ABSTRACT

WILLIAMS, LARISSA MORIARTY. Signatures of Selection in Natural Populations Adapted to Chronic Pollution. (Under the direction of Dr. Damian Shea and Dr. Marjorie F. Oleksiak).

Understanding the genetic basis of ecologically important traits and how such variation translates into functional phenotypes is a central goal of ecological and evolutionary genetics. For this thesis, the genetic basis of resistance to pollution was explored in populations of the estuarine minnow, *Fundulus heteroclitus*, which lives in extremely polluted estuaries along the east coast of the United States of America. Using molecular and genetic techniques, genetic variation within and among several *F. heteroclitus* populations exposed to environmental contamination was determined in order to identify signatures of natural selection.

In the first study, the genetic technique called amplified fragment length polymorphisms (AFLP) was used to generate hundreds of genome-wide markers. Then statistical tests were used to determine whether any of the genome-wide markers were under selection in three polluted *F. heteroclitus* populations compared to each of their reference populations. In total, 24 loci were determined to have outlying FST values in polluted populations. Several of these loci were outliers in two different, geographically separated populations, indicating a shared evolutionary response to contaminant exposure.

In the second study, high-throughput sequencing and genotyping technologies were used to determine genome-wide single nucleotide markers in order to assess levels of polymorphism along the genome. SNPs were polymorphic and showed latitudinal, clinal variation separating northern and southern *F. heteroclitus* populations. These markers also were used to differentiate *F. heteroclitus* from several other *Fundulus* species: *majalis*,...
and similis. This study established a technique to quickly and accurately scan the genome for polymorphisms for use in population genetics.

The third study analyzed the SNP data collected from the high-throughput sequencing and genotyping with the goal of identifying SNPs which were under selection in polluted populations. In contrast to the AFLP study, the nucleotide polymorphisms as well as surrounding nucleotide sequences were known which allowed additional statistical tests for neutrality. In total, one to four percent of the SNPs studied showed signatures of selection in any one polluted population. One SNP in the proximal promoter of the xenobiotic metabolizing enzyme, CYP1A, was identified as being under selection in all three polluted populations that were studied.

In order to determine the functional significance of the CYP1A SNPs under selection in polluted populations, the fourth study focused on determining genetic variation in the CYP1A promoter in one polluted and two reference populations and functionally characterizing the inducibility of the promoter in vitro. Overall, the nucleotide variability of the CYP1A promoter was high in all populations. There was also no significant, selective signature on the promoter as a whole, or any one portion (functional or non-functional) of the promoter suggesting that it is evolving through neutral processes in all three populations. The CYP1A promoter constructs were shown to be inducible in a dose-dependent manner in vitro, and induction was significantly higher in the polluted New Bedford Harbor population as compared to both reference populations. This result is surprising because previous studies have shown that CYP1A is refractory to induction by prototypic inducers in the New Bedford Harbor population.
Overall, this research highlights the role of natural selection due to exposure to anthropogenic contamination in shaping variation in natural populations of *F. heteroclitus*. This research was the first to complete several genome-wide scans for polymorphisms in this species and was also the first to describe signatures of selection in independent *Fundulus* populations exposed to contamination. It also established that the genetic basis for the refractory CYP1A transcriptional phenotype in polluted *F. heteroclitus* populations is not solely due to the CYP1A promoter. Ultimately, this work advanced our knowledge of how natural populations adapt to changing environmental conditions.
Signatures of Selection in Natural Populations Adapted to Chronic Pollution

by
Larissa Moriarty Williams

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Toxicology

Raleigh, North Carolina
2010

APPROVED BY:

_______________________________  _______________________________
Damian Shea                Marjorie F. Oleksiak
Committee Chair            Co-Committee Chair

_______________________________  _______________________________
Robert C. Smart          W. Owen McMillan

_______________________________
Patricia McClellan-Green
DEDICATION

I dedicate this work to my parents as well as the many strong women who have supported and guided me through the years.
BIOGRAPHY

Larissa Moriarty Williams was born May 13, 1983 in Ithaca, New York. She was welcomed by her parents David J. Williams and Karen E. Andrésen. Her family later expanded to include her sister Talia A. Williams. Growing up, Larissa was an avid horseback rider, spending endless hours at the barn with her riding instructor Barbara Watts. Through riding, she was introduced to Dr. Jacqueline Sinclair of Dartmouth University, who offered her a research internship in her toxicology laboratory following her first year at Smith College. At Smith College, located in Northampton, Massachusetts, she majored in Biological Sciences with a minor in Marine Science. It was in college that Larissa became an avid fan of marine biology and, at the suggestion of Larissa’s favorite Smith professor, Dr. Paulette Peckol, attended a summer program at the Duke University Marine Laboratory (DUML) during her second summer in college where she took classes and completed an independent research project with Dr. Daniel Rittschof. The following summer Larissa worked with Dr. Patricia McClellan-Green at DUML on her summer project, funded by the Bookhout Fellowship Larissa was awarded, which ultimately evolved into her senior honors thesis. Larissa graduated in 2005 from Smith College with Highest Honors in the Biological Sciences. She went on to join the Ph.D. program at North Carolina State University in Environmental Toxicology under the direction of Dr. Margie Oleksiak. Two years into her Ph.D. program she moved to Miami, Florida to join Dr. Oleksiak at her new institution, the University of Miami’s Rosenstiel School. Larissa’s time in North Carolina and Miami has been fun, fruitful, and memorable.
ACKNOWLEDGMENTS

I would first like to thank my advisor, Dr. Margie Oleksiak, for her invaluable guidance, inspiration, and support through the Ph.D. process. Margie works harder than anyone I know, and I admire her extraordinary talents as a researcher. Margie is also one of the most patient advisors I have ever known, a quality that will benefit her and her future students to come.

I would also like to thank my committee at North Carolina State University, which allowed me to move to Miami and have been extremely supportive during my entire Ph.D. I could not have achieved this degree without them. I would especially like to thank Dr. Damian Shea who made it administratively and financially feasible for me to remain at NCSU and Dr. Patricia McClellan-Green for her endless support.

To Goran: Thanks for all your support, patience, laughs and guidance through this whole process. You are an incredible human being whom I admire, and I am a better person because of you.

Lastly, without the love and support of Michael, my family, and my friends, this thesis would not have been possible. Thanks for always being there for me; I am eternally grateful.
TABLE OF CONTENTS

LIST OF TABLES ........................................................................................................... vii
LIST OF FIGURES ......................................................................................................... viii
INTRODUCTION ........................................................................................................... 1
  References ................................................................................................................. 15

CHAPTER ONE: SIGNATURES OF SELECTION IN NATURAL POPULATIONS ADAPTED TO CHRONIC POLLUTIONS ................................................................. 34
  Abstract ...................................................................................................................... 35
  Background ................................................................................................................. 37
  Methods ...................................................................................................................... 40
  Results ......................................................................................................................... 45
  Discussion ................................................................................................................. 49
  Conclusions ................................................................................................................. 55
  Acknowledgements ..................................................................................................... 56
  References ................................................................................................................. 57

CHAPTER TWO: SNP IDENTIFICATION, VERIFICATION, AND UTILITY FOR POPULATION GENETICS IN A NON-MODEL SPECIES .................................................. 78
  Abstract ...................................................................................................................... 79
  Background ................................................................................................................. 80
  Methods ...................................................................................................................... 82
  Results ......................................................................................................................... 90
  Discussion ................................................................................................................. 95
  Conclusions ............................................................................................................. 101
  Acknowledgements .................................................................................................. 102
  References ............................................................................................................. 103

CHAPTER THREE: ECOLOGICALLY AND EVOLUTIONARILY IMPORTANT SNPS IDENTIFIED IN NATURAL POPULATIONS SPECIES ..................................... 122
  Abstract .................................................................................................................... 123
  Introduction .............................................................................................................. 124
  Materials and Methods .......................................................................................... 125
  Results and Discussion .......................................................................................... 127
CHAPTER FOUR: CYTOCHROME P4501A PROMOTER POLYMORPHISMS AND ACTIVITY IN NATURAL POPULATIONS ................................................................. 149

Abstract ............................................................................................................. 150

Introduction ........................................................................................................... 151

Materials and Methods .......................................................................................... 155

Results .................................................................................................................... 159

Discussion .............................................................................................................. 166

Acknowledgements ............................................................................................... 175

References ............................................................................................................. 176

SUMMARY AND DISCUSSION .............................................................................. 218

References ............................................................................................................. 232
LIST OF TABLES

CHAPTER ONE

Table 1. Sample locations of *F. heteroclitus* along the east coast of the United States.69
Table 2. Primer sequences used in AFLP analyses.70
Table 3. Outlier loci shared among the Superfund site Fundulus populations.71
Table 4. Pairwise $F_{ST}$ values with and without outlier loci.72

CHAPTER TWO

Table 1. Adapters and primers used in the amplification of genomic DNA.113
Table 2. Genotyping success of SNP markers using the MassARRAY multiplex assay.114
Table 3. Genetic Parameters of sampled populations in two species of *Fundulus*.115
Table 4. SNP minor allele frequencies (MAF) within *F. heteroclitus* and *F. grandis* populations.116

CHAPTER THREE

Table 1. Genes and loci with SNPs identified as outliers using the FST modeling approach, association test, and Minor Allele Frequency-$F_{\text{MAX}}$ test (MAF-$F_{\text{MAX}}$) 138

CHAPTER FOUR

Table 1. Measures of population divergence between *F. heteroclitus* populations.186
Table 2. Neutrality test for the pattern of sequence variation.187
Table S1. Predicted transcription factor binding sites along the CYP1A promoter.202
LIST OF FIGURES

CHAPTER ONE

Figure 1. Sample locations of *F. heteroclitus* along the east coast of the United States. Circles represent reference sites and stars are Superfund sites ..........................73

Figure 2. $F_{ST}$ values estimated from approximately 300 variable AFLP loci plotted against mean allele frequency. The solid line represents the 0.99 quantile estimated from a simulation model for each comparison. Loci shared among the same Superfund site are labeled with their primer set (letter) and number. Loci shared between Superfund sites are starred. § Shared loci included in these points are: A2, A19, A34, A56, D87, E118, E127, E137, E150, E156, C186, C194, C205, and C252. E118 also is shared between New Bedford Harbor and Elizabeth River populations ........................................................................74

Figure 3. Outlier loci in comparisons of each Superfund populations to both its clean reference sites; numbers in the unions of circles represent outlier loci shared among populations. A) New Bedford Harbor, MA, Sandwich, MA and Pt. Judith, RI comparison. B) Newark Bay, NJ, Clinton, CT, and Tuckerton, NJ comparison. C) Elizabeth River, VA, Magotha, VA and Manteo, NC comparison. ..........................................................................................................................75

Figure 4. Shared outlier loci among Superfund population comparisons to both clean reference sites; numbers in the unions of circles represent outlier loci shared between two Superfund populations. ........................................................................................................76

Figure 5. Relationship between genetic distance and geographic distance. Genetic distance was calculated from the mean $F_{ST}$ for each pair of populations with (A) and without (B) outlier loci. Circles represent a pairwise comparison of a Superfund *versus* a reference site, squares represent a Superfund *versus* a Superfund site comparison, and crosses represent a reference *versus* a reference site comparison............................................................................................................77

CHAPTER TWO

Figure 1. Sampling sites for *Fundulus* species. *F. heteroclitus* was collected along the east coast of the United States and *F. grandis* was collected along the Gulf of Mexico coast ........................................................................................................117

Figure 2. Design of 454 pyrosequencing contig generated from the digestion of genomic DNA with restriction enzymes (EcoRI and BspEI), the addition of restriction site specific linkers, an individual barcode and a 454 amplicon adapter ....118
Figure 3. Contig totals \textit{versus} number of reads per contig amongst those contigs with identified SNPs (bars) and all contigs (squares). ................................119

Figure 4. Non-amplified and non-polymorphic loci among \textit{Fundulus} species. (A) Numbers of loci, which did not amplify with the MassARRAY platform among the four species of \textit{Fundulus}. Not shown: loci shared between \textit{F. majalis} and \textit{F. similis} (8) and \textit{F. heteroclitus} and \textit{F. grandis} (12). (B) Numbers of loci, which were not polymorphic among the four species. Not shown: loci shared between \textit{F. majalis} and \textit{F. similis} (9) and \textit{F. heteroclitus} and \textit{F. grandis} (1) .................................................................120

