a 6 base pair recognition site) restriction enzyme and a frequent cutter (a 4 base pair recognition site) restriction enzyme (Figure A1). The frequent cutter enzyme generates a large number of fragments of ideal size for separation on a denaturing gel, while the rare cutter enzyme reduces the number of bands amplified (Blears *et al*., 1998; Vos *et al*., 1995). The reduced number of bands simplifies analysis of the fragments. *Eco* RI (with the recognition site GAATTC) is generally preferred as a rare cutter enzyme due to its reliability, which limits partial restriction digest of fragments, preventing false null or plus alleles. Other rare cutters used are *Pst* I, *Ase* I, *Hind* III, and *Apa* I (Blears *et al*., 1998; Vos *et al*., 1995; Wong *et al*., 2001). *Mse* I or *Taq* I are used as the frequent cutter enzyme to generate a large amount of fragments. Due to the AT-rich nature of eukaryotic

genomes, *Mse* I (recognition site TTAA) is ideal to optimize fragment numbers (Blears *et al.*, 1998; Vos *et al*., 1995). In conjunction with digestion, double stranded AFLP adaptor oligonucleotides (10-30 bps long) are ligated to fragments (Figure A1). Two partially complementary, synthetic oligonucleotides form dsDNA adaptors under appropriate conditions (Blears *et al*., 1998). The AFLP adaptors contain a core sequence and enzyme specific sequence, which the primers bind to during selective amplification (Figure A2). This allows the detection of polymorphisms within the genome without knowledge of the genomic sequence. Furthermore, the first base of the restriction site is changed in the adaptor to prevent digestion by the restriction enzyme, as shown in Figure A2 (Blears *et al.*, 1998; Vos *et al*., 1995). To prevent more than one adaptor from ligating to restriction fragments, the adaptors are dephosphorylated (Vos *et al*., 1995). Additionally, the simultaneous digestion and ligation reactions prevent the formation of fragment-fragment products (Blears *et al*., 1998). The specificity of the restriction enzymes contributes to the repetitive nature of AFLP.

Figure A2. Structure of *Mse* I forward AFLP adaptor. The core sequence (blue and red) is used as a binding site for AFLP primers during amplification. The base before the restriction site is changed to prevent recognition by the *Mse* I (red). The sequence of the adaptor-fragment restriction site reads GTAA instead of TTAA. The enzyme sequence binds to *Mse* I generated fragment ends (black).

Digestion with two different restriction enzymes allows a researcher to manipulate the number of fragments generated and to create multiple fingerprint combinations. Many AFLP studies have noted that markers tend to be clustered instead of dispersed throughout the genome (Lindner *et al.*, 2000; Wong *et al.*, 2001; Young *et al.*, 1999). Specifically, AFLP markers often concentrate in centromeric regions of the genome (Lindner *et al.*, 2000; Meudt & Clarke, 2007;

Wong *et al.*, 2001). This occurs because noncoding regions tend to accumulate mutations in contrast to coding regions where mutations tend to be deleterious (Wong *et al.,* 2001). The increase in polymorphisms detected near centromeric regions can further be explained by the suppression of recombination near centromeres resulting in an accumulation of mutations (Lindner *et al.*, 2000). In studies of oat and barley, AFLP markers tended to flank RFLP markers, which were both generated from cDNA libraries. RFLPs predominantly represented the cDNA libraries, which generally represent coding regions, indicating AFLP markers tend to represent non-coding regions. Since functional genes are relatively rare near centromeres, more fragments are generated near centromeric regions (Wong *et al.*, 2001). Because a high level of C methylation occurs in heterochromatic regions surrounding centromeres, variation in enzyme inhibition by methylation can manipulate the dispersal of markers. In a study of the soy bean genome by Young *et al*. (1999), *Eco* RI generated makers in clusters in contrast to *Pst* I, which generated less AFLP markers in the clustered regions because of the enzyme's sensitivity to C methylation. Thus, selecting enzymes sensitive to methylation can result in avoidance of hypermethylated sites, which can be either beneficial or detrimental to a study. Through the careful selection of two different enzymes, a researcher can manipulate the number of generated fragments (Blears *et al*., 1998; Vos *et al*., 1995). In addition, digestion with two different restriction enzymes allows labeling of one strand of the dsPCR product, which prevents the occurrence of doublets on gels due to difference in mobility of the two strands in the amplified fragment (Vos *et al*., 1995). Finally, a large number of different fingerprints can be generated from combinations of only a small number of primers when two different enzymes are used (Blears *et al.*, 1998; Vos *et al*., 1995). By digesting with two different enzymes and careful selection of the enzymes, researchers can manipulate the number of fragments generated and the

dispersal of the markers. Additionally, only a small number of primers can generate many different fingerprints.

