Investigation of the population genetics of crayfish (Orconectes virilis) using AFLP markers

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
by

________________________________________________
Jessica McKniff

________________________________________________
Kortni Violette

April 26, 2012

Approved:

________________________________________________
Professor Michael Buckholt

________________________________________________
Professor Lauren Mathews
Abstract

Crayfish are one of many freshwater organisms that are subject to biodiversity changes. Environmental factors play an important role in the genetic composition of a population. Using AFLP markers we were able to better understand the population genetics of 418 *O. virilis* crayfish from 23 sites in the Blackstone River Valley by identifying interactions between these various populations and lineages.
**Acknowledgements**

We would like to thank our project advisors, Professor Mathews and Professor Buckholt for their guidance and support throughout the year. We would also like to thank Alyce Buchenan, Hiral Dutia and Emily Scott–Solomon for working with us in conjunction on this project. We would like to acknowledge David Rolle for technical assistance using the program Structure.
# Table of Contents

Abstract ................................................................................................................................. 2

Acknowledgements ................................................................................................................ 3

Table of Figures ..................................................................................................................... 5

Tables .................................................................................................................................... 5

Table of Equations .................................................................................................................. 5

1. Introduction .......................................................................................................................... 6
   1.1 Evolutionary Processes ...................................................................................................... 6
   Natural Selection .................................................................................................................. 7
   Genetic Drift ....................................................................................................................... 7
   Gene Flow ........................................................................................................................... 8
   1.2 Landscape Ecology ......................................................................................................... 9
   Landscape ecology, freshwater systems, and the application of molecular markers .......... 9
   1.4 Biology & Ecology of Freshwater Crayfish .................................................................. 14
   1.5 The Blackstone River Valley ......................................................................................... 19
   1.6 Objective & Hypothesis ............................................................................................... 22

2. Materials and Methods ...................................................................................................... 23
   2.1 Crayfish Collection & DNA extraction .......................................................................... 23
   2.2 AFLP Protocol .............................................................................................................. 23
      Digestion Ligation ............................................................................................................ 24
      Preselective PCR ........................................................................................................... 26
      Selective PCR ............................................................................................................... 27
      Sample Preparation ...................................................................................................... 28
   2.3 Fragment Analysis ........................................................................................................ 29
   2.4 Allele Selection ............................................................................................................ 30
   2.5 Analysis of Population Genetics .................................................................................. 31
      Structure ......................................................................................................................... 31
      CLUMPP and Distruct ................................................................................................. 32

3. Results and Discussion ...................................................................................................... 34
   3.1 Estimating the Number of Populations ......................................................................... 34
   3.2 Aligning the Populations ............................................................................................. 35
   3.3 Conclusions .................................................................................................................. 39

References .............................................................................................................................. 41
Table of Figures

Figure 1: Family Distribution of Freshwater Crayfish (Carnegie Museum, 2006) ................................................................. 15
Figure 2: Natural Range of O. virilis in North America (Hobbs, 1988) .................................................................................. 15
Figure 3: Species localities for Orconetes virilis in New England (Crocker, 1979) ............................................................. 16
Figure 4: Dorsal and ventral views of general male crayfish (Hobbs, 1972) & Dorsal and ventral views of O. virilis 18
Figure 5: Twenty three collection sites in the Blackstone River Valley .................................................................................. 21
Figure 6: Diagram showing how EcoRI enzyme is used to produce sticky ends. On the left is the DNA fragment with arrows indicating where EcoRI will cut. On the right it shows the two sticky ends with the overhang that are produced after the EcoRI cut. ............................................................... 24
Figure 7: Image from a set of successful digestion-ligation reactions. The two outer bands are the 100 bp ladders used for reference. Inside lanes show significant smearing which indicates that there are many DNA fragments present in each sample. ....................................................................................................................... 25
Figure 8: Plot of ΔK vs. K, obtained from Formula 1. The large peak at K=2 indicates that there are 2 distinct populations that can be observed in the Blackstone River Valley. ................................................................. 34
Figure 9: Bar plot from one run of Structure at K=2. The two colors represent alleles from each of the 2 different populations. Each individual crayfish is represented by a vertical line that represents that percent of alleles they exhibit from each assumed population. ............................................................................................................... 35
Figure 10: Graph of the individual Q-matrix (membership coefficients) obtained from CLUMPP and Distruct. ....... 35
Figure 11: Map of the Blackstone River Valley with histograms for each location. .......................................................... 37

Tables

Table 1: Collection Sites .......................................................................................................................................................... 20

Table of Equations

Equation 1: This is the equation used to calculate the ΔK that is plotted for each K in order to determine the most likely number of true populations present. It is the second order rate change of Ln K, the log likelihood of the probability of the data. For this calculation we used data from 5 runs of Structure. .................................................. 32
1. Introduction

Population genetics is the study of the genetic composition of biological populations (Okasha, 2008). Population genetics is not an exact field of study; it involves both using known genetic inheritance to construct theoretical models of evolution and analyzing actual data sets obtained from populations (Charlesworth, 2001). Studies may be based on phenotypic or genotypic data; however, while genotypic data are harder to obtain, they provide more evidence for variation both within and among populations. There are several methods used for extracting and analyzing genetic data, but this study will focus on amplified fragment length polymorphisms (AFLP). We employ these markers to obtain insights into the population genetics of an invasive species of crayfish, Orconectes virilis. We focused on a set of populations of freshwater crayfish in New England because freshwater organisms are subject to biodiversity changes; therefore, this investigation will add to our understanding of interactions between populations and lineages.