Figure 5. Population structure as assessed by STRUCTURE. Bar plot was generated by DISTRUCT and depicts the classifications of the populations with the highest probability under the model. \textit{K} indicates the number of clusters that maximized the probability of the model. Each individual is shown as a vertical bar. (B) Principal components PC1, PC2 and PC3 from all SNPs (as calculated in JMP Genomics 3.2) among all individuals. Species are separated from each other as well as northern and southern \textit{F. heteroclitus} populations. Colors represent different species. (C) Principal components PC1, PC2, and PC3 from all SNPs among \textit{F. heteroclitus} individuals. Colors represent different populations .................................................................121

\textbf{CHAPTER THREE}

Figure 1. Polluted sites are starred and flanked north and south by clean reference sites (circles) to form a triad. Venn diagrams indicate the number of SNPs exhibiting non-neutral behavior using the three statistical tests: the \textit{F}_{ST} modeling approach (\textit{F}_{ST}), Association (Assoc.), and MAF-\textit{F}_{MAX} (MAF). .....143

Figure 2. Empirical \textit{F}_{ST} values are plotted against heterozygosity. The line demarks the 99\textsuperscript{th} percentile estimated from a simulation model. Blue diamonds indicate SNPs that are significantly different between the polluted population and both reference populations but not different for reference \textit{versus} reference. Red dots are superimposed on blue diamonds if the SNP was also significant in the other two statistical tests. Less interesting are the crosses and open diamonds. Black crosses are outliers also in the reference \textit{versus} reference comparison. Open diamonds represent outliers where the polluted population was only significant in comparison to one reference population.................................144
Figure 3. Association test for detection of selection. Likelihood of association of each SNP with either the polluted site (red line) or reference sites (blue line) as a \( \log_{10} p \)-value. The \(- \log_{10} p\)-value of 2 is marked by a black line, and the Bonferroni correction for multiple testing is marked by the dotted grey line (\( \log_{10} p \)-value of 4.56). SNPs are identified as outliers in polluted sites versus reference sites if the polluted association value is greater than 2 and the likelihood ratio test \( p \)-value of polluted versus reference association is \( \leq 0.01 \) (Table 1). Black squares indicate those SNPs where the likelihood model for pollution significantly exceeds the model based on divergence among reference sites. Red dots are superimposed on black squares if the SNP was also significant in the other two statistical tests.

Figure 4. MAF-\( F_{\text{MAX}} \) test for detection of differences in SNP allele frequencies between polluted and reference sites. The allele frequency of the triad-wide minor allele was calculated and plotted for all SNPs. Columns are collection sites arranged north to south, and each row represents an individual SNP. SNPs with allele frequencies significantly different in an ANOVA using \( F_{\text{MAX}} \) to control for type I errors among iterations (\( F_{\text{MAX}} \): empirical \( F \)-value exceeds the top 1% of all permutated \( F \)-values assuming random population differentiation) between polluted and both reference sites are plotted. Red dots denote SNPs exhibiting non-neutral behavior in all three statistical tests. The SNP exhibiting non-neutral behavior in all three triads and using all tests (CYP1A +268) is boxed.

CHAPTER FOUR

Figure 1. \( F. \) heteroclitus collection sites. Superfund site (New Bedford Harbor, MA) is denoted by a star, and reference sites north and south denoted by a circle.

Figure 2. 221 parsimony informative sites among and between \( F. \) heteroclitus populations and between \( F. \) heteroclitus and \( F. \) grandis in the sequenced portion of the CYP1A promoter, exon and intron 1. Within the promoter region (upstream of basepair 1630), there are 157 parsimony informative sites. There are no fixed differences between \( F. \) heteroclitus populations. 20 fixed differences between \( F. \) heteroclitus and \( F. \) grandis are marked by a star. SNPs (929 and 1892bp) found to be under selection in Williams and Oleksiak (2010) are starred over the nucleotide number.

Figure 3. Linkage disequilibrium (LD) among polymorphic nucleotides along the proximal promoter and first exon and intron of CYP1A for Sandwich. Green indicates no significant LD, red indicates significant LD (\( p \leq 0.05 \)), and the blue line indicates the diagonal.

Figure 4. Linkage disequilibrium (LD) among polymorphic nucleotides along the proximal promoter and first exon and intron of CYP1A for New Bedford Harbor. Green indicates no significant LD, red indicates significant LD (\( p \leq 0.05 \)), and the blue line indicates the diagonal.
Figure 5. Linkage disequilibrium (LD) among polymorphic nucleotides along the proximal promoter and first exon and intron of CYP1A for Point Judith. Green indicates no significant LD, red indicates significant LD (p≤0.05), and the blue line indicates the diagonal.

Figure 6. Sliding window comparisons of average nucleotide substitutions per site within *F. heteroclitus* populations (Pi) and between populations (Dxy). Plots for the sliding window use 50bp-wide window and 10-bp step size.

Figure 7. Tajima’s D calculations (panels A, C, E) and Fu and Li’s D calculations (panels B, D, and F) using a sliding window with a 50bp-wide window and a 10-bp step size for each *F. heteroclitus* population.

Figure 8. Phylogenetic analyses of the CYP1A proximal promoter using PAUP*4.0.
A. Phylogenetic tree using all nucleotides of the proximal promoter. B. Phylogenetic tree using non-functional regions of the proximal promoter. C. Phylogenetic tree using functional regions of the proximal promoter. Bootstrap values (N=500) are listed. Probabilities are maximum parsimony values.

Figure 9. Induction of luciferase activity from *F. heteroclitus* CYP1A reporter gene constructs from a New Bedford and Sandwich individual by 3-MC in PLHC-1 cells plotted on a non-logarithmic (A) and logarithmic scale (B). A star indicates a significant fold induction over control (p<0.05).

Figure 10. Average fold induction of the CYP1A promoter over control for three populations (N=4) at a 1μM dose of 3-MC. A one-way ANOVA found a significant difference in the average fold induction between populations (p<0.0001) and a Tukey’s HSD post-hoc test determined that New Bedford Harbor was significantly different from Sandwich (p<0.0001) and Point Judith (p<0.0001), but Sandwich and Point Judith were not significantly different from each other (p=0.258). Significant induction in the New Bedford Harbor population over the reference populations is denoted by a star.

Figure S1. Full length proximal promoter and exon 1 and intron 1. Star indicates conservation in the nucleotide across all individuals, P indicates a parsimony informative site, and F indicates a fixed difference between *F. heteroclitus* and *F. grandis*. GRE and XRE binding sites are noted, as well as the TATA box, start of transcription, and intron and exonic boundaries. Big stars on nucleotides 929 and 1892 indicate SNPs under selection as described in Williams and Oleksiak (2010).
INTRODUCTION

Determining the molecular basis of adaptation is fundamental to understanding evolution. Adaptive evolution proceeds primarily through the process of natural selection whereby individuals with phenotypes that enhance their survival and reproductive rate will pass on heritable genetic characteristics at a higher rate in comparison to other individuals in a population. In successive generations, the genetic basis for traits associated with greater fitness in a particular environment can become more common in a population. It has long been a goal of biologists to identify the basis of adaptation to understand how organisms have adapted to a wide range of environments as well as assess the potential of populations to respond to environmental change. A powerful approach to study adaptation utilizes population genomics to study genome-wide allele frequency distribution and change under the influence of natural selection, genetic drift, mutation and gene flow. Population genomics characterizes molecular variation within and between populations and can be used to identify signatures of selection in populations.

The teleost fish, *Fundulus heteroclitus* has adapted to a diverse assortment of environments. They inhabit estuaries along the east coast of North America where they experience daily fluctuations in temperature, salinity and oxygen. In the last 150 years or so, portions of the east coast of the United States became industrialized and subsequently polluted with a variety of contaminants including persistent organic pollutants, metals, and pesticides, among others. Populations of *F. heteroclitus* have adapted to sites contaminated with these persistent toxicants. While the basis of adaptation for this species has been
studied intensely for over 20 years, the exact mechanism(s) by which *Fundulus heteroclitus* has been able to survive and thrive in such harsh environments is not yet known. In order to start understanding this phenomenon, my thesis has involved the use of population genetics to study the genetic basis of adaptation across three independent, polluted populations of *F. heteroclitus*.

**The common mummichog: *Fundulus heteroclitus***

*Fundulus heteroclitus* is a small minnow, which resides along the east coast of the United States in tidal marshes from the Gulf of St. Lawrence to northeastern Florida. *F. heteroclitus* are mainly carnivorous although algae and detritus are often found in their gut contents (Allen et al., 1994). Due to their significant detrital dietary component and their coastal habitat, they are exposed to a wide variety of environmental contaminants from the water and sediments. With a small home range of only a few hundred yards from its natal creek (Lotrich, 1975), these fish are indicative of local conditions. Nevertheless, migration of *Fundulus* is large enough to minimize genetic drift but small enough that natural selection can still occur (Brown, Chapman, 1991). Local populations of *F. heteroclitus* are calculated at $10^3$ to $10^6$ individuals, lending to large standing genetic variation (Adams et al., 2006). The species is tolerant to a wide-range of environments ranging from hypo- to hypersalient conditions (Griffith, 1974), lives along the steepest thermal cline in the world (Crawford et al., 1999), and inhabits extremely polluted estuaries (Weis, Weis, 1989). *F. heteroclitus* are easy to collect in the wild and maintain in the laboratory. Their life stages are hardy and individuals can withstand experimental manipulations. For these reasons, *F. heteroclitus*
makes an optimal model species to study the population genetic effects and genetic basis of adaptation to persistent anthropogenic contamination in natural populations.

**EPA Superfund Collection Sites**

New Bedford Harbor, located in Massachusetts, has been polluted since the growth in population during the mid 1800s where human waste was disposed of into the estuary (Voyer *et al.*, 2000). Following the demise of the textile industry in the early 1900s, the city of New Bedford recruited several electrical component manufacturers, who in turn polluted the Acushnet River estuary and adjoining harbor with polychlorinated biphenyls (PCBs). In addition to PCBs, dyes and chemicals used in rubber processing as well as metals also were dumped into the estuary and harbor following the expansion of the New Bedford industrial scene (Weaver, 1984). The abundance and persistence of the complex mix of contaminants placed New Bedford Harbor on the Superfund National Priority List (NPL) by the Environmental Protection Agency in 1983. The EPA manages the *NPL* as sites where known releases or threatened releases of hazardous substances, pollutants, or contaminants have occurred throughout the United States and its territories. The Superfund program is run through the US federal government to clean up the nation's uncontrolled hazardous waste sites.

In New Jersey, at the Diamond Alkali Company site, pesticides were manufactured including dichlorodiphenyltrichloroethane (DDT) and phenoxy herbicides from the mid 1940s to the 1970s (Trustees, 2007) on the Passaic river upstream of Newark Bay. In addition to the Diamond Alkali company site, several other industrial Superfund sites are
located in or upstream of Newark Bay: the Beazer East Inc. site carried out wood treatment from 1940-1991 leading to the polluting of polycyclic aromatic hydrocarbons (PAHs), Federated Metals Corporation operated from 1943-1984 and manufactured magnesium, aluminum, and zinc cathode protection anodes used on steel structures, and the Tidewater Baling site processed a variety of scrap metals. These industrial activities have contaminated sediments of the lower Passaic River and Newark Bay with dioxins, PCBs, mercury, DDT, pesticides, PAHs and heavy metals (Albrecht et al., 1999; Brown et al., 1994; Finley et al., 1997; Huntley et al., 1997; Huntley et al., 1994; Iannuzzi et al., 2005; Mitra et al., 1999; Stackelberg, 1997; Wenning et al., 1994; Yang, Sanudo-Wilhelmy, 1998). Newark Bay was first placed on the National Priority list in 1984.

From 1926 until 1992 the Atlantic Wood Industry, located on the banks of the Elizabeth River, Virginia, treated wood with the complex PAH mixture, creosote, and pentachlorophenol (PCP). The Navy also leased this land from 1960 to 1977 and disposed of abrasive blast media and sludge from the production of acetylene (Agency, 2009). As a result of these industrial operations, the Elizabeth River is contaminated with high concentrations of PAHs, PCP, dioxins, and several metals including arsenic, chromium, copper, lead and zinc (Conrad, Chisholm-Brause, 2004; Conrad et al., 2007; Padma et al., 1998). Elizabeth River was listed as a Superfund site in 1990.