Figure A3. Pre-selective and selective *Mse* I primers. AFLP primers contain a core sequence (blue) and an enzyme specific sequence (purple) complementary to AFLP adaptors. In addition, the primers contain selective nucleotides on the 3' end (orange). In the pre-selective primer (top), one selective nucleotide is used to amplify a subset of fragments. Selective PCR (bottom) uses two selective nucleotides to amplify a subset of the pre-selective fragments.

Selective amplifications of a subset of fragments generated by double digestion permits identification of polymorphisms within and beyond the restriction sites. AFLP primers contain a core sequence and an enzyme specific sequence identical to adaptors, but also include selective nucleotides on the 3' end (Figure A3). The selective nucleotides in the primers result in only a subset of fragments amplified since the fragments must contain the nucleotide extension (Figure A4). Each selective nucleotide reduces the number of fragments generated by about four fold (Meudt & Clarke, 2007; Vos *et al*., 1995). For simple genomes (10^6-10^7), one to two selective nucleotides are used while more complex genomes (10^8-10^9) require the use of three or more nucleotides (Blears *et al*., 1998). However, when a primer contains more than two selective nucleotides, specificity starts to be lost, with complete loss of sensitivity occurring at four selective nucleotides. More tolerance of mismatching occurs at four selective nucleotides because the 1st selective base is farther from the 3' end of the primer (Vos *et al*., 1995). Amplification of fragments with one selective base during pre-selective PCR followed by amplification of fragments with a total of three selective nucleotides increases the quality of the amplified product and is recommended for complex genomes, as shown in Figure A4 (Blears *et al*., 1998; Vos *et al*., 1995). Selective nucleotides can control the number of fragments

amplified. While there is no standard for selecting nucleotides, rare di-or tri- combinations can reduced the number of amplified fragments (Blears *et al*., 1998). For example, selecting CG combinations can reduce fragment number since the combination is relatively rare due to the tendency of a C before a G to mutate to an A. Primers using a CG selective nucleotide combination can reduce generated fragments by up to two-thirds (Bensch & Akesson, 2005). Additionally, when designing primers, they must start with a G on the 5' to prevent double banding, which results from an incomplete addition of an extra nucleotide on the synthesized strand. Furthermore, the synthesized strand should avoid ending in a C, which can also cause double banding due to a terminal transferase (Vos *et al*., 1995).

While digestion/ligation creates a mix of fragments with different cut sequences on the ends, fragments cut by both enzymes are preferentially amplified even though they represent a minority of the total fragments. Primers for the rare cutter enzyme have a higher annealing temperature than the frequent cutter enzymes, making amplification of rare-frequent cutter fragments more efficient then frequent-frequent cutter fragments, which represent the majority of generated fragments. Additionally, frequent-frequent cutter fragments have a greater tendency to form stem-loop structures by base pairing of the fragment ends because they were amplified by one primer. This competes with primer annealing (Blears *et al.*, 1998; Vos *et al*., 1995). Thus, rare cutter and frequent cutter primers can be used in the same reaction because fragments cut by both enzymes are preferentially amplified. Since fragments containing the rare cutter restriction site represent the minority of total fragments, the rare cutter primer generally is labeled with either radioactive phosphate or a fluorophore during selective PCR (Meudt & Clarke, 2007; Vos *et al*., 1995). This primer also tends to be limiting since the labeled primers are completely consumed by the end of the reaction; thus, amplification ceases when the labeled primer is used

up regardless of further thermo cycling. This results in equal concentration of fragments/alleles regardless of initial DNA concentrations (Blears *et al.*, 1998; Vos *et al.*, 1995). The use of different fluorophores to label primers also means that multiple primer combinations can be tested at a time (Meudt & Clarke, 2007). Generated fragments from selective PCR are then run on a denaturing gel and bands are either scored as plus or null (Blears *et al.*, 1998; Vos *et al.*, 1995). The inclusion of selective nucleotides in the primer design allows for the detection of polymorphisms near restriction sites because base substitutions or indels (insertions/deletions) in selective primer binding sites will result in a loss or gain of a band during amplification (Wong *et al.*, 2001). Pre-selective and selective PCR allow for amplification of subsets of fragments which can generate multiple fingerprints from a small number of primers.