1.1 Evolutionary Processes

Evolution is the change in allele frequencies across generations. Population genetics integrates genetics with evolution in order to analyze the genetic variation in populations. There are various factors that affect the genetic composition of a population such as natural selection, genetic drift, gene flow, and mutations. Mutations are important in population genetics because they are capable of incorporating new alleles into a population, and are due to an error or alteration of a genome during the replication of a nucleotide sequence (Futuyma, 2009). Mutation is especially crucial when it comes to dispersal and differentiation of populations. Numerous empirical studies examine the role of evolutionary forces within populations; the role of several or just one of these processes is predicted by making simplified assumptions about the
rest. In this section, we briefly describe the evolutionary processes relevant to our project and provide empirical examples of how previous investigations have investigated them.

**Natural Selection**

Natural selection, or the changes in allele frequency driven by differential reproductive success, is known to play an important role in population genetics. Alleles with high fitness levels are essentially selected for, becoming more frequent throughout subsequent generations (Masel, 2011). Studies in evolutionary ecology look into the power of natural selection in increasing fitness criteria within limitations imposed by constraints. Natural selection can result in local adaptations if selection pressures differ among populations. For example, in a study of the copepod *Diaptomus sanguineus*, researchers documented shifts in timing of diapause induction due to natural selection. In a pond containing predatory fish, copepods switch from making hatching eggs to diapausing eggs each year at the end of March; diapause induction starts a month later in a neighboring pond without fish. Due to the close proximity of the ponds, the researchers inferred that dispersal most likely occurs. This means the copepods are laying diapausing eggs at a time of the year that is suboptimal for the new pond; therefore, dispersers will have decreased fitness. Differences among ponds in growth rates, hatching cues, and adaptations to timing of reproduction result from natural selection (Hairston and Walton, 1986; Bohonak and Jenkins, 2003).

**Genetic Drift**

Genetic drift is the random change in allele and genotype frequencies in a population from generation to generation due to sampling error (Futuyma, 2009). Genetic drift is eliminates alleles and would in most cases produce a decrease in the heterozygosity of the population. The
strength of genetic drift depends on the number of adults in a population that will reproduce (Mitton, 2001). Typically, genetic drift is stronger in smaller populations; the fewer individuals in a population, the more genetic drift affects the population. Neraas and Sruell (2001) described a set of populations of bull trout that are likely to be subjected to strong genetic drift. In the Clark Fork river system, Twin Creek enters the river downstream of the Cabinet Gorge Dam, but is genetically more similar to sites above the dam. The habitat in the creek may have led to a small population subject to the impacts of genetic drift (Neraas and Spruell, 2001). Genetic drift is important in population genetics because it drives polymorphic loci to fixation for one allele. If an allele becomes fixed it is essentially permanently lost (unless reintroduced by a mutation) which causes a reduction in genetic diversity. Also, it is an important concept because it increases genetic divergence between gene pools, especially when there is a lack of gene flow, because they are opposing forces.

**Gene Flow**

Gene flow is the movement of alleles among populations (Mitton, 2001). This is caused by migration of individuals from one population to another, adding or removing different alleles from the respective gene pools. Gene flow is an important concept because it allows scientists to understand how populations adapt and how mutations may be spreading (Manel et al.2003). When gene flow occurs between populations it tends to homogenize them (Mitton, 2001). This is because gene flow can result in alleles being reintroduced in populations from which they had been lost through drift or selection. Extensive gene flow would theoretically create one large gene pool, a situation known as “panmixia”.

8
1.2 Landscape Ecology

Landscape ecology is closely related to the field of population genetics because it seeks to identify how environmental features influence evolutionary processes; therefore, environmental features can potentially correlate with genetic discontinuities among populations (Nevillel et al. 2006). As previously mentioned, there are several factors that affect the genetic composition of a population. The most frequently observed genetic patterns are metapopulations, random patterns, isolation by distance, clines, and genetic boundaries. Metapopulations are populations that are for the most part spatially separate, but that still have some sort of interaction. Isolation by distance simply refers to the fact that populations that are distant have more genetic differentiation because they do not interact, as nearby populations do. Genetic boundaries refer to anything that may separate populations, including clines, dams, or any other potential barriers of gene flow. These are caused by landscape variables and evolutionary processes discussed earlier. Due to the evolutionary forces, the majority of populations of species reveal certain levels of genetic differentiation as structured by various historical processes, environmental barriers, and life histories (Storfer et al. 2010) Studying the landscape ecology of a population helps us to understand the genetic structure of a population in a specific area, and gives insight into the micro evolutionary processes occurring (Manel et al. 2003).

Landscape ecology, freshwater systems, and the application of molecular markers

Molecular markers are used to gain insight into the role of landscape features in structuring populations. Population structure is determined by analysis of the genetic variation among individuals at molecular markers. Once discontinuities in the geographical pattern of genetic variation have been identified, the impact of landscape features on population subdivision can be considered. The yellow perch in the freshwater section of the Saint Lawrence River were
analyzed with microsatellites markers (Leclerc et al. 2008). Once the genetic diversity was quantified, 4 distinct populations were defined. To test the influence of landscape on genetic discontinuities, information was gathered about each landscape feature – dams, water masses, turbidity, temperature, and spawning habitats- and converted into a measure of ecological distance between sampling sites. Mantel tests were used to show the relationships between genetic distance, geographical distance, and environmental distances; the results revealed their potential influence in shaping the genetic structure of the yellow perch. The presence of physical barriers played a significant role in the pattern of genetic populations structure. We can infer that the identified genetic discontinuities in the yellow perch are associated with dams because genetic isolation can develop over time due to barriers to gene flow. All other factors showed a low, but almost significant correlation between genetic and environmental differences; therefore a weak effect on genetic differentiation of populations (Leclerc et al. 2008).

Molecular markers are important in the field of evolutionary biology because they can provide useful information that cannot be obtained with phenotypic data, and can have a wide array of applications. Some potential problems scientists may try to address using genetic markers are risk of extinction, identifying hybridization, and determining population structure (Crawford et al. 2011). With advances in technology continuously occurring, there are now several methods available to collect genetic data for analysis. These methods include amplification and sequencing of mtDNA, amplification and characterization of microsatellite loci, and AFLP.