**Resistance to contaminants by Fundulus heteroclitus**

*F. heteroclitus* from all three superfund sites have been shown to be resistant to the contaminants mixtures to which they are continuously exposed. In New Bedford Harbor,
fish are approximately 82 times less sensitive to the embryo-larval toxicity of PCB-126 than reference fish and 194% less responsive to cytochrome P4501A induction by prototypical inducers, PAHs and PCBs (Nacci et al., 1999). A more recent study reported that early life stage LC20 values for PCB126, an indicator of relative sensitivity, were 2,000 less among NBH embryos as compared to reference sites (Nacci et al., 2010). In adult fish, certain PCBs are accumulated far less than in reference fish (Gutjahr-Gobell et al., 1999), differential gene expression patterns between polluted and reference sites have been documented (Fisher, Oleksiak, 2007; Oleksiak, 2008), and refractory CYP1A induction to PAHs and PCBs has been shown (Bello, 1999; Bello et al., 2001). NBH fish are also more susceptible to challenges by marine bacterial pathogens in comparison to reference sites (Nacci et al., 2009) but form fewer DNA adducts upon exposure to benzo(a)pyrene (Nacci et al., 2002).

In Newark Bay, embryo and adult F. heteroclitus are less sensitive to dioxins and other toxicologically similar compounds (Arzuaga, Elskus, 2002; Bozinovic, Oleksiak, 2010; Elskus et al., 1999; McArdle et al., 2004; Nacci et al., 2010; Prince, Cooper, 1995a). Similar to the New Bedford Harbor F. heteroclitus, lack of induction of CYP1A by prototypic inducers (dermal exposure to TCDD) has also been shown (Prince, Cooper, 1995b). Gene expression patterns between Newark and reference site F. heteroclitus show significantly different genes (Fisher, Oleksiak, 2007), many of which in the liver are in the oxidative phosphorylation pathway suggesting significant energy metabolism in polluted fish (Oleksiak, 2008). While seemingly resistant to contaminants, Newark Bay fish exhibit molecular and morphological changes indicative of impaired reproductive health and endocrine disruption compared to the reference population (Bugel et al., 2010).

Elizabeth River embryos also exhibit 10,000 times less sensitivity towards PCB-126 (as measured by LC20) as compared to reference sites, a value that is five times greater than LC20 values for NBH and 14 times greater than Newark Bay embryos (Nacci *et al.*, 2010). Elizabeth River fish show elevated levels of proteins involved in phase II and III metabolism, namely glutathione-S-transferases (Van Veld *et al.*, 1991) and P-glycoproteins (Cooper *et al.*, 1999). Like that of the other Superfund sites, microsomal (Van Veld, 1995; Wills *et al.*, 2010b) and mitochondrial (Jung D., 2010) CYP1A is refractory to induction in *F. heteroclitus* from the Elizabeth River. B(a)P-induced damage was less on mitochondrial and nuclear DNA in Elizabeth River fish as compared to reference fish (Jung *et al.*, 2009), a similar finding to that reported in New Bedford Harbor fish (Bello, 1999; Bello *et al.*, 2001; Nacci *et al.*, 1999). Differential gene expression patterns are also found between Elizabeth River and its reference sites (Fisher, Oleksiak, 2007; Oleksiak, 2008), in particular CYP2N2 which is potentially involved in metabolism of xenobiotics (Oleksiak *et al.*, 2000) is up/down regulated in *F. heteroclitus* from the Elizabeth River. The Elizabeth River population also seems to have concomitantly developed resistance to chronic effects, including cancer, as compared to a reference population (Wills *et al.*, 2010a).

While there are many molecular and biochemical measurements of resistance in *F. heteroclitus*, there is still a lack of understanding of the underlying mechanisms driving adaptation. Since a gene candidate approach, such as characterizing CYP1A, has not
provided concrete answers, a global genome approach to identify selectively important loci may be a more fruitful alternative.

**Population Genetics and Genome-Wide Scans**

Population genomics is the study of numerous loci or genome regions to better understand the roles of evolutionary processes, such as mutation, random genetic drift, gene flow and natural selection, that influence variation across genomes and populations (Luikart *et al.*, 2003). Sampling of the genome across loci is used to determine locus-specific effects like that of selection, mutation, assertive mating and recombination, from genome-wide effects such as drift, bottlenecks, gene flow, and inbreeding. Locus specific effects can be used to identify genes important to fitness and adaptation whereas genome-wide effects provide information about population demography and phylogenetic history. To characterize and measure locus and genome-wide effects genetic or gene-product markers are used.

The earliest markers used to study populations and adaptation were alloenzymes (Hamrick *et al.*, 1979; Lewontin, 1974; Nevo, 1978; Nevo *et al.*, 1984), or variant forms of an enzyme that are coded by different alleles. Alloenzymes are readily translated across species (Nevo, 1990), but lack variability and number of loci which limits their power in statistical analyses for population genetics (Oleksiak, 2010). Since the age of DNA sequencing, alloenzymes have been replaced with more informative DNA markers such as microsatellites, mitochondrial DNA, amplified fragment length polymorphisms (AFLPs), and single nucleotide polymorphisms (SNPs). Microsatellites are repeating sequences of one to six base pairs of DNA and are codominant (homozygote and heterozygotes can be identified).
markers (Bahram, Inoko, 2007) which are highly variable and provide high statistical power for population genetic tests. Amplified fragment length polymorphisms are genomic fragments flanked by restriction sites which are amplified with the polymerase chain reaction (PCR) (Vos et al., 1995) and are dominant (cannot discern homozygotes from heterozygotes) markers. AFLPs are widely used due to their ease of development and scoring in non-model species (Bensch, Akesson, 2005). Single nucleotide polymorphisms are point mutations which are abundant throughout coding and noncoding DNA and are codominant markers (Falque et al., 2005). SNPs are abundant and widespread throughout the genome, can be typed reproducibly by many different methods (sequencing and high-throughput genotyping platforms), and provide high statistical power for population genetic tests.

For a population genomics study, a marker type is selected and tens to hundreds of loci are genotyped in a process called a genome-wide scan (GWS). Genotyping markers is necessary to determine both genome and locus specific effects, and to test for the neutrality of each marker. Neutral loci should not be under selection, are inherited according to Mendelian laws (transmission of heredity characteristics from parent organisms to their offspring), and are in Hardy-Weinberg proportions. Hardy-Weinberg is a law or model in which allele and genotype frequencies will reach equilibrium in one generation and remain constant from generation to generation in large, random-mating populations with no mutation, migration or selection (Luikart et al., 2003). In order to estimate population genetic parameters, such as effective population size ($N_e$), mutation-drift equilibrium, bottlenecks, population subdivision, and migration between populations, markers should be neutral to avoid biasing results. Selectively important loci, or outlier loci, are generally
identified by comparing a data set against a theoretical (simulated) or empirical (observed) null distribution. The null distribution is founded on Kimura’s theory of neutrally evolving mutations that states the vast majority of evolutionary changes at the molecular level are caused by random drift of selectively neutral mutants (Kimura, 1983). The majority of studies use theoretical null distributions because it is difficult to generate robust, empirical null distributions. Thus, robust null distributions are used to model random effects in order to determine loci that are selectively important or linked to selectively important loci (through hitchhiking/linkage) or are outliers due to demography or population dynamics.

Several established tests can be used to detect directional selection in multilocus markers. The most widely used is the $F_{ST}$ or fixation index test (Weir, 1984; Wright, 1951) which calculates the amount of genetic variation (reduction in heterozygosity in subpopulations relative to that expected in the population as a whole with random mating) among groups relative to a panmictic state. When examining multiple populations $F_{ST}$ is often calculated for each locus between a pair of populations (Beaumont, Balding, 2004; Beaumont, Nichols, 1996; Bowcock et al., 1991; Lewontin, Krakauer, 1973; McDonald, Golding, 1994; Porter, 2003; Vitalis et al., 2001). To detect selection at any one locus, simulations are used to generate a null distribution of the $F_{ST}$ summary statistic. Empirical $F_{ST}$ values for each locus are then compared with the null distribution to determine statistical significance. Relative levels of diversity within populations, calculated though Theta ($\Theta$)-ratios tests (ratio between lnRV (variance) and lnRH (heterozygosity) statistics), can also be tested to detect selection (Schlotterer, 2002; Vigouroux et al., 2002; Wootton et al., 2002). Theta is defined as $4N_e\mu$, where $N_e$ is the effective population size and $\mu$ is the mutation rate.
Theta-ratio tests assume that selective sweeps dramatically reduce genetic diversity in target
genomes or genomic regions. Theta-ratios are compared for each locus between populations
and a p-value is calculated for that ratio based on a null distribution. A third test is known as
the Ewens-Watterson test, which tests whether the observed Hardy-Weinberg homozygosity
for a sample of size with \( n \) different alleles is significantly different from homozygosity
expected under the neutral theory if the locus is at mutation-drift equilibrium (Ewens, 1972;
Vigouroux et al., 2002; Watterson, 1978). In practice, genome scans and their associated
statistical tests have been used to identify selection in populations of Drosophila (Harr et al.,
2002; Kauer, 2003; Kauer et al., 2003; Kingan et al.; Orengo, Aguade, 2007; Schofl,
Schlotterer, 2004; Tracy et al.), parasites (Mu et al.; Nair et al., 2003; Wootton et al., 2002),
maize (Vigouroux et al., 2002), intertidal snails (Wilding et al., 2001), lake whitefish
(Campbell, Bernatchez, 2004), cod (Beaumont, Nichols, 1996; Moen et al., 2008b; Pogson et
al., 1995), guppies (Willing et al., 2010), salmon (Vasemagi et al., 2005), three-spined
stickleback (Makinen et al., 2008), frogs (Bonin et al., 2006), oysters (Murray, Hare, 2006),
seagrass (Oetjen, Reusch, 2007), Arabidopsis (Cork, Purugganan, 2005), mosquitos (Bonin et
al., 2009), walking-sticks (Nosil et al., 2008), beech trees (Jump et al., 2006), white spruce
(Namroud et al., 2008), sunflowers (Kane, Rieseberg, 2007), cattle (Gibbs et al., 2009),
horses (Gu et al., 2009), mice (Storz, Dubach, 2004; Storz, Nachman, 2003), rats (Kohn et
al., 2003; Tollenaere, 2010), and humans (Amato et al., 2009; Bowcock et al., 1991; Kayser
et al., 2003; Ronald, Akey, 2005; Storz et al., 2004). These studies all identified regions of
genomes that have been under diverse selective pressures or are linked to areas of the
genome under selection.
Technology has enhanced efforts to sequence and genotype hundreds of genome-wide markers in non-model species. One such technology is called 454 pyrosequencing; it is a highly paralleled DNA sequencing approach that can provide over 20 Mbp of sequence in a four-hour run (Margulies et al., 2005). Average read lengths are about 350-400 base pairs which are derived from approximately 300,000 samples (Huse et al., 2007). This method does not require cloning of the DNA and therefore avoids certain biases that can be introduced by enzymatic steps or by instability of sequences in E. coli while cloning (Weber et al., 2007). Pyrosequencing technology relies upon enzyme cascades and CCD luminescence detection capabilities to measure the release of inorganic pyrophosphate with every nucleotide incorporation (Ronaghi et al., 1998). Sequencing takes place within an oil and water emulsion containing a capture bead. PCR amplification is performed prior to sequencing whereby one copy of the sample is captured on the bead and amplified to millions of copies. The beads are then distributed on a solid-phase sequencing substrate with 1.6 million wells; each well holds a bead and additional reagents including polymerase, luciferase, and ATP sulfurylase. Four nucleotide triphosphates are cycled over the plate, and when a nucleotide is incorporated, a pyrophosphate is released. This pyrophosphate is the substrate for luciferase, so luminescence occurs and is recorded. The number of reads per run makes it possible to have many different copies of the same gene, creating the opportunity for quality control. Further, studies have shown that 454 pyrosequencing is appropriate for many genomic applications, from SNP analysis ((Alderborn et al., 2000; Moen et al., 2008a; Namroud et al., 2008; Novaes et al., 2008; Ronaghi, Elahi, 2002; Wiedmann et al., 2008)) to transcriptome profiling (Bainbridge et al., 2006; Moustafa et al.;
Shin et al., 2008; Vera et al., 2008; Weber et al., 2007). 454 pyrosequencing can be coupled to high-throughput genotyping platforms to expand genotyping to additional individuals or to verify SNPs initially identified with 454.