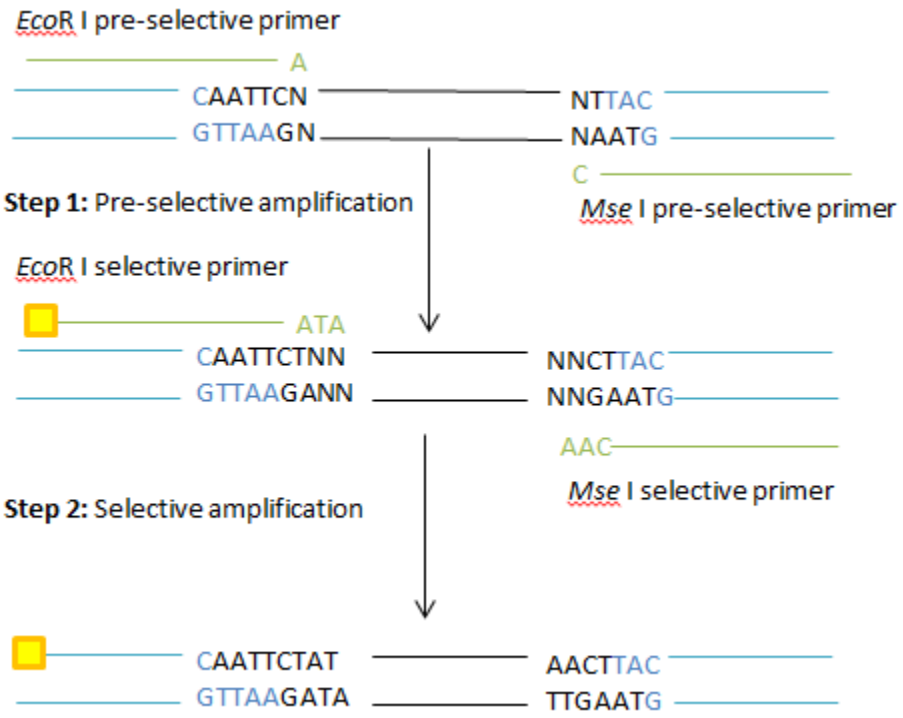

Figure A4. Schematic representation of fragment amplification. A pre-selective primer amplifies a subset of fragments from the digestion/ligation by using one selective nucleotide. A selective primer amplifies a subset of the pre-selective PCR fragments by using a total of three selective nucleotides. The rare cutter (*Eco* RI) primer has a label which identifies amplified fragments from selective PCR.

Appendix B: Sequence Information for Adaptors and Primers

Name of adaptor or primer	Sequence
Mse I F-adaptor	5’-GAC GAT GAG TCC TGA G-3’
Mse I R-adaptor	5’-TAC TCA GGA CTC AT-3’
Eco RI F-adaptor	5’-CTC GTA GAC TGC GTA CC-3’
Eco RI R-adaptor	5’-AAT TGG TAC GCA GTC TAC-3’
Mse-A primer	5’GAT GAG TCC TGA GTA AA-3’
Mse-C primer	5’-GAT GAG TCC TGA GTA AC-3’
Eco-A primer	5’GAC TGC GTA CCA ATT CA-3’
Eco-C primer	5’GAC TGC GTA CCA ATT CC-3’
Eco-ACG primer	5’-6FAM-ACTGCGTACCAATTCACG-3’
Eco-CAG primer	5’-VIC-ACTGCGTACCAATTCCAG-3’
Eco-CTC primer	5’-NED-ACTGCGTACCAATTCCTC-3’
Mse-ATC primer	5’-GAT GAG TCC TGA GTA AAT C-3’
Mse-CTC primer	5’-GAT GAG TCC TGA GTA ACT C-3’

Appendix C: CLUMPP Supplementary Manual

The following provides a summary and example of how to use the program CLUMPP to generate an output file for the graphical program DISTRUCT. In this example, Microsoft Excel will be used to create the tables for the CLUMPP input files using an individual data type.

1. CLUMPP can use two types of input files for analysis: a populations file containing a membership coefficients (Q) matrix or an individual's file containing a Q matrix. The Q matrix can be found in the simulations results page (Figure A6) or it can be found in the "q" file if the setting "Print Q-hat" was selected under the "Advanced" tab when defining parameters. The latter is the preferred method because the data has to be manually typed into Microsoft Excel if no "q" files are printed.