In all of these methods, loci are typically sampled in the genomes of several individuals in the population. It is important to sample as many loci as possible because this provides insight into
the genetic differences among and within certain populations. Additionally, using numerous loci addresses the challenge AFLP shows when utilized for assigning parentage and excluding parentage. To achieve this, AFLP requires more loci, unlike microsatellites that require only a few loci (Bensch and Akesson, 2005). Loci under selection allow scientists to compare the dispersal of neutral and selected markers and can aid in understanding local adaptation and speciation (Manel et al. 2003). As previously mentioned, microsatellites, as opposed to AFLPs, were used as molecular markers to investigate the landscape genetics of yellow perch in the St. Lawrence River (E. LeClerc et al., 2008). This study aimed to identify the population structure of yellow perch by using microsatellites to determine the extent of genetic variation. This genetic data was assessed along with the landscape features and environmental barriers to examine how the environment influences gene flow in the St. Lawrence River. Using microsatellites the researchers identified a pattern of isolation by distance. Dams, along with other factors such as water speed and turbidity, were shown to have a significant impact on the population structure as well.

The AFLP technique of fingerprinting DNA is a strongly developed molecular marker approach, with numerous uses in population genetics, linkage mapping, development of markers in single locus PCR, shallow phylogenetics, and in analysis of parentage. Advances in technology have provided new opportunities for analysis of data, and current experiments have addressed certain areas of the approach, including comparison to other processes of genotyping, homoplasy, error assessment and phylogenetics signal, among others. The technique has become one of the commonly used methods of study and analysis of plants, bacteria, animals, and fungi (Meudt and Clarke, 2007). We will now review recent studies using molecular markers to highlight the
importance features of these various methods, and to demonstrate why we chose to conduct our study using AFLPs.

The method we will be using to collect genetic data is AFLP PCR. AFLP was developed in the 1990’s and has many advantageous aspects. The main benefits of using AFLP are the convenience, reproducibility, and cost. AFLPs are convenient because unlike microsatellites, no previous knowledge of genetic information of a species is required (Campbell et al. 2003). Also, AFLPs are capable of producing many genetic markers from one assay, and can be used in numerous applications (Trybush et al. 2004). This is a considerable advantage over microsatellites because microsatellites loci are harder to obtain because of the cost and time constraints, and a small number of loci drastically decreases the statistical power of analysis (Campbell et al. 2003).

In the study of the freshwater insect *Calopteryx splendens*, AFLP markers were used to evaluate the genetic differentiation across several watercourses. Recent studies showed that AFLPs provide tools for addressing questions that require the estimation of relatedness between individuals and for estimating genetic structure (A. Chaput-Bardy et al. 2008). The question that they sought to address was whether or not overland dispersal was occurring between watercourses. They used this question to form two different hypotheses. The first hypothesis was that dispersal was only occurring in-stream, which would create genetic differentiation between watercourses and an isolation by distance pattern to form within the watercourses as a result of downstream gene flow. Conversely, they also hypothesized that dispersal was also due to overland dispersal, which would create almost no genetic differentiation between watercourses.
Analysis of the AFLP markers by several statistical methods such as model based clustering gave reasonable data to support the hypothesis of overland dispersal, due to the lack of genetic differentiation between watercourses. Another example of the application of the AFLP technique was in the study of two subspecies of willow warblers that are closely related. The technique was first utilized to test a group of extreme phenotypes of the two species. Bands that were observed to have major dissimilarities were put under new tests to prove that the skewed profile was not an error. The technique thus is critical in identifying hybrid species even in cases where microsatellite techniques have been found limiting. Once more, it is possible to create numerous polymorphic markers in a short period; something that makes the AFLP approach more preferable for making out hybrids (Bensch and Akesson, 2005).

However, there are also several drawbacks associated with AFLP. For instance, AFLPs often fail to amplify fragments with high molecular weights, which is problematic because polymorphic fragments are often large. Another disadvantage is that even when AFLP methods do generate long fragments, they can have a decrease in signal and uneven peak intensity, leading to unreliable data. Also, AFLPs are sometimes considered problematic because they are dominant markers, meaning they cannot differentiate between heterozygous or homozygous individuals (Foll et al. 2010). Unlike AFLPs, microsatellites are codominant markers, which make them useful in evaluating deviations from the Hardy-Weinberg equilibrium (Campbell et al. 2003). However, our application of AFLP will be examining presence or absence of fragments, which will not require dominance information. Because AFLP have uses and application when it comes to generation of sufficient polymorphic markers to address or determine extremely small individual dissimilarities on the level of DNA fingerprints, it is usually more suitable and
applicable for identification of individuals, analysis of parentage and measurement of pair wise relations, among other applications (Bensch and Akesson, 2005).

1.4 Biology & Ecology of Freshwater Crayfish

The study of landscape ecology typically focuses on the study of individuals. We will be focusing on the species *O. virilis* and assuming all sampled individuals are of this species. Crayfish are decapod crustaceans divided among three families: Parastacidae, Astacidae, and Cambaridae (Hobbs, 1989). The main distributions are Astacidae, and Cambaridae in the Northern Hemisphere, and Parastacidae in the Southern Hemisphere (Figure 1). North America is one of the two centers of biodiversity for freshwater crayfishes; this is where 80% of the cambarid species can be found (Carnegie Museum, 2006). Cambaridae represents more than 75% of the over 500 species of crayfish documented (Taylor, 2002). There are 3 major genera in the Cambaridae family; *Procamburus*, *Cambarus*, and *Orconetes*. *Orconetes* is comprised of 11 subgenera, 81 species, and 11 subspecies (Harm, 2002).
The species *Orconetes virilis* has a widespread native range in Canada from Saskatchewan to Ontario and in the United States from Montana to New York (Figure 2); it is considered an invasive species over the rest of the United States (Hobbs, 1974; Global Invasive Species Database, 2006).
The origins of *O. virilis* in New England are unknown, but Crocker (1979) reported on localities in which he had collected members of the *O. virilis* species complex (Figure 3). Introduction of this species to the area have occurred, but whether its distribution was from a natural or anthropogenic effect has not been determined (Crocker, 1979). Ballast water and the use of crayfish as fish bait have led to anthropogenic introductions of the species in freshwater systems (Lodge, 2000). Nonindigenous species introduction, species established outside of their native range, leads to introgression with the endemic gene pool (Perry, 2001; Perry, 2002).