One platform useful for genotyping SNPs in non-model species due to limited sequence information needs and high multiplexing across SNPs and individuals (which ultimately lowers cost per nucleotide) is MassARRAY. MassARRAY uses an initial locus-specific PCR reaction followed by single base extension using mass-modified dideoxynucleotide terminator of an oligonucleotide primer which anneals immediately upstream of the single nucleotide polymorphic site. Using MALDI-TOF mass spectrometry, the distinct mass of the extended primer identifies the SNP allele (Gabriel et al., 2009). MassARRAY has been used successfully to identify and verify SNPs in a wide array of organisms (Abel et al., 2006; Berard et al., 2009; Buggs et al.; Craig et al., 2009; Goddard et al., 2007; Nakai et al., 2002; Rohrer et al., 2007).

Once SNPs have been identified and verified for population genetic studies, the next step is to characterize selectively important loci by definitively linking genotype to phenotype, moving forward from population genomics to functional genomics and defining how much of the variation is biologically important (Oleksiak, 2010). Several studies have proof of concept between selective sweeps on populations and functional genomic outcomes: coat color and shortened limbs in dog breed formation (Pollinger et al., 2005); chloroquine resistance in the malaria-inducing parasite Plasmodium falciparum (Wootton et al., 2002); cyclodiene insecticide resistance (Ffrench-Constant, 1994); dichlorodiphenyltrichloroethane (DDT) insecticide resistance (Daborn et al., 2002); organophosphate insecticide resistance
warfarin resistance in rats (Kohn et al., 2003); adaptive color pattern in beach mice (Hoekstra et al., 2006); malarial resistance in humans with sickle cell (Currat et al., 2002; Ohashi et al., 2004); differential susceptibility of humans to cerebral malaria (Aitman et al., 2000; Pain et al., 2001); and variation in zebrafish and human pigmentation (Lamason et al., 2005). In these cases, one locus had a large effect on the phenotype. The paucity of examples linking genotype to phenotype may reflect the fact that selection often acts on multiple loci (Lee et al., 2008), especially under heterogeneous environmental conditions (Weinig et al., 2003). It is this latter point, where multiple selectively important loci are contributing to a phenotype, which makes it difficult to directly link genotype to phenotype to fitness.

**Thesis Objectives**

For my thesis, I integrated molecular biology and population genetics in order to define regions of the genome subject to selection in *F. heteroclitus* populations adapted to high concentrations of anthropogenic contaminants. This thesis represents the first, comprehensive, genome-wide study of the molecular variation and signatures of selection among polluted and reference populations of *F. heteroclitus*. To this end, I conducted an AFLP analysis of genome-wide variation on both sensitive and resistant populations of *F. heteroclitus* to assess neutral variation and attempt to identify regions of the genome under selection (chapter 1). Because AFLPs are anonymous, I next used high-throughput sequencing and genotyping methods to assess the technology’s utility in determining SNPs within AFLP fragments and expressed sequence tags (ESTs). Once identified and verified, I
analyzed SNPs among and between populations and species of the *Fundulus* genus to establish population parameters (chapter 2). To establish SNPs under selection in the same polluted and reference *F. heteroclitus* populations used in the first study, I performed three different statistical tests (chapter 3). In the last data chapter (chapter 4), I sought to characterize the potentially adaptive phenotypic outcome of one SNP, a SNP in the CYP1A promoter, that was under selection in all polluted populations. In total, this thesis describes one of the first genome-wide scans of a single species throughout many, variable, natural populations across geographies and environmental conditions. It elucidates selectively important regions of the genome, which may contribute to the phenotypic resistance of *F. heteroclitus* to anthropogenic pollution.
References


phylogeography, Pleistocene history and secondary contact hypotheses for the


Aitman TJ, Cooper LD, Norsworthy PJ, et al. (2000) Malaria susceptibility and CD36

Albrecht ID, Barkovskii AL, Adriaens P (1999) Production and dechlorination of 2,3,7,8-
tetrachlorodibenzo-p-dioxin in historically-contaminated estuarine sediments.
*Environmental Science & Technology* 33, 737-744.

polymorphisms by real-time pyrophosphate DNA sequencing. *Genome Research* 10,
1249-1258.

eheteroclitus* L., in a restored impounded march and in natural reference marshes
*Estuaries* 17, 462-471.


Wills LP, Matson CW, Landon CD, Di Giulio RT (2010b) Characterization of the recalcitrant CYPI phenotype found in Atlantic killifish (Fundulus heteroclitus) inhabiting a Superfund site on the Elizabeth River, VA. Aquatic Toxicology In Press, Corrected Proof.


CHAPTER 1

Signatures of selection in natural populations adapted to chronic pollution

Larissa M. Williams¹ and Marjorie F. Oleksiak²

1. Department of Environmental and Molecular Toxicology
   Box 7633, North Carolina State University
   Raleigh, NC 27695-7633 USA

2. Rosenstiel School of Marine and Atmospheric Sciences
   University of Miami
   4600 Rickenbacker Causeway
   Miami, FL 33149 USA

Corresponding Author:
Marjorie Oleksiak
Rosenstiel School of Marine and Atmospheric Sciences
University of Miami
4600 Rickenbacker Causeway
Miami, FL 33149
Fax: 305-421-4600
Email: moleksiak@rsmas.miami.edu

Published in BMC Evolutionary Biology 2008, 8:282.
Abstract

Background: Populations of the teleost fish Fundulus heteroclitus appear to flourish in heavily polluted and geographically separated Superfund sites. Populations from three Superfund sites (New Bedford Harbor, MA, Newark Bay, NJ, and Elizabeth River, VA) have independently evolved adaptive resistance to chemical pollutants. In these polluted populations, natural selection likely has altered allele frequencies of loci that affect fitness or that are linked to these loci. The aim of this study was to identify loci that exhibit non-neutral behavior in the F. heteroclitus genome in polluted populations versus clean reference populations.

Results: To detect signatures of natural selection and thus identify genetic bases for adaptation to anthropogenic stressors, we examined allele frequencies for many hundreds of amplified fragment length polymorphism markers among populations of F. heteroclitus. Specifically, we contrasted populations from three Superfund sites (New Bedford Harbor, MA, Newark Bay, NJ, and Elizabeth River, VA) to clean reference populations flanking the polluted sites. When empirical $F_{ST}$ values were compared to a simulated distribution of $F_{ST}$ values, 24 distinct outlier loci were identified among pairwise comparisons of pollutant impacted F. heteroclitus populations and both surrounding reference populations. Upon removal of all outlier loci, there was a strong correlation ($R^2 = 0.79$, $p < 0.0001$) between genetic and geographical distance. This apparently neutral evolutionary pattern was not evident when outlier loci were included ($R^2 = 0.092$, $p = 0.0721$). Two outlier loci were shared between New Bedford Harbor and Elizabeth River populations, and two different loci were shared between Newark Bay and Elizabeth River populations.
**Conclusion:** In total, 1% to 6% of loci are implicated as being under selection or linked to areas of the genome under selection in three *F. heteroclitus* populations that reside in polluted estuaries. Shared loci among polluted sites indicate that selection may be acting on multiple loci involved in adaptation, and loci shared between polluted sites potentially are involved in a generalized adaptive response.
Background

The genetic basis of adaptation is a fundamental issue in evolutionary biology. Much of the research in this field has been focused on the classic model systems of *Drosophila* (Beaumont, Balding, 2004; Beaumont, Nichols, 1996; Daborn et al., 2002; Feder, 1999; Hoffmann et al., 2003; Kauer et al., 2002; Kopp et al., 2000; Nuzhdin et al., 2004; Posthuma, Vanstraalen, 1993; Presgraves et al., 2003; Riley et al., 2003; Vitalis et al., 2001; Wiehe et al., 2007) and *Arabidopsis* (Fowler, Thomashow, 2002; Maloof et al., 2001; Pigliucci et al., 2003; Tian et al., 2003; Weinig et al., 2002). Recently, insight into adaptation in non-model species has become possible due to advances in molecular biology and statistics (Allendorf, Seeb, 2000; Beaumont, Nichols, 1996; Bradshaw et al., 2000; Frary et al., 2000; Kohn et al., 2003; Mock et al., 2002; Parsons, Shaw, 2001; Peichel et al., 2001; Pogson et al., 1995; Storz, Dubach, 2004; Storz, Nachman, 2003; Whitehead et al., 2003; Yan et al., 1999). This recent expansion into studies of non-model systems allows further development of evolutionary inferences (Luikart et al., 2003), such as the role that selection, mutation, gene flow, and drift play in adaptation (Wang et al., 2003). A powerful approach to understand genome-wide adaptation is to investigate independent natural populations that inhabit environments with strong selective pressures.

One species that has adapted to a wide range of estuarine environments is the teleost fish, *Fundulus heteroclitus* (Griffith, 1974). *F. heteroclitus* is widely distributed along the United States’ eastern seaboard from the Gulf of St. Lawrence to northeastern Florida (Duvernell et al., 2008). Subpopulations of *F. heteroclitus* inhabit clean estuaries as well as those heavily impacted by chemical pollutants (reviewed in (Wirgin, Waldman, 2004)).
Three well-known polluted sites where *F. heteroclitus* reside are New Bedford Harbor (Massachusetts), Newark Bay (New Jersey), and Elizabeth River (Norfolk, VA). All three sites have been identified by the Environmental Protection Agency (EPA) as Superfund sites (part of the federal government's program to clean up the nation's uncontrolled hazardous waste sites) and contain high levels of a variety of lipophilic, persistent and toxic contaminants worthy of remediation using Federal funds. All three Superfund sites are highly contaminated with chemical pollutants that are broadly classified as aromatics. New Bedford Harbor is polluted with extremely high levels of polychlorinated biphenyls (Pruell *et al.*, 1990) as well as polychlorinated dibenzo-p-dioxins (PCDD), polychlorinated dibenzofurans (PCD), polycyclic aromatic hydrocarbons (PAH), and several trace metals (Bergen *et al.*, 1998; Pruell *et al.*, 1990). Newark Bay is most notorious for containing 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) as well as other dioxins (Prince, Cooper, 1995; Weis, 2002) and also is contaminated with heavy metals, pesticides, PCBs and PAHs (Iannuzzi *et al.*, 2005). The Elizabeth River is predominantly contaminated with creosote, comprised of a complex mixture of PAHs (Bieri *et al.*, 1986; Huggett *et al.*, 1992; Padma *et al.*, 1998).

*F. heteroclitus* from these chronically polluted areas are resistant to the aromatic hydrocarbons in their environment as compared to nearby fish from relatively clean environments (Black *et al.*, 1998; Elskus *et al.*, 1999; Meyer, Di Giuliuo, 2003; Meyer *et al.*, 2002; Nacci *et al.*, 1999; Nacci *et al.*, 2002; Ownby *et al.*, 2002; Vogelbein *et al.*, 1990). Resistance in first and second generation embryos from New Bedford Harbor and Elizabeth River and first generation embryos from depurated Newark Bay fish suggests that differential
survival is due to genetic adaptation rather than physiological induction. Investigating and comparing *F. heteroclitus* from these three sites provides the opportunity to study similarities and differences in adaptation to differing chemical pollutant and resistance to general stress conditions among populations.

Previous work to elucidate mechanisms of resistance and the underlying genetic basis in *F. heteroclitus* from these three sites has investigated the refractory phenotype of the xenobiotic metabolizing enzyme cytochrome P4501A (CYP1A) in polluted populations (Bello *et al.*, 2001; Elskus *et al.*, 1999; Nacci *et al.*, 1999; Prince, Cooper, 1995; Van Veld, Westbrook, 1995), epigenetic silencing through CpG methylation of promoter regions of the CYP1A 5’ promoter region (Timme-Laragy *et al.*, 2005), and elimination of contaminants through the induction of other phase I, II, and III enzymes ((Armknecht *et al.*, 1998; Bard *et al.*, 2002; Bello *et al.*, 2001; Cooper *et al.*, 1999)), many by way of the aryl hydrocarbon receptor (AHR) pathway (reviewed in (Hahn, 1998)). Yet, none of these research efforts has completely accounted for the differences in the resistance phenotypes between polluted and reference site fish in New Bedford Harbor, Newark Bay and Elizabeth River, nor has the genetic basis for resistance been elucidated.

In contrast to a candidate gene approach, our strategy to begin to understand the genetic mechanisms that enable *F. heteroclitus* populations to inhabit these highly polluted sites was to screen the genome for selectively important loci. The premise is that loci under selection will have patterns of variation statistically different from the majority of neutral loci (Lewontin, Krakauer, 1973). Loci that have a large difference in allele frequencies between populations with respect to what would be expected under the neutral expectation are
outliers. The identification of these outliers provides evidence for which and how many loci may be involved in the evolutionary adaptation to anthropogenic pollution.