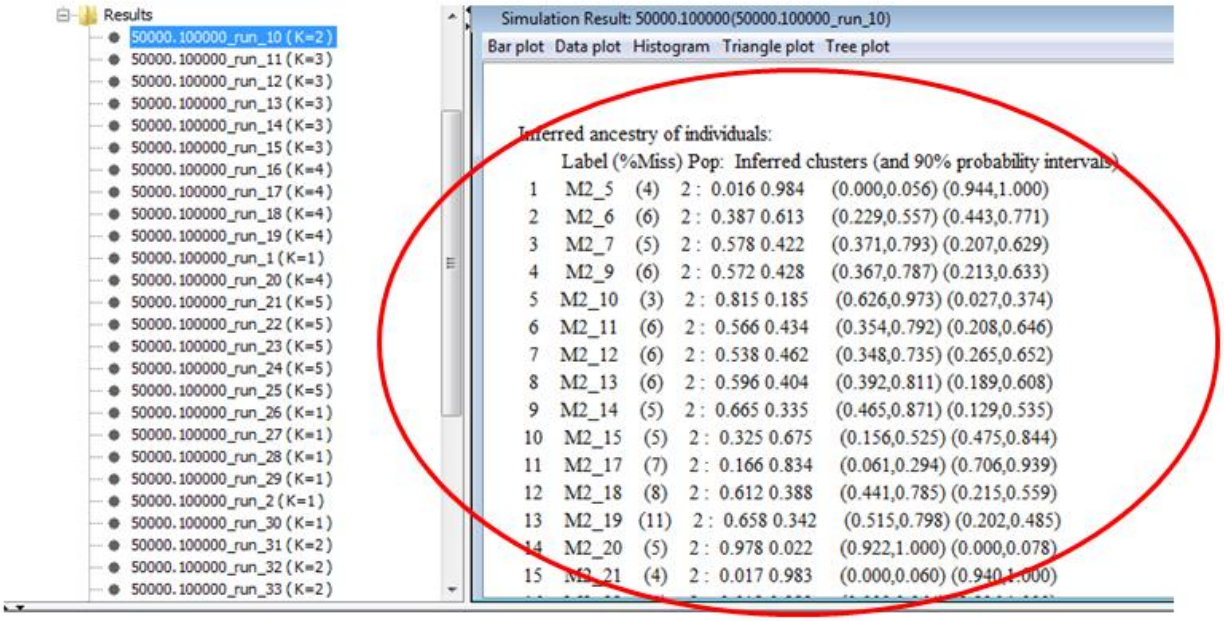

Figure A6. Location of Q matrix. The Q-matrix can be found under the "Simulation Results" summary for each individual run. The values have to be manually transferred to Microsoft Excel. The example shown is a Q matrix for an individual's file.

2. The file type for individuals is an "indfile" while it is a "popfile" for populations. The format for these files is described in the CLUMPP manual. The following will explain

one method to set up an "indfile." The Q matrix requires modifications before it can be used as in input file. First, replicated runs must be organized into one file. The results from the first run must be followed by the results from the second run (below the first run) (Table A1). CLUMPP expects a file with CXR rows (where C= number of individuals and R=number of replicas) and K+5 columns. The program will ignore columns 1, 3, 4, and 5, but column 2 will be used. Column 2 is the unique number that identifies individuals and column 4 is the putative population number. Column 5 must contain ":" and column 3 contains "(x)". The values in column 1, 2, and 4 must also be integers (Table A1). Table A1 was created in Microsoft Excel and saved as a "txt" file.

Table A1. Example of "indfile" set up. C=146, K=2, R=2

1	199	(x)	2	:	0.0164	0.9836
2	200	(x)	2	:	0.3852	0.6148
3	201	(x)	2	:	0.5835	0.4165
4	202	(x)	2	:	0.5769	0.4231
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
146	447	(x)	1	:	0.7323	0.2677
1	199	(x)	2	:	0.016	0.984
2	200	(x)	2	:	0.3899	0.6101
3	201	(x)	2	:	0.5866	0.4134
4	202	(x)	2	:	0.5803	0.4197
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
146	447	(x)	1	:	0.7242	0.2758

3. In a text editing program such as Notepad or Notepad+++, open the individuals text file saved from Microsoft Excel. To save the file as an "indfile," go to File →Save As…. Under the "Save as type" menu, select "All Files (*.*)." Then give the file the extension "indfile" instead of "txt" and save to the CLUMPP folder (Figure A7).