**Figure 3**: Species localities for *Orconetes virilis* in New England (Crocker, 1979)
This species inhabits rivers, streams, lakes, ponds, and marshes. They prefer water with rocky substrates, vegetation, and debris like logs, presumably to hide from predators. They are most active when the water is warm from May to September; therefore, mating occurs in the autumn and the eggs are laid in the spring. The sperm is stored over the winter in the annulus ventralis, the sperm receptacle, of female crayfish. Unlike most decapod crustaceans, larval stages occur within the egg and newly hatched juvenile crayfish stay attached to their mother as they continue to mature before they leave; also known as direct development. This developmental cycle, unlike other crustaceans whose hatched eggs are free-living larval, presents a unique dispersal pattern. Larval are moved downstream by the mother, which promotes dispersal and genetic connectivity between populations (Reynolds, 2002; Scholtz et al. 2002).

Crayfish grow by molting, a process in which their exoskeleton sheds to allow expansion. Mature *O. virilis* males alternate between two morphological forms; they molt from Form II (a sexually immature form) to Form I (a sexually mature form) during mating season then back to Form II outside of mating season (Harm, 2002; Global Invasive Species Database, 2005). *O. virilis* is a crustacean identified by its reddish-brown to olive-brown body and abdomen; the chelae and legs have a blue tiny with yellow tubercles. Adults range from 10 to 12 centimeters in size, excluding the chelipeds and antennae (Figure 4).
Crayfish are an ideal model system for studying population genetics and landscape ecology. They are invertebrates, which mean there are fewer legal restrictions. Also, crayfish are relatively easy to collect and allow for a large number of samples (Cowie and Holland, 2008). More specifically, North America is a center of high crayfish biodiversity. The crayfish have
undergone substantial radiation, though the causes for this are unknown. The described range of *O. virilis* is the broadest of any crayfish in North America (Perry, 2002; Taylor, 2002; Mathews and Warren, 2008); the invasive tendencies of *O. virilis* allows for the study of interactions between populations and lineages.

### 1.5 The Blackstone River Valley

The Blackstone River Valley is a watershed that runs from central Massachusetts to Rhode Island. The entire area of the river and its tributaries is made up of a water system, forests, and wetlands. There are dams throughout the water system; these not only modify the stream flow patterns, but cause fragmentation of river populations because dams do not allow fish species to freely migrate throughout the river system (Neraas, 2001; Barbaro and Zarriello, 2006). *O. virilis* is one of several species of crayfish found in the Blackstone River Valley. Our investigation focuses on crayfish collected from 23 sites from the part of the watershed in Massachusetts (Table 1, Figure 5). All 23 of these sites are located in the same drainage basin, but are subdivided among different rivers in the basin. In landscape ecology it is important to collect and analyze data both across and within selection gradients (Manel et al., 2003). In the Blackstone River Valley the selection gradients will mainly be dams, but there could potentially be other barriers such as water quality, temperature, and currents.
<table>
<thead>
<tr>
<th>Sire</th>
<th>Waterbody</th>
<th>Location</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>M2</td>
<td>Salisbury Pond</td>
<td>Worcester</td>
<td>Lentic</td>
</tr>
<tr>
<td>M4</td>
<td>Quinsig. R.</td>
<td>Grafton</td>
<td>Loctic</td>
</tr>
<tr>
<td>M9</td>
<td>Blackstone R.</td>
<td>Millbury</td>
<td>Loctic</td>
</tr>
<tr>
<td>M10</td>
<td>Tatnuck Brook</td>
<td>Worcester</td>
<td>Loctic</td>
</tr>
<tr>
<td>M23</td>
<td>White Pond</td>
<td>Leicester</td>
<td>Lentic</td>
</tr>
<tr>
<td>M34</td>
<td>Ramshorn Brk</td>
<td>Auburn</td>
<td>Lentic</td>
</tr>
<tr>
<td>M55</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M66</td>
<td>Mumford R.</td>
<td>Whitinsville</td>
<td>Loctic</td>
</tr>
<tr>
<td>M70</td>
<td>Meadow Pond</td>
<td>Whitinsville</td>
<td>Lentic</td>
</tr>
<tr>
<td>M90</td>
<td>Mill R.</td>
<td>Blackstone</td>
<td>Loctic</td>
</tr>
<tr>
<td>M91</td>
<td>Muddy Brk</td>
<td>Hopedale</td>
<td>Loctic</td>
</tr>
<tr>
<td>M93</td>
<td>Mill R.</td>
<td>Hopedale</td>
<td>Loctic</td>
</tr>
<tr>
<td>M95</td>
<td>Kettle Brk</td>
<td>Leicester</td>
<td>Loctic</td>
</tr>
<tr>
<td>M96</td>
<td>Kettle Brk</td>
<td>Auburn</td>
<td>Loctic</td>
</tr>
<tr>
<td>M98</td>
<td>Mumford R.</td>
<td>Douglas</td>
<td>Loctic</td>
</tr>
<tr>
<td>M102</td>
<td>Blackstone R.</td>
<td>Milleville</td>
<td>Loctic</td>
</tr>
<tr>
<td>M103</td>
<td>Peters R.</td>
<td>Milleville</td>
<td>Loctic</td>
</tr>
<tr>
<td>M106</td>
<td>Newton Pond</td>
<td>Worcester</td>
<td>Lentic</td>
</tr>
<tr>
<td>M107</td>
<td>West R.</td>
<td>Upton</td>
<td>Loctic</td>
</tr>
<tr>
<td>M108</td>
<td>Crystal Lake</td>
<td>Douglas</td>
<td>Lentic</td>
</tr>
<tr>
<td>M109</td>
<td>West R.</td>
<td>Upton</td>
<td>Loctic</td>
</tr>
<tr>
<td>M110</td>
<td>Ramshorn Brk</td>
<td>Auburn</td>
<td>Loctic</td>
</tr>
<tr>
<td>R1</td>
<td>Blackstone R.</td>
<td>Woonsocket</td>
<td>Loctic</td>
</tr>
</tbody>
</table>
Figure 5: Twenty three collection sites in the Blackstone River Valley
1.6 Objective & Hypothesis