Loci can have significantly different frequencies relative to other neutral loci for many reasons. To obviate the detection of outliers due to genetic drift rather than selection, our sampling scheme contrasted each polluted population with two reference populations that were geographically more distant from each other than either was to the polluted population. This provides a control for each Superfund site by identifying which loci are significant outliers relative to two reference sites that are demographically distant from each other. To provide extensive coverage of the genome, we used approximately 300 amplified fragment length polymorphisms (AFLP) (Vos et al., 1995) to genotype 288 individuals from nine *F. heteroclitus* populations and used a modeling approach to reveal significant outliers. Furthermore, we investigated whether outlier loci were shared among polluted populations, suggesting similar patterns of selection on the genome despite differences in pollutant compositions and local conditions.

**Methods**

*F. heteroclitus* were collected using minnow traps during the spring of 2005. Fin clip samples from 32 individuals were sampled from each of the nine collection sites along the east coast of the United States (Fig. 1; Table 1). Three of the collection sites were Superfund sites: New Bedford (EPA ID: MAD980731335), Newark (EPA ID: NJD980528996), and Elizabeth River (EPA ID: VAD990710410). Two non-polluted reference sites flanked each
Superfund site, approximately equidistant on either side of each polluted site (Fig. 1; Table 1).

Genomic DNA was extracted from fin clips using a modified version of Aljanabi and Martinez ((Aljanabi, Martinez, 1997)). Fin clips were incubated at 55º C for two hours in 300 µL of 75 mM NaCl, 25 mM EDTA, and 1% SDS with Proteinase K (3 µL of 20 mg/mL). Following incubation, 0.5 volumes of 7.5M ammonium acetate were added and DNA was precipitated on ice with the addition of 0.7 volumes of isopropanol. Subsequently, DNA was pelleted through centrifugation and washed with 70% ethanol. DNA was resuspended overnight at 4º C in 0.1x TE.

The AFLP analysis was performed in replicate following the ligation of the DNA for each individual using a modified version of Vos et al. ((Vos et al., 1995)) to generate approximately 300 loci. Genomic DNA (500 ng) was digested with 5U EcoRI (New England Biolabs, MA) and 5U MseI (New England Biolabs, MA) overnight at 37º C in a total volume of 45 µL containing 1X T4 DNA ligase buffer (Epicentre) supplemented with 100 µg/mL BSA. Following incubation, 50 pmol adaptor oligonucleotides (Applied Biosystems) and 1U T4 DNA ligase (Epicentre) were added and incubated overnight at 16º C. Preselective PCRs were performed in a 15 µL volume using 5 µL of diluted (1:10) ligation product with EcoRI + (C/A) primer (Integrated DNA Technologies; 10 pmol), MseI + (C/A) primer (Integrated DNA Technologies; 10 pmol) and 1U Taq. PCR conditions were 20 cycles of 94º C for 10 sec, 56º for 30 sec, and 72º C for 2 min. Selective Eco + 3NT primers (Integrated DNA Technologies; 10 pmol) labeled with FAM dye at the 5’ end and MseI + 3NT primers (Integrated DNA Technologies; 10 pmol) were added to diluted (1:10)
pre-selective PCR product in a 15 µL volume. PCR conditions in the first cycle were 94°C for 10 sec, 65°C for 30 sec, and 72°C for 2 min with the annealing temperature reduced by 0.4°C for 12 cycles, then 30 cycles of 94°C for 10 sec, 56°C for 30 sec, and 72°C for 2 min. Semi-automated analysis of the selective PCR products was performed on MegaBACE 1000 DNA sequencing system (GE Healthcare). Peak patterns were calculated using MegaBACE Geneprofiler software v. 1.0 (GE Healthcare). The criteria for distinct peaks were a size between 50 and 400 base pairs and an absolute intensity greater than or equal to 1000. Replicated fragments were obtained from all samples (the same template was used for independent PCRs) and replicate fragments were scored as being present or absent using Peakmatcher software (DeHaan et al., 2002). Peakmatcher software automatically creates marker categories and generates a binary table for the presence and absence of markers based on the minimum 75 percent repeatability of markers across replicates.

**Statistical Analysis**

The frequency of band presence allele was calculated using the formula $P = 1 - ((N - C)/N)^{0.5}$ where $N$ equals the sample size and $C$ is the number of individuals with the band (Wilding et al., 2001a). This formula assumes Hardy-Weinberg equilibrium. However, because AFLPs are dominant markers and heterozygotes are not observed, Hardy-Weinberg equilibrium cannot be directly tested. Due to strong selection or increased mutational rates, some of the loci may not be in Hardy-Weinberg equilibrium. Though not directly comparable, microsatellites are in Hardy-Weinberg equilibrium in these *F. heteroclitus* populations (Adams et al., 2006). This calculation also assumes that shared band presence or absence between two individuals is due to common evolutionary origin and not homoplasy.
Pairwise $F_{ST}$ values between populations were calculated for each locus by the method of Nei (Nei, 1977) with the correction of Nei and Chesser (Nei, Chesser, 1983) for finite sample sizes, and a null distribution of $F_{ST}$ values versus allele frequency was simulated using the Winkles program ((Wilding et al., 2001b), Fig. 2).

Winkles is based on the model described in Beaumont and Nichols ((Beaumont, Nichols, 1996b)) which employs coalescent simulations using the Island model and an infinite alleles mutational model. Samples of the same size and number as the data are simulated, where each sample is taken from a different island. This simulation uses two populations of size $N$ diploid individuals, with a set mutation rate, $\mu$, and a migration rate, $m$, per generation. Parameters for the simulation are estimated through the calculation $F_{ST} = 1/(1 + 16Nm + 16N\mu)$. The $F_{ST}$ value is found by calculating the mean $F_{ST}$ from any given pairwise comparison and adjusting that value by -0.0093 to account for the upward bias in the model reported by Wilding et al. ((Wilding et al., 2001b)); this bias is consistent with previous simulations using Nei’s methods to calculate pairwise $F_{ST}$ values (Slatkin, Barton, 1989). The $Nm$ factor is calculated by solving for that parameter in the above equation. Each simulation used $10^3$ and $10^{-4}$ as estimates of $N$ and $\mu$, respectively. Simulated $F_{ST}$ values are relatively unaffected by changing either the sample size of the simulated population or the mutation rate (Beaumont, Nichols, 1996b). Five simulations were run on each pairwise comparison to generate an expected null distribution of 25,000 values. Each simulation started with 500 simulation bi-allelic loci in each of the two populations with uniform random distribution and was allowed to drift for $10N$ generations. The 99th percentile of $F_{ST}$ values within each of the 40 binned mean allele frequency values (each bin
representing a set of 0.025 frequency values from 0 to 1) was calculated after removing monomorphic loci because $F_{ST}$ is strongly dependent on allele frequencies (Beaumont, Nichols, 1996b).

The model we used (Beaumont, Nichols, 1996b) is robust to a wide range of alternative models such as colonization and stepping-stone (Vitalis et al., 2001). It is likely to detect outliers with unusually high $F_{ST}$ values and will identify adaptive selection at one or many loci through pairwise comparisons of populations (Beaumont, Balding, 2004; Vitalis et al., 2001). This model is not able to identify loci under balancing selection and tends to generate discrepancies when numbers of immigrants per generation are unequal, the true population history consists of repeated branching events, or the connectivity of populations is uneven (Vitalis et al., 2001). Isolation, population bottlenecks, and populations which are heterogeneous with respect to their demographic parameters further bias to the model (Beaumont, Nichols, 1996b). There is no evidence for isolation and bottleneck history (Adams et al., 2006) or reduced genetic diversity (McMillan et al., 2006) in our populations. However, if non-homogenous demographic parameters exist (e.g., skewed age structure or sex ratios), this model may be biased. Given the relative robustness of the model to identify loci under adaptive selection, we used theoretical versus experimentally derived allele frequencies for loci to determine significant deviations from the neutral expectation.
Results

Total number of loci among populations

Five different primer combinations (Table 2) were used to amplify approximately 300 loci from 288 individuals from nine different *F. heteroclitus* populations. Among New Bedford Harbor and its reference sites, Sandwich and Point Judith, a total of 296 loci were scored. Of those 296 loci, 11 bands were found to be monomorphic (3.7%). Newark and its two reference sites, Tuckerton and Clinton, had a total of 336 loci, of which 7 loci were monomorphic (2.1%). Elizabeth River and its two reference sites, Magotha and Manteo, had a total of 299 loci, with 4 loci found to be monomorphic (1.3%). Among all populations, 450 distinct loci were scored.

Outlier loci among populations

In comparisons of the three Superfund sites and their clean reference sites, twenty-four loci show patterns indicative of selection. The criteria for identifying these selective loci are that they were identified as outliers in pairwise comparisons of each Superfund site population relative to its two reference site populations (polluted *versus* both references, analyzed separately, i.e. the union of polluted *versus* reference 1 and polluted *versus* reference 2) but not in comparisons between the reference site populations. Eighteen of these twenty-four loci were found in the New Bedford Harbor comparisons, four were found in the Newark Bay comparisons, and six were found in the Elizabeth River comparisons (Fig. 3). Four of these loci were shared between two Superfund site populations suggesting conserved mechanisms of adaptation (Fig. 4).
In the northern most Superfund site, New Bedford Harbor, 42 loci representing 14% of total analyzed loci were located above the simulated 0.99 quantile in the polluted versus one of the references’ comparisons. That is, these 42 loci have $F_{ST}$ values that lie outside the expected neutral distribution of 99% of all loci. This is more than 10 fold greater than the 3 that are expected by chance from the approximately 300 amplified loci. These 42 loci are outliers in the New Bedford Harbor comparison to the Point Judith, RI reference population (36 loci), the Sandwich, MA reference population (23 loci) or relative to both reference sites (18 loci). The 18 outlier loci found in the comparisons of New Bedford Harbor to both of its reference populations were amplified from three different primer combinations, spanning a 100 base pair range (Fig. 3A). The joint probability ($<0.01$ squared or $<0.0001$) indicates that less than one locus should be different in both clean sites versus the Superfund site. These 18 loci are thus implicated as separate loci under selection or linked to areas of the genome under selection. There are 16 loci that are outliers when comparing the two reference populations to each other. Only one of these 16 outlier loci is specific to the clean reference sites; the other 15 are also found in the comparison to the New Bedford Harbor Superfund site to one of these reference sites. No locus was an outlier in all pairwise comparisons.

Newark Bay, NJ is close to the phylogeographic boundary that separates northern and southern populations of *F. heteroclitus* (Adams *et al*., 2006; Smith *et al*., 1998). The Clinton reference population is on the northern side and the Tuckerton reference population is on the southern side. The Newark Bay Superfund site has 26 outlier loci (8% versus 1% expected) relative to these two reference sites: 18 (5%) in the comparison with the Clinton reference
population and 13 (4\%) in the comparison with the Tuckerton reference site population. Four outlier loci are found in both comparisons between the Newark Bay Superfund site and its two clean reference sites (Fig. 3B) and not among clean sites. These four loci are greater than that predicted from the joint probability of differences in both clean sites versus the Superfund site. In pairwise comparisons of the two clean reference sites, 18 loci are outliers. Ten of these 18 loci are common outliers between a northern and two different southern populations i.e., Clinton and Newark Bay populations and Clinton and Tuckerton populations.

Elizabeth River is the most southern Superfund site. The Elizabeth River population, in comparisons to its two reference site populations, had 9 outlier loci (3\%). The Elizabeth River and Magotha reference site comparison had 8 outlier loci (2.7\% of the total loci) whereas the Elizabeth River and Manteo reference site comparison had 7 (2.4\% of the total loci). Six outlier loci were found in both comparisons (Fig. 3C) and not found in the comparison among clean sites. Among the two reference sites (Magotha and Manteo) only three loci were outliers and none of these were unique to the reference-reference comparison. Two loci were in common with outliers from the Elizabeth River-Magotha comparison and one locus was in common with the Elizabeth River-Manteo comparison.

Among the twenty-three loci that were outliers in comparisons only among Superfund sites and both reference sites, four loci are outliers in two of the three Superfund sites (Fig. 4; Table 3). Two of these four outlier loci are shared between New Bedford Harbor and Elizabeth River populations, and two are shared between Newark Bay and Elizabeth River.
None is shared between New Bedford and Newark Bay, nor are any shared among all three Superfund site populations.