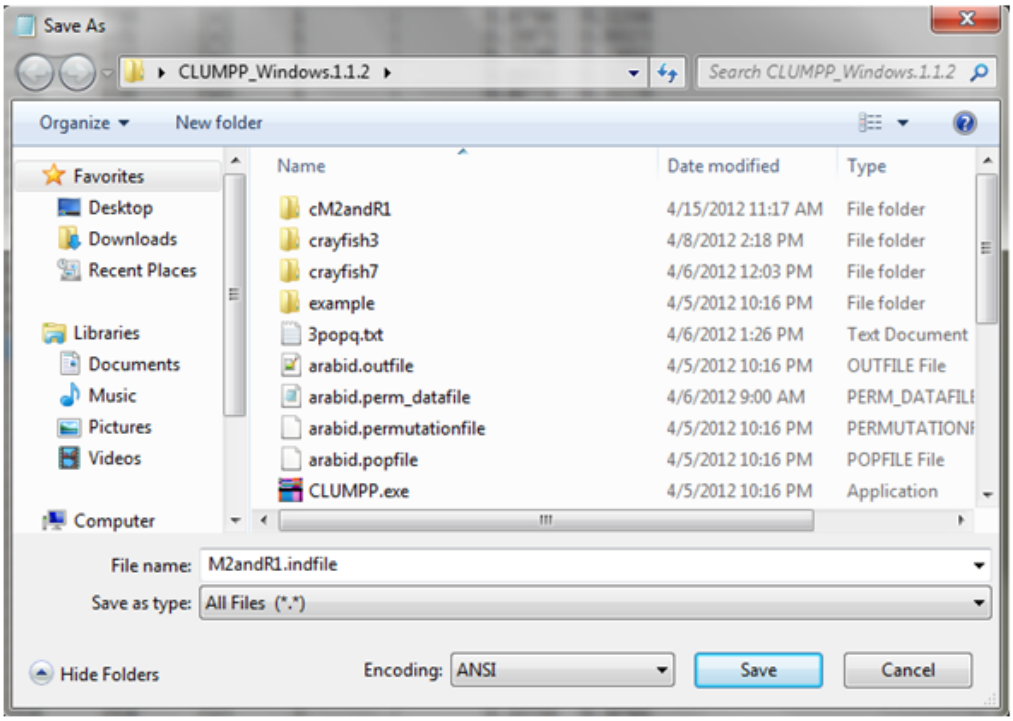

Figure A7. Saving "indfile" in Notepad.

4. In the "paramfile," the type of data and the algorithm to use will be defined. The CLUMPP manual supplies information on how to select the best algorithm to use. Use the paramfile supplied by the authors to set up the parameters by editing the file in either Notepad or Notepad+++ and saving it under a new name. The "paramfile" must be in the same directory as CLUMPP (same folder location) or the program will use the default "paramfile." The "paramfile" will be saved in the same manner as the "indfile" except with no extension. For example, the parameter file in Figure A8 is saved as "crayfishM2R1K2paramfile".
5. Figure A8 shows the settings used to analyze R1 and M2 where K=2, C=146 (146 individuals were used), and R=14 (14 independent runs were performed in STRUCTURE). "M" defines the method to be used (1, 2, or 3), "S" defines the pairwise matrix similarity statistic to be used, and "W" tells the program whether to weight by the

number of individuals in each population. The CLUMPP manual describes advanced settings based on the method selected. Give unique names for the “outfile”, “miscfile”, and “perm_data” or it will save to the example files provided by CLUMPP.