We have used AFLP markers to gain insight into the genetic relationships among crayfish populations over most of the area of the Blackstone River Valley. In addition, we have used this dataset to test a hypothesis that is based on much background research on landscape ecology and crayfish. As mentioned previously, this watershed is composed of many streams feeding into the main river, which flows north to south. We hypothesize that there will be genetic differentiation between the tributaries that feed into the Blackstone River.
2. Materials and Methods

The following section describes in detail the materials and methods that we used to perform our experiments and analyze the data we obtained. We will begin with a summary of crayfish collection and DNA extraction, which were performed in previous projects. Next we will provide a complete protocol of AFLP PCR, including digestion ligation, pre-selective PCR, selective PCR and sample preparation. To conclude this section we will discuss our analysis methods.

2.1 Crayfish Collection & DNA extraction

Live crayfish were collected from 23 different sites from the Blackstone River Valley watershed in Massachusetts and Rhode Island between June and October of 2008 as described in Saltzman & McMurrough (2009). For most sites, species of the particular samples were not identified, but all of them belong to the *O. virilis* species complex. Extraction of genomic DNA was performed as described in Saltzman & McMurrough (2009) and Becker et al. (2009).

2.2 AFLP Protocol

The following protocols were used in order to generate AFLP profiles for each crayfish in the collection. This process included many steps including digestion ligation, preselective PCR, selective PCR and sample preparation.
Digestion Ligation

The first step in AFLP PCR is digestion ligation, which fragments the DNA using restriction enzymes. For this experiment we used EcoRI and MseI for our restriction enzymes. MseI is a four-cutter enzyme that restricts the DNA at 5’-T/TAA-3’, and EcoRI is a six cutter enzyme that restricts at 5’-G/AATTC-3’. Digestion ligation with these restriction enzymes produces many fragments of DNA with overhanging sticky ends that are used for ligation of adaptors. In this study, we used adaptors with sequences as follows: Eco RI-adaptor (5’-CTCGTAGACTGCGTACC-3’; 3’-CTGACGCAATGGTTAA-5’), Msel-adaptor (5’-GACGATGAGTCCTGAG-3’; 3’-TACTCAGGACTCAT-5’).

![EcoRI enzyme and sticky ends diagram](image)

**Figure 6:** Diagram showing how EcoRI enzyme is used to produce sticky ends. On the left is the DNA fragment with arrows indicating where EcoRI will cut. On the right it shows the two sticky ends with the overhang that are produced after the EcoRI cut.

We carried out restriction digestion and ligations of adaptors in a single reaction as follows. First we prepared a master mix using the following components in each reaction tube:

- T4 ligase buffer (10X) 1.25 µl
- NaCl (0.5 M) 1.0 µl
- BSA (1mg/ml) 0.5 µl
- MseI adaptors (50mM) 0.5 µl
- EcoRI adaptors (5mM) 0.5 µl
- Msel enzyme (10,000U/mL) 0.1 µl
- EcoRI enzyme (20,000U/mL)  0.15 µl
- T4 DNA ligase (400,000U/mL)  0.05 µl
- distilled water  5.95 µl

All reagents were mixed and briefly centrifuged, then 10 µl of master mix and 5 µl of genomic DNA were added to each respective tube, then incubated for 2 hours at 37˚C and held at 4˚C. Gel electrophoresis was performed on a subset of samples to ensure the reactions were successful and was performed using 2g of agarose, 100mL of TAE, and 2µl of ethidium bromide. Gels were run with 4µl of 100bp ladder and 4µl of diluted product in loading buffer for approximately 45 minutes at 120V and maximum amperage. Products were then diluted 1:10 with autoclaved water in preparation for pre-selective PCR.

**Figure 7:** Image from a set of successful digestion-ligation reactions. The two outer bands are the 100 bp ladders used for reference. Inside lanes show significant smearing which indicates that there are many DNA fragments present in each sample.
Preselective PCR

The next step in this process is preselective PCR. This reaction uses primers that match to the adaptor sequences and have an additional selective base, which serves to reduce the total number of amplicons from the genome. In this step the primers are MseI-A, MseI-C, Eco-A, and Eco-C, with the following sequences, respectively: 5’-GATGAGTCCTGAGTAAA-3’, 5’-GATGAGTCCTGAGTAAC-3’, 5’-GACTGCGTACCAATTCA-3’ and 5’-GACTGCGTACCAATTCC-3’.