$F_{ST}$ values were calculated for comparisons between all sites with and without outlier loci (Table 4). As would be expected, average $F_{ST}$ values were higher in all comparisons before the removal of the outliers. The average $F_{ST}$ value (with outliers) between New Bedford Harbor and its reference sites is 0.038, between Newark and its reference sites it is 0.039, and between Elizabeth River and its reference sites it is 0.018. Upon removal of the outliers, average $F_{ST}$ values fall to 0.010, 0.016, and 0.011 for New Bedford Harbor, Newark Bay, and Elizabeth River, respectively. These values were plotted against log-ten of geographic distance between sites versus genetic distance $[F_{ST}/(1 - F_{ST})$, (Rousset, 1997)].

There is no apparent pattern in the distribution of pairwise comparisons corresponding to reference-reference, polluted-reference, or polluted-polluted sites. When outliers were included in the calculation of average $F_{ST}$ and plotted against distance, there was no significant linear relationship ($R^2 = 0.092$, $p = .0721$). Upon removal of the outliers, there was a significant and strong linear relationship ($R^2 = 0.79$, $p < 0.0001$) between geographic and genetic distance (Fig. 5). Mantel tests that account for multiple comparisons confirmed the significance of both relationships (data not shown). This relationship indicates that 79% of the variability in the neutral genetic distance (without outlier loci) between sites can be explained by geographic distance.
**Discussion**

Multiple *F. heteroclitus* populations have independently evolved adaptive resistance to complex suites of pollutants (Black *et al.*, 1998; Elskus *et al.*, 1999; Gutjahr-Gobell *et al.*, 1999; Meyer, Di Giuliuo, 2003; Meyer *et al.*, 2002; Nacci *et al.*, 1999; Nacci *et al.*, 2002; Ownby *et al.*, 2002; Powell *et al.*, 2000; Vogelbein *et al.*, 1990). These different populations provide independent contrasts for identifying loci involved in adaptation. We identified loci suggestive of adaptation for each polluted population by identifying outlier loci in the polluted population relative to two nearby reference populations. These loci are outliers because they are statistically different from the neutral distribution among populations. Only loci exhibiting a non-neutral distribution in comparisons of the polluted population *versus* both a north and south reference population were considered to be adaptive. Through this comparison, we are more likely to identify loci whose non-neutral distribution is due to pollution rather than geography. Similarly, while the model used to identify outlier loci has a false positive rate of approximately 7% (Beaumont, Balding, 2004), it is unlikely that the same loci will be falsely identified in multiple comparisons (*i.e.*, in the polluted population *versus* both a north and south reference populations). In each of the Superfund sites, 1% to 6% (four to 18 loci out of approximately 300) of amplified fragments were identified as being loci under selection or linked to areas of the genome under selection. Four of these loci were outliers in two separate Superfund population comparisons.

We only consider loci exhibiting a non-neutral distribution in comparisons of the polluted population *versus* both a north and south reference population to be adaptive. These populations make up a geographic triangle formed among the northern and southern clean
reference populations and a latitudinally intermediate polluted population (Fig. 1). This
double comparison ensures that we are not identifying loci that differ simply due to genetic
drift or clinal variation common to this species. This contrast, in addition to the joint
comparison among populations, address most of the possible neutral or demographic models.
Population isolation can alter allele frequencies among populations. One would expect that a
single population that suffered from unique isolation would have significantly greater $F_{ST}$
values among many loci in comparison to similarly geographically distance populations that
were not uniquely isolated. This demographic explanation does not fit the data for two
reasons: 1) it is the statistically different $F_{ST}$ value for a few loci in comparison to all other
loci that we define as being important, and 2) all non-outlier loci follow the more common
demographic trend of isolation by distance (Fig. 5). However, differences in $F_{ST}$ values also
can result if loci under functional constraints evolve more slowly than loci without functional
constraints. Thus, loci with large $F_{ST}$ values would have few, if any constraints, relative to
the hundreds of other AFLP loci. However, our comparisons were based on both a
significant $F_{ST}$ between both reference sites versus a polluted site and insignificant
differences among reference sites (as well a a difference from the permutation model, see
methods). Because we are using three criteria (significant difference versus the joint
distribution in two reference sites, lack of a difference among reference sites, and a statistical
difference from a neutral permutation model), it seems most parsimonious to suggest that
these outlier loci are due to natural selection. However, lack of Hardy-Weinberg equilibrium
or recent mutations also might cause loci to be outliers. We suggest that the most obvious
cause for this evolved difference is chronic exposure to the aromatic hydrocarbons and other
anthropogenic pollutants; yet, we cannot explicitly control every variable in natural environments. Other selective forces also could be different between the three sites. For instance, site complexity differs among the nine sites with the three polluted sites tending to be less complex (have less edges) than the reference sites. Thus, predation or food availability might differ among sites. Similarly, salinity might affect food availability or absorption, and although all populations inhabit brackish waters, the Elizabeth River population is less coastal than the reference populations to which it is compared. Under controlled laboratory conditions, survival differs among fish from clean populations exposed to polluted sediments and fish from polluted populations exposed to clean sediments. This phenomenon points towards adaptation to anthropogenic contaminants rather than differing local conditions for the differences seen between polluted and reference populations. Thus we postulate that outlier loci are due to pollution, especially those loci shared among separate Superfund populations.

Most of the outlier loci are unique to a single polluted population rather than shared across polluted populations (Fig. 4). One explanation for the lack of shared loci is that different loci are involved in the adaptation to a particular pollutant or stress. Alternatively, some of these outliers might be linked to the same locus in the different populations and only appear to be different because the locus under selection dragged different polymorphisms to fixation. This could occur because different polymorphisms existed in the different ancestral populations.

Resistance to pollution is a modern phenotype in *F. heteroclitus* due to recent exposure (approximately within the last 60 years), suggesting that *F. heteroclitus* have rapid
evolutionary responses with respect to their environment. Our data and other data on survival and development indicate that populations of *Fundulus* have adapted to local pollutants and thus selection has favored a few alleles. Resistance phenotypes resulting from rapid evolution have been well documented in plants (Forbes, 1999) and benthic invertebrates (Klerks, Levinton, 1989) in response to metals as well as in insects in response to pesticides (McKenzie, 1996) and depend both on population dynamics as well as the strength of selection. *F. heteroclitus* populations residing in chronically polluted areas provide an advantageous situation whereby strong selective pressures and rapid evolution can be studied. *F. heteroclitus* have high standing genetic variation (Mitton, Koehn, 1975), high reproductive potential (Weis, 2002), limited home ranges (Lotrich, 1975) and large population sizes exceeding 10,000 in a single tidal creek (Adams et al., 2006). These attributes can and have resulted in locally adapted *F. heteroclitus* populations. Adaptation due to positive selection often reduces genetic variation among natural populations because of selective sweeps. For example, reduced genetic variation has occurred in brown rats resistant to the rodenticide, warfarin (Kohn, Pelz, 1999; Kohn et al., 2000; Kohn et al., 2003), tobacco budworm exposed to the pyrethroid insecticide (Taylor et al., 1995), and the human malarial parasite, *Plasmodium falciparum*, exposed to antimalarial agents (Wootton et al., 2002). However, genetic diversity is not reduced in the polluted *F. heteroclitus* populations compared to the reference site populations for either neutral markers (McMillan et al., 2006; Mulvey et al., 2003; Roark et al., 2005) or gene expression (Fisher, Oleksiak, 2007). Maintenance of genetic diversity in these populations subjected to significant selection most likely represents steady influx of alternative alleles by migration. If migration
and resulting gene flow is strong enough to prevent the reduction of genetic diversity at non-selected loci, it suggests that selection at adaptively important loci is equally strong. Importantly, with constant influx of allelic variation at loci without adaptive value, there should be fewer spurious allelic differences among populations. Thus, shared loci between Superfund populations are likely to be affected by selection and therefore biologically important.

Among three *F. heteroclitus* populations inhabiting highly polluted Superfund sites and flanking reference populations, 63 different loci (14% of the collective 450 loci) have \( F_{ST} \) values outside the 99% quantile. Using all loci (*i.e.*, including outliers) our \( F_{ST} \) values based on AFLP (0.038, 0.039, and 0.018 for New Bedford Harbor, Newark Bay and Elizabeth River, respectively) are approximately one-half of those found for microsatellites (0.077, 0.068, and 0.043, respectively (Adams *et al.*, 2006)) although these genetic measures are difficult to compare due to differences in genomic coverage and mutation rates (Gaudeul *et al.*, 2004). Using AFLPs, McMillan *et al.* (McMillan *et al.*, 2006) found similar \( F_{ST} \) values for the New Bedford Harbor population (0.056). For the Elizabeth River population, Mulvey *et al.* (Mulvey *et al.*, 2003) also found similar \( F_{ST} \) values (0.014) using allozymes. Notice that these calculated \( F_{ST} \) values use all loci and do not distinguish between neutral and non-neutral loci. If selection affects the frequency of alleles among these molecular markers, the perceived genetic distance (\( F_{ST} \)) will be exaggerated.

The neutral hypothesis is a powerful tool to explore differences among populations (Kreitman, 1996). However, in order to test evolutionary hypotheses, one needs to distinguish between neutral and non-neutral loci. Among populations for each Superfund
site, the genetic distances among local populations are affected by the outlier loci. New Bedford Harbor and Newark Bay populations are more differentiated in comparison to their reference site populations than the Elizabeth River populations ($F_{ST}$ values of 0.038 and 0.039 versus 0.018) because the Elizabeth River population has the fewest outlier loci (2.4% - 2.7%) in comparison to neutral loci. These differences among Superfund sites do not exist upon removal of outliers: $F_{ST}$ values among loci without outlier values are similar for New Bedford Harbor, Newark Bay and Elizabeth River (0.01, 0.016, and 0.011, respectively).

With outliers, there is no relationship between $F_{ST}$ values and geographic distance. However, upon removal of outlier loci, there is a strong relationship between genetic and geographical distance indicating an equilibrium model of isolation-by-distance. Similar findings have been shown in other *F. heteroclitus* studies (Adams et al., 2006; Roark et al., 2005), with the intertidal snail (Wilding et al., 2001b), and sea trout (Hansen, Mensberg, 1998). Not surprisingly, these data indicate that loci with unusually large $F_{ST}$ values have a large and potentially misleading effect on the perceived genetic distance among populations. The 63 outliers exhibit this effect; once removed from the data set, the neutral expectation of increasing genetic distance with geographic distance holds true. For twenty-four of these outlier loci, this non-neutral distribution is most likely caused by evolution by natural selection due to pollution or another strong selective force unique to the polluted sites since the geographical effect was taken into account through the comparison of the polluted sites with both a north and south reference population. Ten other loci have a larger than expected distance at the north-south phylogenetic boundary and likely reflect the historic split among northern and southern *F. heteroclitus* populations (Cashon et al., 1981; Gonzalezvillasenor,
Powers, 1990; Ropson et al., 1990). Outlier loci in reference-reference pairwise comparisons likely reflect genetic drift although some may be due to selection. While we can only speculate why these and the remaining 29 loci affect the relationship between genetic and geographic distance, this illustrates the need to distinguish among potentially selected and neutral loci to determine expected differences and posit hypotheses.

**Conclusions**

Contrasting populations that experience different selective pressures provides insight into evolution by natural selection. Our goal is to understand the genetic basis of adaptive resistance to pollution in chronically contaminated natural populations. Future analyses will address whether polymorphisms between populations are functional and potentially responsible for conferring resistance in populations adapted to chronic exposure to chemical pollutants in the different Superfund sites. We have shown that between 1 to 6% of loci are implicated as being under selection or linked to areas of the genome under selection in three distinct *F. heteroclitus* populations that reside in polluted Superfund estuaries. Shared loci affected by natural selection among polluted sites indicate that there may be a similar mechanism of resistance in these different populations. This study suggests that multiple loci may be involved in adaptation and a few of these loci have a generalized adaptive response.
Acknowledgements

The Authors thank G. Bozinovic for assistance in field collections and D. Crawford for valuable comments on an earlier version of this manuscript. Funding was partially provided by NIEHS Training Grant ES007046 award from the Department of Environmental and Molecular Toxicology at North Carolina State University to LMW and NIH 5 RO1 ES011588 to MFO.
References


*Aquatic Toxicology* **45**, 99-113.


Fowler S, Thomashow MF (2002) Arabidopsis transcriptome profiling indicates that multiple regulatory pathways are activated during cold acclimation in addition to the CBF cold response pathway. 


Table 1. Sample locations. Locations along the east coast of the United States where *F. heteroclitus* was collected.