6. CLUMPP is run in a command prompt. Go to the Windows Menu and search “cmd” to find the command prompt and open it. Go the directory where CLUMPP is located using the command “cd”. For example, if the file is on the Desktop the command line would look like this: C:\Users\Name>cd Desktop\CLUMPP_Windows.1.1.2 . To run CLUMPP, type ““CLUMPP” paramfile.” The command line for the paramfile in this example is: “CLUMPP” crayfishM2R1K2paramfile. The “output” file that can be used in DISTRUCT will be saved to the CLUMPP folder under the name defined in the paramfile. Do not close the command prompt window until the program says the job finished, it can take a few hours depending on the size if the file and the algorithm chosen.

```
# --------------- Main parameters -------------------------------------------

DATATYPE 0                      # The type of data to be read in.
                              # 0 = individual data in the file
                              # specified by INDFILE, 1 = population
                              # data in the file specified by
                              # POPFILE.

INDFILE M2andR1.indfile             # The name of the individual datafile.
                              # Required if DATATYPE = 0.


OUTFILE M2andR1.outfile                 # The average cluster membership
                              # coefficients across the permuted runs
                              # are printed here.

MISCFILE M2andR1.miscfile           # The parameters used and a summary of
                              # the results are printed here.

K 2                             # Number of clusters.

C 146                           # Number of individuals or populations.

R 14                                # Number of runs.

M 2                           # Method to be used (1 = FullSearch,
                              # 2 = Greedy, 3 = LargeKGreedy).

W 0                           # Weight by the number of individuals
                              # in each population as specified in
                              # the datafile (1 if yes, 0 if no).

S 2                                 # Pairwise matrix similarity statistic
                                    # to be used. 1 = G, 2 = G'.


# - Additional options for the Greedy and LargeKGreedy algorithm (M = 2 or 3) -

GREEDY_OPTION 2                 # 1 = All possible input orders,
                              # 2 = random input orders,
                              # 3 = pre-specified input orders.

REPEATS 1000                    # If GREEDY_OPTION = 2, then REPEATS
                              # determines the number of random input
                              # orders to be tested. If GREEDY_OPTION
                              # = 3, then REPEATS is the number of
                              # input orders in PERMUTATIONFILE.

```

Figure A8. Example of set up in a "paramfile."

Appendix D: DISTRUCT Supplementary Manual

The following provides explanations and demonstrations for the program DISTRUCT, which is to be used in conjunction with the author supplied manual. This example was performed in Windows 7, using Microsoft Excel and text editors to create input files for DISTRUCT.

1. DISTRUCT requires a "popq" file with the option of also including an "indivq" file. The "indivq" file is the membership coefficient (Q) matrix generated by the program CLUMPP. The "popq" file contains a Q matrix for the populations. The Q for populations is the mean of Q for all individuals comprising the population. For example, if there were 3 individuals in a population with Qs of 0.25, 0.30, and 0.15, the Q for the population would be 0.23. Table A2 shows an example of the format for two populations.

Table A2. Format of "popq" file. The first column is the unique population identifier (must be an integer) followed by a colon. K=2 for this dataset; columns 2 and 3 are the population Q matrix. Column 4 is the number of individuals within the population.

1:	0.361958824	0.638039706	68
2:	0.685184615	0.314815385	78

2. The Q matrix for the populations and individuals must have the extension "popq" and "indivq" respectively. The Q matrix for the populations can be made in Microsoft Excel and saved as a "txt" file. This file can be opened in a text editing program such as Notepad or Notepad+++. If Microsoft Excel prevented any column formatting (such as not allowing a ":" after the population identifier), the modifications can be made in the editor. To save the population Q matrix, select File→Save as…. Select "All Files (*.*)" for the file type and save the file with the extension "popq" typed into the name bar

(Figure A9). Save the file to the DISTRUCT folder. Use the text editor to convert the "outfile" from CLUMPP to an "indivq" file and save in the DISTRUCT folder.

3. DISTRUCT provides the option of labeling the populations in a " names" file. The format is simple. The unique population identifier is put in the first column and the second column contains the population name (Table A3). A second label can also be used if desired. The user has the option of putting the labels on top and/or bottom of the figure. "1" will turn an option on and "0" will turn an option off.
4. The program gives the option of defining the colors for each cluster group. Table A4 shows the format for a "perm" file. The first column is the cluster identifier and the second column is the color code. The code for the colors can be found in the manual included with the program. The file can be made in a text editor and saved with a "perm" extension. If the user does not define the colors, the program will default and assign color 1 to cluster 1, color 2 to cluster 2, and so forth.