For preselective reactions, we prepared four master mixes using:

- Thermopol I buffer 10X 1.5 µl
- dNTP (2.5 mM) 0.75 µl
- Primer 1 (10 mM) 0.4 µl
- Primer 2 (10 mM) 0.4 µl
- Taq (5000U/mL) 0.075 µl
- distilled water 9.4 µl

The volumes are listed per reaction. All reagents were mixed and briefly centrifuged, then 12.5 µl of master mix and 2.5µl of diluted digestion ligation product were added to each respective tube. The four preselective primers were used in all four possible combinations (preselective primer set A (PS-A): MseI-A and Eco-A, PS-B: MseI-A and Eco-C, PS-C: MseI-C and Eco-A, and PS-D: MseI-C and Eco-C). Samples were run in the thermocycler program as follows: 95 °C for 2 min, followed by 40 cycles of 95 °C for 30 sec, 48 °C for 30 sec, and 72 °C for 1min, followed by a terminal step of 72 °C for 10 min. Samples were then held at 4 °C, and then were
moved to -20 °C. Gel electrophoresis was performed to ensure the reactions were successful and was performed using 2g of agarose, 100mL of TAE, and 2µl of ethidium bromide. Gels were run with 4µl of 100bp ladder and 4µl of diluted product in loading buffer for approximately 45 minutes at 120V and maximum amperage. Products were then diluted 1:10 with autoclaved water in preparation for selective PCR.

Selective PCR

This is the last step in the process and uses the diluted products from preselective PCR as a template. The primers used have 3 selective bases, and one primer in each pair is labeled with a fluorescent dye that can be detected by the automated sequencer. The primers used in this step are Mse-ATC, Mse-CTC, Eco-ACG (6-FAM label), Eco-CAG (VIC label), and Eco-CTC (NED label). The sequences of these primers are as follows, respectively: 5’-GATGAGTCTGCTACTAATC-3’, 5’-GATGAGTCTGCTACTAATC-3’, 5’-6FAM-ACTGCGTACCAATCCAG-3’, 5’-VIC-ACTGCGTACCAATCCAG-3’ and 5’-NED-ACTGCGTACCAATCCAG-3’. In this step, the reactions were multiplexed, which means that each reaction used multiple primer sets. This is important because it allows us to generate a more complete AFLP profile while minimizing time and cost per crayfish. Production of amplicons of different sizes and of varying sequences that are specific to certain DNA sequences. Ultimately, this allows us to gain more information on the genomes of each individual crayfish.

Two master mixes are prepared using:

- Thermopol I buffer 10X 1.5 µl
- dNTP 0.75 µl
- Mse primer (10mM) 0.6 µl
- Eco-ACG primer (10mM) 0.2 µl
- Eco-CAG primer (10mM) 0.2 µl
- Eco-CTC primer (10mM) 0.2 µl
- Taq (5,000U/mL) 0.075 µl
- Distilled water 9.0 µl

The volumes are listed per reaction. All reagents were mixed and briefly centrifuged, then 12.5 µl of master mix and 2.6 µl of diluted preselective products were added to each respective tube. Diluted preselective products were combined as 1.3 µl of pre-selective primer set A. and 1.3 µl of pre-selective primer set B or as 1.3 µl of pre-selective primer set C and 1.3 µl of pre-selective primer set D. The master mixes were made with primer combination as follows: selective primer combination A (SE-A) included MseI-ATC and all three Eco primers and selective primer combination B (SE-B) included MseI-CTC and all three Eco primers. Samples were run in the thermocycler program as follows: 95 °C for 2 min, followed by 40 cycles of 95 °C for 30 sec, 48 °C for 30 sec, and 72 °C for 1 min, followed by a terminal step of 72 °C for 10 min. Samples were then held at 4 °C, and then were moved to -20 °C. Gel electrophoresis was performed to ensure the reactions were successful and was performed using 2g of agarose, 100mL of TAE, and 2µl of ethidium bromide. Gels were run with 4µl of 100bp ladder and 4µl of diluted product in loading buffer for approximately 45 minutes at 120V and maximum amperage. Products were then diluted 1:10 with autoclaved water in preparation for sample preparation for shipment.

**Sample Preparation**

This was the final step in preparing the samples for shipment to the facilities at either Cornell University or Yale University for fragment analysis. In order to prepare the samples for
processing, 0.4 µl of LIZ 600 size standard and 13.1 µl of HiDi formamide was added to 1.5 µl of each SE-PCR product. Plates were then labeled, packaged and shipped to a facility, and an online submission form was completed.

2.3 Fragment Analysis

Fragment analysis was performed using the program GENEMAPPER v4.0 by Applied Biosystems. Preliminary analysis was done to determine the success of our AFLP PCR reactions. First, we evaluated the size standard by hand and made any necessary adjustments to allow the automated analysis to proceed. We then examined the peak heights of the 3 different dyes to ensure the samples met our minimum requirements. After examining all the samples, we decided on a threshold of 200 florescence units as a standard for judgment. This threshold was used because after collectively examining all the samples, all the clear, defined peaks appeared to be above 200 florescence units, and anything below was considered to be background fluorescence. Any sample containing all peaks below 200 units were discarded and the AFLP protocol was again performed on these samples beginning with the digestion-ligation step.

Once we reduced the total data set to include only those samples that met our minimum requirements for inclusion, we began more in-depth analysis. Using GENEMAPPER we created our own analysis method which we named “Crayfish AFLP”. This analysis method started at 50bp and ran to 500bp. We only analyzed peaks over 200 fluorescence units but flagged reoccurring peaks from 50-200 fluorescence units to be analyzed by hand to ensure we did not miss any important peaks that may not have fully fluoresced. A panel was generated using the samples, and had a bin width of 1bp. Also, it deleted common alleles to allow us to focus on loci of interest, which are the ones with variability among and within populations. We analyzed the data in 6 batches, each reflecting one SE primer pair. First, we separated samples into SE-A and
SE-B groups, and then we analyzed the samples separately by each dye present (6FAM, VIC and NED). This analysis was done in batches because the different dyes produced various alleles at differing lengths of base pairs. After this analysis was completed, we were able to form a master list of every sample, and included information such as location and presence or absence data for each allele.