<table>
<thead>
<tr>
<th>Reference/Superfund</th>
<th>Abbreviation</th>
<th>Geographical location</th>
<th>Latitude (N)</th>
<th>Longitude (W)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>SAND</td>
<td>Sandwich, MA</td>
<td>41°44.0’</td>
<td>70°23.0’</td>
</tr>
<tr>
<td>Superfund</td>
<td>NBH</td>
<td>New Bedford, MA</td>
<td>41°34.0’</td>
<td>70°54.9’</td>
</tr>
<tr>
<td>Reference</td>
<td>PTJ</td>
<td>Point Judith, RI</td>
<td>41°21.7’</td>
<td>71°28.9’</td>
</tr>
<tr>
<td>Reference</td>
<td>CLI</td>
<td>Clinton, CT</td>
<td>41°15.3’</td>
<td>72°32.8’</td>
</tr>
<tr>
<td>Superfund</td>
<td>NEW</td>
<td>Newark, NJ</td>
<td>40°41.2’</td>
<td>74°06.7’</td>
</tr>
<tr>
<td>Reference</td>
<td>TUCK</td>
<td>Tuckerton, NJ</td>
<td>39°32.2’</td>
<td>74°19.4’</td>
</tr>
<tr>
<td>Reference</td>
<td>MAG</td>
<td>Magotha, VA</td>
<td>37°10.6’</td>
<td>75°56.5’</td>
</tr>
<tr>
<td>Superfund</td>
<td>ER</td>
<td>Elizabeth River, VA</td>
<td>36°48.5’</td>
<td>76°17.7’</td>
</tr>
<tr>
<td>Reference</td>
<td>MAN</td>
<td>Manteo, NC</td>
<td>35°53.8’</td>
<td>75°36.9’</td>
</tr>
</tbody>
</table>
Table 2. Primer sequences used in AFLP analyses.

<table>
<thead>
<tr>
<th>Primers</th>
<th>Sequence (5'-3')</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Eco +1</strong></td>
<td></td>
</tr>
<tr>
<td>Eco +A</td>
<td>GACTGCGTACCAATTCA</td>
</tr>
<tr>
<td>Eco +C</td>
<td>GACTGCGTACCAATTCC</td>
</tr>
<tr>
<td><strong>Mse +1</strong></td>
<td></td>
</tr>
<tr>
<td>Mse +A</td>
<td>GATGAGTCCTGAGTAAA</td>
</tr>
<tr>
<td>Mse +C</td>
<td>GATGAGTCCTGAGTAAC</td>
</tr>
<tr>
<td><strong>Eco +3</strong></td>
<td></td>
</tr>
<tr>
<td>Eco +ACT</td>
<td>GACTGCGTACCAATTCACT</td>
</tr>
<tr>
<td>Eco +ACC</td>
<td>GACTGCGTACCAATTCACC</td>
</tr>
<tr>
<td>Eco +AAG</td>
<td>GACTGCGTACCAATTCAGG</td>
</tr>
<tr>
<td><strong>Mse +3</strong></td>
<td></td>
</tr>
<tr>
<td>Mse +AGT</td>
<td>GATGAGTCCTGAGTAAGG</td>
</tr>
<tr>
<td>Mse +ATC</td>
<td>GATGAGTCCTGAGTAAATC</td>
</tr>
<tr>
<td>Mse +CAA</td>
<td>GATGAGTCCTGAGTAACAA</td>
</tr>
<tr>
<td>Mse +CGA</td>
<td>GATGAGTCCTGAGTAACGA</td>
</tr>
</tbody>
</table>

**Combinations**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Eco+ACT and Mse+AGT</td>
</tr>
<tr>
<td>B</td>
<td>Eco+ACC and Mse+ATC</td>
</tr>
<tr>
<td>C</td>
<td>Eco+AAG and Mse+CAA</td>
</tr>
<tr>
<td>D</td>
<td>Eco+ACT and Mse+CGA</td>
</tr>
<tr>
<td>E</td>
<td>Eco+ACC and Mse+CAA</td>
</tr>
</tbody>
</table>
Table 3. Outlier loci shared among the Superfund site *Fundulus* populations.

<table>
<thead>
<tr>
<th>Population 1 and locus number</th>
<th>Population 2 and locus number</th>
<th>Primer Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Bedford Harbor, 19</td>
<td>Elizabeth River, 120</td>
<td>A</td>
</tr>
<tr>
<td>New Bedford Harbor, 98</td>
<td>Elizabeth River, 190</td>
<td>B</td>
</tr>
<tr>
<td>Newark Bay, 78</td>
<td>Elizabeth River, 147</td>
<td>A</td>
</tr>
<tr>
<td>Newark Bay, 335</td>
<td>Elizabeth River, 194</td>
<td>C</td>
</tr>
</tbody>
</table>
Table 4. Pairwise $F_{ST}$ values with and without outlier loci. Mean $F_{ST}$ between populations of *Fundulus heteroclitus* with and without outlier loci. Below diagonal: mean $F_{ST}$ including outlier loci. Above diagonal: mean $F_{ST}$ without outlier loci. Average of $F_{ST}$ values below diagonal is 0.034 and 0.023 after the removal of outlier loci.

<table>
<thead>
<tr>
<th></th>
<th>SAND</th>
<th>NBH</th>
<th>PTJ</th>
<th>CLI</th>
<th>NEW</th>
<th>TUCK</th>
<th>MAG</th>
<th>ER</th>
<th>MAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>SAND</td>
<td>0.0090</td>
<td>0.0157</td>
<td>0.0258</td>
<td>0.0238</td>
<td>0.0213</td>
<td>0.0220</td>
<td>0.0315</td>
<td>0.0298</td>
<td></td>
</tr>
<tr>
<td>NBH</td>
<td>0.0361</td>
<td>0.0112</td>
<td>0.0168</td>
<td>0.0249</td>
<td>0.0299</td>
<td>0.0318</td>
<td>0.0318</td>
<td>0.0277</td>
<td></td>
</tr>
<tr>
<td>PTJ</td>
<td>0.0399</td>
<td>0.0399</td>
<td>0.0149</td>
<td>0.0228</td>
<td>0.0219</td>
<td>0.0250</td>
<td>0.0338</td>
<td>0.0308</td>
<td></td>
</tr>
<tr>
<td>CLI</td>
<td>0.0313</td>
<td>0.0245</td>
<td>0.0190</td>
<td>0.0211</td>
<td>0.0278</td>
<td>0.0309</td>
<td>0.0288</td>
<td>0.0370</td>
<td></td>
</tr>
<tr>
<td>NEW</td>
<td>0.0286</td>
<td>0.0294</td>
<td>0.0283</td>
<td>0.0584</td>
<td>0.0112</td>
<td>0.0239</td>
<td>0.0338</td>
<td>0.0328</td>
<td></td>
</tr>
<tr>
<td>TUCK</td>
<td>0.0270</td>
<td>0.0309</td>
<td>0.0264</td>
<td>0.0584</td>
<td>0.0197</td>
<td>0.0230</td>
<td>0.0218</td>
<td>0.0247</td>
<td></td>
</tr>
<tr>
<td>MAG</td>
<td>0.0365</td>
<td>0.0388</td>
<td>0.0303</td>
<td>0.0355</td>
<td>0.0310</td>
<td>0.0322</td>
<td>0.008</td>
<td>0.0159</td>
<td></td>
</tr>
<tr>
<td>ER</td>
<td>0.0387</td>
<td>0.0393</td>
<td>0.0460</td>
<td>0.0294</td>
<td>0.0722</td>
<td>0.0320</td>
<td>0.0140</td>
<td>0.0144</td>
<td></td>
</tr>
<tr>
<td>MAN</td>
<td>0.0349</td>
<td>0.0330</td>
<td>0.0350</td>
<td>0.0464</td>
<td>0.0607</td>
<td>0.0318</td>
<td>0.0243</td>
<td>0.0217</td>
<td></td>
</tr>
</tbody>
</table>
Figure 1. Sample locations. Sampling locations for *Fundulus heteroclitus* populations. Circles are reference sites and stars are Superfund sites.
Figure 2. $F_{ST}$ versus allele frequency values. $F_{ST}$ values estimated from approximately 300 variable AFLP loci plotted against mean allele frequency. The solid line represents the 0.99 quantile estimated from a simulation model for each comparison. Loci shared among the same Superfund site are labeled with their primer set (letter) and number. Loci shared between Superfund sites are starred. § Shared loci included in these points are: A2, A19, A34, A56, D87, E118, E127, E137, E150, E156, C186, C194, C205, and C252. E118 also is shared between New Bedford Harbor and Elizabeth River populations.
Figure 3. Venn diagrams of shared outlier loci in each Superfund comparison. Outlier loci in comparisons of each Superfund populations to both its clean reference sites; numbers in the unions of circles represent outlier loci shared among populations. A) New Bedford Harbor, MA Sandwich, MA and Pt. Judith, RI comparison. B) Newark Bay, NJ, Clinton, CT, and Tuckerton, NJ comparison. C) Elizabeth River, VA, Magotha, VA and Manteo, NC comparison.
Figure 4. Venn diagram of shared outlier loci among Superfund populations. Shared outlier loci among Superfund population comparisons to both clean reference sites; numbers in the unions of circles represent outlier loci shared between two Superfund populations.
Figure 5. Geographic versus genetic distance. Relationship between genetic distance and geographic distance. Genetic distance was calculated from the mean F_{ST} for each pair of populations with (A) and without (B) outlier loci. Circles represent a pairwise comparison of a Superfund versus a reference site, squares represent a Superfund versus a Superfund site comparison, and crosses represent a reference versus a reference site comparison.
CHAPTER 2

SNP identification, verification, and utility for population genetics in a non-model genus

Larissa M. Williams¹, Xin Ma², Adam R. Boyko², Carlos D. Bustamante², and Marjorie F. Oleksiak³

1. Department of Environmental and Molecular Toxicology
   Box 7633, North Carolina State University
   Raleigh, NC 27695-7633 USA

2. Department of Biological Statistics and Computational Biology
   Cornell University
   Ithaca, NY 14853 USA

3. Rosenstiel School of Marine and Atmospheric Sciences
   University of Miami
   4600 Rickenbacker Causeway
   Miami, FL 33149 USA

Corresponding Author:
Marjorie Oleksiak
Rosenstiel School of Marine and Atmospheric Sciences
University of Miami
4600 Rickenbacker Causeway
Miami, FL 33149
Fax: 305-421-4600
Email: moleksiak@rsmas.miami.edu

Published in BMC Genetics 2010, 11:32.
Abstract

Background: By targeting SNPs contained in both coding and non-coding areas of the genome, we are able to identify genetic differences and characterize genome-wide patterns of variation among individuals, populations and species. We investigated the utility of 454 sequencing and MassARRAY genotyping for population genetics in natural populations of the teleost, *Fundulus heteroclitus* as well as closely related *Fundulus* species (*F. grandis*, *F. majalis* and *F. similis*). Results: We used 454 pyrosequencing and MassARRAY genotyping technology to identify and type 458 genome-wide SNPs and determine genetic differentiation within and between populations and species of *Fundulus*. Specifically, pyrosequencing identified 96 putative SNPs across coding and non-coding regions of the *F. heteroclitus* genome: 88.8% were verified as true SNPs with MassARRAY. Additionally, putative SNPs identified in *F. heteroclitus* EST sequences were verified in most (86.5%) *F. heteroclitus* individuals; fewer were genotyped in *F. grandis* (74.4%), *F. majalis* (72.9%), and *F. similis* (60.7%) individuals. SNPs were polymorphic and showed latitudinal clinal variation separating northern and southern populations and established isolation by distance in *F. heteroclitus* populations. In *F. grandis*, SNPs were less polymorphic but still established isolation by distance. Markers differentiated species and populations. Conclusions: In total, these approaches were used to quickly determine differences within the *Fundulus* genome and provide markers for population genetic studies.
Background

High throughput sequencing and genotyping has become increasingly faster, less expensive and more accurate. In recent years this has lead to the establishment of myriad data sets ranging from increased coverage of variation in the human genome at the individual level (Bordoni et al., 2008; Garber et al., 2009; Ingman, Gyllensten, 2009; Turner et al., 2009; Zheng et al., 2009) to the sequencing of non-model prokaryotic and eukaryotic genomes and transcriptomes (Bontell et al., 2009; De Schutter et al., 2009; Iacono et al., 2008; Novaes et al., 2008; Vera et al., 2008; Worden et al., 2009). For many organisms sequencing of entire genomes is still unattained, but smaller, more targeted portions of the genome can be easily sequenced and genotyped. Such data can provide genome-wide sequence information which can be used to characterize population and selection pressure parameters as well as provide evolutionary insights that are broadly applicable (Luikart et al., 2003).