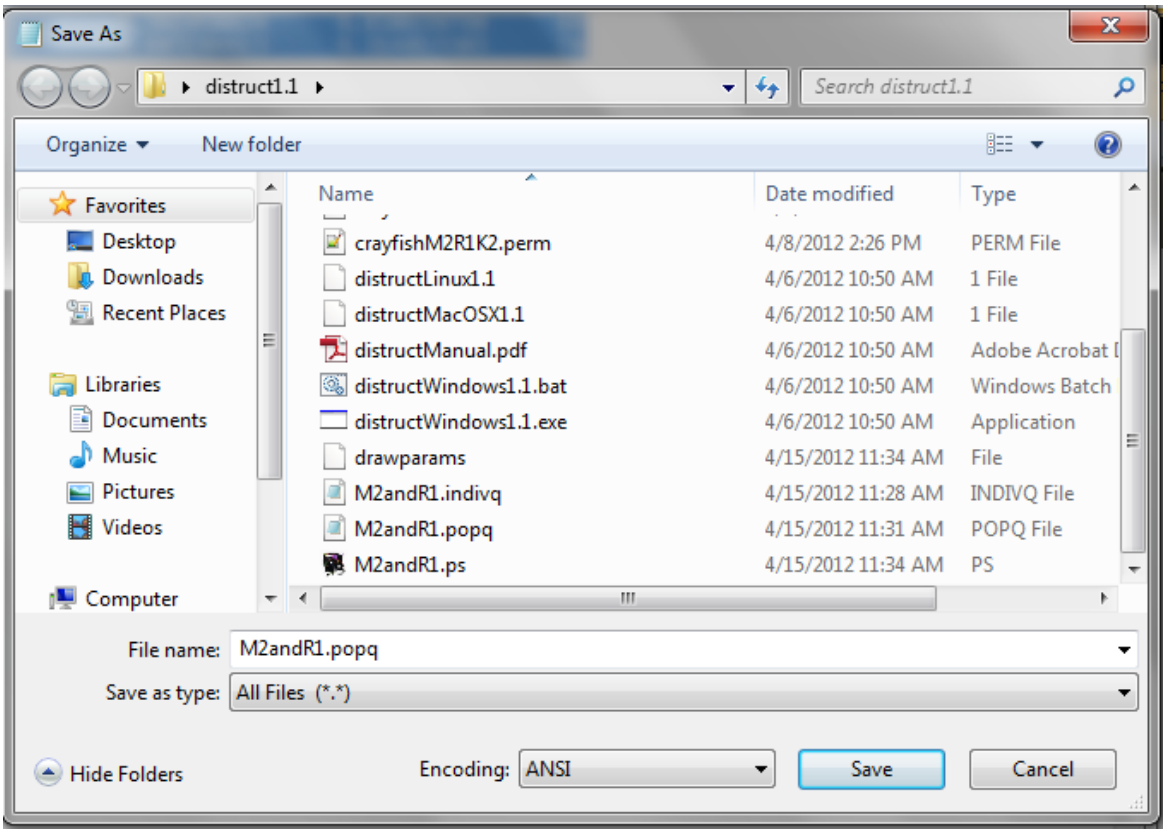

Figure A9. Saving "popq" file. Save the "popq" and "indivq" file as an "All File" type and give the appropriate extension.

5. DISTRUCT does not give the option of using a separate "paramfile," therefore the included "drawparams" file must be modified for each individual run. Drawparams is most easily edited in Word Pad or Notepad+++. Input and output file names should be changed. "Ps" is the output file and will be an image. Parameters such as K, NUMPOPS (number of populations), and NUMINDS (number of individuals) must also be changed. (Figure A10).

Table A3. Names file. A names file can be made in a text editor program such as Notepad. It should be saved with the extension "names" similar to how "popq" files are saved.

1 *O. virilis*
2 *O. quinebaugensis*

Table 4. APerm file for defining colors. "Perm" file for a dataset where K=2.

1 color40 2 color32

```
Data settings

#define INFILE_POPQ        M2andR1.popq      // (str) input file of population q's
#define INFILE_INDIVQ      M2andR1.indivq    // (str) input file of individual q's
#define INFILE_LABEL_BELOW crayfishM2R1K2.names     // (str) input file of labels for below figure
#define INFILE_CLUST_PERM  crayfishM2R1K2.perm     // (str) input file of permutation of clusters to print
#define OUTFILE            M2andR1.ps       //(str) name of output file

#define K   2    // (int) number of clusters
#define NUMPOPS 2    // (int) number of pre-defined populations
#define NUMINDS 146  // (int) number of individuals
```

Figure A10. Drawparams file name and value changes. These are the file names and values that must be changed for the input data.