2.4 Allele Selection

Once presence/absence data were obtained for each allele in each individual, ratios of the presence of an allele to the total number of individuals for that allele were calculated; ratios of the presence of an allele to the number of individuals in a known population were also calculated. Alleles that were exhibited in less than 5% of the population were discarded because they are not common enough to provide diagnostic information. A chi-squared test was performed on the remaining alleles in the M2 and R1 populations in order to identify loci that showed significant associations with one of two evolutionary lineages that we expected to be present in our collection. Specifically, previous research had revealed that central Massachusetts has at least two lineages in the O. virilis species complex (Mathews et al., 2008). We were interested in excluding any alleles that showed signs of being diagnostic for either species so that we could examine only the remaining loci that showed non-lineage specific variation. Another project focused on the species-specific markers in order to investigate hybridization in the collection (Buchanan et al., 2012). The chi squared test was only performed on M2 and R1 crayfish because we assume the R1 population to be nearly pure O. virilis, and the M2 population to be nearly pure O. quinebaugensis. This test indicates the likelihood that these two populations share alleles by chance, or whether there is some statistical correlation with one or the other lineage. A 95% confidence was used for each allele in the dataset for the null
hypothesis that the allele was not significantly associated with one or the other lineage. From this, we selected the loci that failed to reject the null hypothesis. Using this set of 150 loci, we treated the collection of 418 crayfish as if they were members of a single interbreeding lineage; this allowed us to gain preliminary insight into the population genetic structure of the Blackstone River Valley.

2.5 Analysis of Population Genetics

In order to analyze our data and draw conclusions about the population genetics of crayfish in the Blackstone River Valley, several bioinformatics programs were used. This software uses algorithms to statistically sort and cluster our data into likely population models. The three programs used were Structure, CLUMPP and Distruct.

Structure

In order to determine how many distinct gene pools of crayfish can be found in the Blackstone River Valley, and to gain insight into the geographic distribution of genetic differentiation, we performed genetic analysis using the program Structure (Pritchard et al. 2000). This program uses a model-based clustering method to deduce possible population structure based on genotypic data. By analyzing our data in this program we were able to determine the probability of membership for each different gene pool. Because this program does not know where each genotype originated from, we should be able to use the information Structure outputs along with our knowledge of the environment to logically determine which organisms fall into each gene pool in a specific area. Overall, we were then able to determine if each site has a pure population (contains one species), has two pure populations (contains two species), has a hybridized population, or any combination these possibilities.
Data was entered into Structure in binary form. Since AFLP markers cannot distinguish between heterozygotes and dominant homozygotes, presence data was entered as 1, -9 and absence data was entered as 0,0. In this program -9 represents a missing data value. A project was created using the following parameters: a burnin period of 50,000 with 100,000 reps after burnin. An admixture model was used, and allele frequencies were assumed to be correlated. The program was run several times, with K values (assumed number of populations) ranging from 2-25. The output from this program gave the value of the estimated Ln probability of the data for each K from Structure was used. Structure determines this value by calculating the log likelihood of the data at each Markov chain Monte Carlo (MCMC), a parameter set at the beginning of the analysis that correlates with the burn-in and number of repetitions performed. These values were then averaged and half of their variance was subtracted (Evanno et al., 2004). The method applied in Evanno et al. (2004) was used to estimate K, because this method was shown to be most accurate, as opposed to the Wilcoxon test, which is a non-parametric statistical test used to compare two related samples. This new method uses the calculation of ΔK, the second order rate change of L(K), to form a plot that peaks at the true value of K (Equation 1).

\[\Delta K = m(|L''(K)|)/s[L(K)]\]

**Equation 1:** This is the equation used to calculate the ΔK that is plotted for each K in order to determine the most likely number of true populations present. It is the second order rate change of Ln K, the log likelihood of the probability of the data. For this calculation we used data from 5 runs of Structure.

**CLUMPP and Distruct**

CLUMPP (Jakobsson and Rosenberg, 2007) is a program used to align the replicate runs of Structure and stands for ‘cluster matching and permutation program’. Although Structure is an effective program for clustering individuals into populations, there is often variation seen in the results from run to run. This variation is considered to be caused by the ‘unsupervised’ nature of
the program, which means that the labels are not known at the start of the analysis so each
permutation may come out differently (Jakobsson and Rosenberg, 2007). This can have different
effects on the data set and can cause problems such as label switching and multimodality. These
are complex variations that can be observed over multiple runs of Structure which make it more
difficult to analyze the data, and that can be reduced by algorithms in CLUMPP. We employed
CLUMPP to align the 5 replicate runs of Structure for our highest ΔK value. Once we obtained
the output file from CLUMPP, Distruct was used to form a graphical representation of these
results. Distruct is a program that creates a plot of the clusters aligned from CLUMPP, and is
simply used for visualization of the data to make interpretation of population structure easier
(Rosenburg, 2004). The graph produced shows each individual grouped by each site, and
displays by color the membership coefficient for each individual. The membership coefficient is
the percentage of an individual’s genome that is assumed to belong to a particular cluster.
3. Results and Discussion

Our AFLP development protocol yielded a total of 150 loci that showed non-lineage-specific variability in the collection of 418 individuals across 23 sites. Using these variable loci, we determined the most likely true number of gene pools in our collection. We then analyzed these data with respect to collection site to visualize the genetic differentiation in the Blackstone River Valley. This allowed us to draw conclusions about the speciation and admixture occurring within this expansive area.

3.1 Estimating the Number of Populations

As seen below in Figure 8 and 9, a K of 2 was determined.