6. The user can decide to either print individuals or populations for the graphical output. In addition, the user can print labels above and/or below the figure and lines to separate populations (Figure A11).

```

Main usage options

#define PRINT_INDIVS      1  // (B) 1 if indiv q's are to be printed, 0 if only population q's
#define PRINT_LABEL_ATOP  0  // (B) print labels above figure
#define PRINT_LABEL_BELOW 1  // (B) print labels below figure
#define PRINT_SEP         1  // (B) print lines to separate populations
```

Figure A11. Output options. These are the commands that can be modified to control DISTRUCT output.

7. The layout of the graphical output can also be controlled based on the size of the samples. The appearance of the figure can be modified to increase the height of the box, the font size of labels and the width of the segments representing individuals (Figure A12). In addition, the orientation of the figure can be changed to fit the output page (standard 8.5 X 11 sheet of paper). For example, the figure can either be oriented vertically (1) or horizontally (0). The figure is moved in pixel units and the font size is also changed in

pixel units (Figure A13). When modification of the "drawparam" file is complete, it must be saved under the same name.

```
30  Figure appearance
31
32  #define FONTHEIGHT 14    // (d) size of font
33  #define DIST_ABOVE 5     // (d) distance above plot to place text
34  #define DIST_BELOW -7    // (d) distance below plot to place text
35  #define BOXHEIGHT  50    // (d) height of the figure
36  #define INDIVWIDTH 1.7   // (d) width of an individual
37
```

Figure A12. Figure appearance commands.

```
39  Extra options
40
41  #define ORIENTATION 1        // (int) 0 for horizontal orientation (default)
42                   //       1 for vertical orientation
43                   //       2 for reverse horizontal orientation
44                                //       3 for reverse vertical orientation
45  #define XORIGIN 160     // (d) lower-left x-coordinate of figure
46  #define YORIGIN 40      // (d) lower-left y-coordinate of figure
47  #define XSCALE 1        // (d) scale for x direction
48  #define YSCALE 1        // (d) scale for y direction
49  #define ANGLE_LABEL_ATOP 60 // (d) angle for labels atop figure (in [0,180])
50  #define ANGLE_LABEL_BELOW 60    // (d) angle for labels below figure (in [0,180])
51  #define LINEWIDTH_RIM  3    // (d) width of "pen" for rim of box
52  #define LINEWIDTH_SEP 1 // (d) width of "pen" for separators between pops and for tics
53  #define LINEWIDTH_IND 0.3   // (d) width of "pen" used for individuals
54  #define GRAYSCALE 0         // (B) use grayscale instead of colors
55  #define ECHO_DATA 1             // (B) print some of the data to the screen
56  #define REPRINT_DATA 1          // (B) print the data as a comment in the ps file
57  #define PRINT_INFILE_NAME 0     // (B) print the name of INFILE_POPQ above the figure
58                                  //     this option is meant for use only with ORIENTATION=0
59  #define PRINT_COLOR_BREWER 1    // (B) print ColorBrewer settings in the output file
60                                  //     this option adds 1689 lines and 104656 bytes to the output
61                                  //     and is required if using ColorBrewer colors
62
```

Figure A13. Command lines to change figure orientation.

8. Verify that all required files are in the same folder as the program DISTRUCT. Run DISTRUCT by simply clicking on the executable file. The output will be a graphical representation of the Q matrix (Figure A14) in the "ps" file.

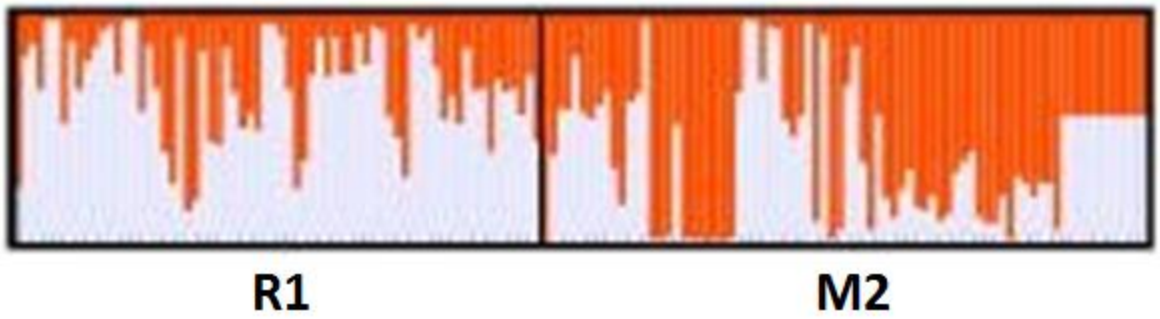

Figure A14. Example of DISTRUCT output.