![Figure 8: Plot of ΔK vs. K, obtained from Formula 1. The large peak at K=2 indicates that there are 2 distinct populations that can be observed in the Blackstone River Valley.](image)
This data shows that there are 2 distinct gene pools found in the Blackstone River Valley. As seen in Figure 9, there is a clear dominance of the population represented by the red. Also, it is clear that admixture is occurring, because several individuals exhibit alleles from both populations (green and red alleles).

### 3.2 Aligning the Populations

In order to align the outcomes from the 5 replicate runs of Structure, the program CLUMPP was used. CLUMPP generated a Q-matrix for each individual, called the membership coefficient, which is the percentage of an individual’s genome that is assumed to belong to a particular cluster. CLUMPP outputs this file which can then be visualized using the program Distruct. Distruct organizes the Q-matrix and cluster data for visual representation. The output of that analysis is shown in Figure 10.
As indicated in Figures 9 and 10, our data showed two distinct populations in the Blackstone River Valley. Previous investigation of this region indicated that there are at least two sister species that are extremely similar in morphology, *O. quinebaugensis* and *O. virilis* (McMurrough and Saltzman, 2009). These previous investigations indicated that the two species have pure or nearly pure populations at sites M2 (*O.q.*) and R1 (*O.v.*). However, as seen in Figure 10, in our dataset, M2 and R1 are nearly indistinguishable. This was expected because the AFLP markers that we chose for analysis were not species-specific, but instead loci that were common in both species. Thus, our data show the presence of 2 previously undetected gene pools, and that may represent a third cryptic species in the collection that had not previously been detected by other means. An alternative explanation is that our data represents strong population subdivision due to geographical barriers separating the population into two sections that do not exhibit much gene flow.

In order to visualize the genetic diversity within the Blackstone River Valley, we show the data displayed in Figure 10 in the form of histograms that display the percentage of alleles among each site. These histograms were placed on the map of the Blackstone River Valley according to each site (Figure 11). This allowed us to envision which sites show similar patterns, and whether or not these sites are geographically close.
From the map in Figure 11 we were able to perform an informal test of the hypothesis that the 2 gene pools are being separated by geographic barriers. Some nearby sites show clear relationships, such as M91 and M93, which is consistent because M91 is downstream of M93, both lotic environments. Sites M103, M90 and R1 also exhibit similarity and are geographically close. Several sites exhibit admixture, such as M4, M34, M55, M70, M96 and M98. Overall there appears to be no clear pattern of dispersal and genetic differentiation, but further analysis
combining our dataset with other sources of information may yield stronger insights. We determined that there is no clear geographic structure detectable in the Blackstone River Valley from our AFLP dataset that would support the hypothesis of separation by geographic barriers. This may suggest that the data supports the hypothesis of cryptic species present in the Blackstone River Valley. This discovery is not uncommon in the field of evolution and landscape ecology. Many empirical studies have been performed that sought to identify the biodiversity among related freshwater species in a particular area, that end up discovering another cryptic lineage that was previously unknown. This can be seen in a study by Martin and Bermingham (2000) that analyzes the molecular and morphological data from a species of catfish. They conclude that the biodiversity of *P. chagresi* catfish is more complex than previously thought, and the many cryptic lineages are likely due to the rich history of this species paired with the geographical structure. One study in particular that correlates to this investigation was performed by Mathews et al. (2008). This study sought to genetically and morphologically investigate the *O. virilis* species complex over parts of its geographic range. Mitochondrial and nuclear genetic data was used and found that many cryptic lineages exist, although the context of this hidden biodiversity is not completely understood. This study would support our hypothesis that there is a cryptic lineage in our collection, because there are still many lineages that have yet to be discovered.

Cryptic lineages can be attributed to a variety of factors such as temporary geographic isolation, historic changes in geography or secondary contact. It is difficult to distinguish between native and introduced populations, but genetic data can provide insight. When populations are introduced, they often exhibit a founder effect with low genetic diversity (Suarez et al, 1999),
while native populations display a higher level of gene flow (Roman, 2006). There is evidence from our data that this cryptic species was not introduced by humans, but is likely endemic and evolved right in the Blackstone River Valley. This is because many sites display admixture, and have several alleles in common with the two previously established populations *O. virilis* and *O. quinebaugensis*. However, it is also important to note the differences in some alleles between this cryptic lineage and the known species because it may suggest that this species represents a genetic intermediate (Daniels et al, 2003). As mentioned earlier, there are no clear geographic barriers that appear to be isolating sites and blocking gene flow, which suggests that the genetic differentiation of crayfish in the Blackstone River Valley is likely due to historic changes in geography. New England has a rich geographic history and the geography could be changing often due to human interference as well as natural events. For example, dams have been built in some areas by humans to control water flow. Natural events such as the extensive cycles of glacial and interglacial periods in recent geological history have had noteworthy effects on North America’s biogeography. Glacial cycles can affect aquatic systems by changing the temperature, which in turn affects water levels, sedimentation, and survival of species (Albert and Reis, 2011).

### 3.3 Conclusions

Our research supports the idea that there may be a previously unknown cryptic lineage in the *O. virilis* species complex of the Blackstone River Valley. However, there are many possibilities as to what could be occurring and causing the genetic differentiation seen in the region, including historic changes in geography and human interference. In addition, if there is a third lineage present, its history in the area is unknown, and it may represent either another endemic lineage or a lineage introduced at some point in time from elsewhere in the range of *O. virilis*. 
Overall, we believe that further research is needed in order to obtain a better understanding of the population genetics of *O. virilis*, as well as other species, in the Blackstone River Valley. We believe it may be of interest to compare our AFLP data with the other genetic markers found in previous projects, such as the mitochondrial data reported in McMurrough et al. (2008). Such a comparison might help resolve what is causing this pattern of genetic differentiation, as well as if this cryptic lineage is an invasive or endemic species.
References