One non-model genus, Fundulus, includes closely related species that range in physiology, environmental and habitat preference, and geographic locales; Fundulus heteroclitus and Fundulus majalis inhabit the Atlantic coast, and Fundulus grandis and Fundulus similis inhabit the Gulf Coast. Many Fundulus species and/or populations have extensive euryhaline capabilities, respond well to varying ranges of hypoxia (Diaz, 2001; Diaz, Rosenberg, 1995; Smith, Able, 2003), live along a steep thermocline, and have adapted to extremely polluted areas (Wirgin, Waldman, 2004). A variety of studies have investigated the underlying genetic basis of this teleosts’ phenotypic plasticity. While some of the transcriptome is known for F. heteroclitus (Fisher, Oleksiak, 2007; Gonzalez et al., 2006;
Meyer et al., 2005; Oleksiak, 2008; Oleksiak et al., 2002; Oleksiak et al., 2005; Paschall et al., 2004; Peterson, Bain, 2004; Roling et al., 2006; Whitehead, Crawford, 2005; Whitehead, Crawford, 2006) much of the genome-wide variation within and between populations and species for this genus is relatively unknown.

Establishing a set of genetic markers, which can be used to assess regions of the genome involved in local adaptation and in speciation is important to understand fundamental similarities and differences between populations and species of Fundulus. Once markers are established they can be further studied to look for signatures of selection to any number of evolutionary forces (e.g., pollution, hypoxia, salinity, temperature). A few studies have established genetic differences between populations of F. heteroclitus mainly with respect to phylogeographic constraints (Adams et al., 2006; Bernardi et al., 1993) or selection (Cashon et al., 1981; Crawford et al., 1989; Crawford, Powers, 1989; Crawford, Powers, 1992; McMillan et al., 2006; Powers, Place, 1978; Powers et al., 1986; Whitehead, 2009; Williams, Oleksiak, 2008). These studies used microsatellite, mitochondrial DNA, and AFLP analyses as well as targeted gene approaches. Single nucleotide polymorphisms (SNPs) are a useful starting point to scan large and disparate regions of the genome due to their abundance in both coding and non-coding regions, their co-dominant nature, and lack of ambiguity.

SNPs have been used to establish differences between individuals (Gill, 2001), populations (Paschou et al., 2007; Weir et al., 2005; Yamaguchi-Kabata et al., 2008) and species (Kong et al., 2008; Primmer et al., 2002). They also are useful markers for propensity to disease (Amos et al., 2008; Johnson et al., 2007; Tomlinson et al., 2007).
disease states (Poehlmann et al., 2007), and evidence of the genetic basis of adaptation (Hoekstra et al., 2006; Mauricio et al., 2003; Moen et al., 2008; Namroud et al., 2008). In vertebrates, a SNP occurs on average every 100 to 1000 base pairs and often is in linkage disequilibrium with many other SNPs along the chromosome, forming strong haplotypes, which can be easily identified (Vignal et al., 2002). Unfortunately, SNP resources are not readily available in the majority of non-model species lacking genomic resources. With this in mind, we set out to establish a set of SNP markers to identify differences between Fundulus populations and species.

Methods

Sample Collection and Extraction

F. heteroclitus were collected using minnow traps during the spring of 2005. Spleen and testes were sampled from 20 individuals from each of ten collection sites along the East coast of the United States (Figure 1). F. grandis were collected using minnow traps during the winter of 2009 (Figure 1). Fin clips were sampled from 15 individuals from each of the six collection sites along the Gulf Coast of the United States. Spleen from F. majalis was extracted from 13 individuals from Woods Hole, Massachusetts and 10 individuals from Sapelo Island, GA. Spleen also was extracted from F. similis collected from Pensacola, Florida (3 individuals) and Corpus Christi, Texas (8 individuals).

Genomic DNA from spleen and testes was extracted by phenol and chloroform as described in Wirgin et al. (Wirgin et al., 1990), and DNA was resuspended in 50 µL 0.1X TE buffer. Genomic DNAs from fin clips were extracted using a modified version of Aljanabi
and Martinez (Aljanabi, Martinez, 1997) and DNA was resuspended in 50 µL 0.1X TE buffer.

**DNA Pyrosequencing**

*F. heteroclitus* genomic DNAs (500 ng) from eight individuals in each of ten collection sites (all sites except Point Judith, RI, Figure 1A) were digested individually with 1U BspE1 (New England Biolabs, MA) and 1U EcoRI (New England Biolabs, MA). Samples were incubated for three hours at 37º C in a total volume of 30 µL containing Buffer 3 (New England Biolabs, MA). Adaptors (Table 1) to each of the restriction sites, 25 mM ATP, and 1U of T4 DNA ligase (Epicentre) were added to reactions and incubated at 16º C overnight. A 2’ O-methyl block was added to the 3’ cytosine base on the adapter. This block assured that only those fragments digested with both BspEI and EcoRI would be amplified with PCR and prevented amplification of fragments with the same type of restriction site on both ends of the fragment.

Preselective PCR reactions with primers specific to adaptors (Table 1) were performed in a total volume of 25 µL containing 2 µL of diluted (1:10 in 0.1x Tris-EDTA buffer) ligation product with EcoRI primer (Integrated DNA Technologies; 10 pmol), BspE1 primer (Integrated DNA Technologies; 10 pmol) and 1U *Taq*. PCR conditions were 20 cycles of 94º C for 10 sec, 49º for 30 sec, and 72º C for one min. Following the preselective amplification, a selective amplification was carried out to decrease the number of fragments amplified in each individual to approximately 200 by extending the primer on the 3’ end. Preselective PCR products were diluted (1:10) and 2 µL of diluted product was amplified with primers (Table 1) to EcoRI+ AAG (Integrated DNA Technologies; 10 pmol) and BspEI
+C (Integrated DNA Technologies; 10 pmol) with 1U Taq in a 25 µL total volume. PCR conditions in the first cycle were 94º C for 10 sec, 65º C for 30 sec, and 72º C for one minute with the annealing temperature reduced by 0.5º C for 20 cycles, then 25 cycles of 94º C for 10 sec, 55º C for 30 sec, and 72º C for one minute.

Primers (Table 1) specific to the EcoRI restriction site were generated with the goals of labeling the DNA fragments from each individual with specific nucleotide barcodes (Parameswaran et al., 2007) and preparing those samples for emulsion-based amplification. Starting at the 5’ end, 19 nucleotides (Table 1) complementary to the primer on the DNA capture beads used in the emulsion PCR reaction (Margulies et al., 2005) were synthesized (Integrated DNA Technologies). Following those nucleotides, each primer had a distinct 10 base pair barcode (Parameswaran et al., 2007) used to identify individuals (ten primers in total). The final 19 base pairs of the primer were specific to the EcoRI adapter. The BspE1 primer (Table 1) started at its 5’ end with 19 nucleotides (Table 1), which were complementary to the primer on the DNA capture beads followed by 18 base pairs specific to the BspE1 adapter (Figure 2). All primers were HPLC purified. Amplified selective fragments were diluted (1:10) and added to both EcoRI and BspE1 primers (Integrated DNA Technologies; 10 pmol) in a 25 µL volume. PCR conditions were 94º C for 10 sec, 50º C for 30 sec, and 72º C for one minute and were carried out for 30 cycles. PCR reactions were pooled into eight wells, where each of the ten distinct barcodes was represented only once in each of the pools. Each pool of PCR products was purified using QIAquick PCR Purification Kit (Qiagen, USA). PCR products were further purified with AMPure (Agencourt).
Emulsion PCR was carried out on PCR products as described (Margulies et al., 2005). Amplification of the PCR product on the bead was controlled for by quantifying and calculating the size of the amplicon pool using a Bioanalyzer 2100 so that there was a minimum of $2 \times 10^6$ copies of DNA that ranged in size from 100 to 700 base pairs. Subsequent products were sequenced on a Roche/454 Life Sciences GS FLX Sequencer at the University of South Carolina’s Environmental Genomics Core Facility. The PicoTiter plate was subdivided into eight regions with an expectation of 30,000 reads per region (Meyer et al., 2008).

**Assembly of pyrosequencing sequences and SNP Detection**

Sequences were trimmed of their barcodes. All 626 sequences with at least one ambiguous base were removed since the presence of even a single ambiguous base is an effective indicator of low-quality sequence (Huse et al., 2007). Because shorter than expected read lengths also correlate strongly with incorrect reads (Brockman et al., 2008), another three percent of the sequences (whose lengths were smaller than 100 bp) were removed. The remaining reads were aligned using CAP3 (Huang, Madan, 1999). Quality scores were rescaled to be comparable to the usual Phred Score using ARACHNE (Batzoglou et al., 2002).

SNPs were called at both the individual level and population level. At the individual level, SNPs were called using both a Bayesian method and a likelihood ratio test (LRT) method. For the Bayesian method, $10^{-4}$ was used as the prior for the mutation rate (Duvernell et al., 2008). At the population level, for each locus on the contig, we simulated the error model and marked a locus as a potential SNP if it had a larger number of second alleles in
comparison to the critical value from the error model. Furthermore, a potential SNP site had to have at least three individuals sequenced to 2X at that locus unless another potential SNP site was within five basepairs or over 90% of the individuals had been classified as heterozygous at the individual level. This was done to minimize the rate of false positives caused by homologs.

**Bayesian and LRT model for SNP calling at individual level**

For the Bayesian model, for each contig, $Prior = 1 \times 10^{-4}$ represents the mutation rate; $N$ represents the total number of unique mapping loci with multiple allelic types; $A^i$ and $a^i$ represent, respectively, the major and minor alleles at locus $i$; $N_i$ represents the total number of alleles observed for locus $i$, and $Y_j$ is the type of the $j^{th}$ allele copy among these $N_i$ alleles where $j = 0 \cdots N_i$; finally, $e_j$ is the probability of error of the $j^{th}$ allele where the error probability is computed as $10^{Q}$ and where $Q$ is the corresponding quality score after rescaling.

The posterior probability for the $i^{th}$ locus being homozygous or heterozygous is:

$$P(Hetero. | data) = \frac{P(data|Hetero.) \times P(Hetero.)}{P(data)}$$

$$\sim P(data | hetero) \times P(Hetero.)$$

$$\sim Prior \times (0.5)^N$$

$$P(Homo. | data) = \frac{P(data|Homo.) \times P(Homo.)}{P(data)}$$

$$\sim P(data | homo) \times P(Homo.)$$
Based on the posterior probabilities from above, we classified each of these N loci as homozygous or heterozygous exclusively. If a locus was classified as heterozygous, it was further tested using a likelihood ratio test (LRT) as follows:

For a particular locus $i$ on the contig:

$$P(X_j = A') = p$$
$$P(X_j = a') = 1 - p$$

where $X_j$ stands for the true allele that we should have observed. For each $Y_j$, we have an error probability of $e_j$ associated with it.

Then we have:

$$P(Y_j = A' \mid X_j = A') = 1 - e_j$$
$$P(Y_j = a' \mid X_j = a') = 1 - e_j$$
$$P(Y_j = a' \mid X_j = A') = e_j$$
$$P(Y_j = A' \mid X_j = a') = e_j$$

Therefore we have:

$$P(Y_j = A' \mid X_j = A') = (1 - e_j) \times p + e_j \times (1 - p)$$
$$P(Y_j = a' \mid X_j = A') = e_j \times p + (1 - e_j) \times (1 - p)$$

and

$$Y_j \sim Bernoulli(1, (1 - e_j) \times p + e_j \times (1 - p))$$
Based on all of the above, the likelihood of locus $I$ was computed as:

$$L = \prod_{j=1}^{N} \left\{ (-e_j) \times p + e_j \times (1-p) \right\}^{I_j} \left\{ e_j \times p + (1-e_j) \times (1-p) \right\}^{L_j}$$

Where $I_j = 1$ if $Y_j = A_i$; and $I_j = 0$ if $Y_j = a_i$

The LRT was performed with the hypothesis of $H_0: p = 0.5$ versus $H_a: p > 0.5$ and

$$-2 \times LRT \sim \chi^2(1).$$

**Error model simulating**

In order to call SNPs at the population level, we simulated the error model for each locus with multiple allelic types; we assumed that a particular locus was homozygous with major allele $A^i$ and randomly simulated $N_i$ number of alleles copies to be $A^i$ or any of the other three allele types from a uniform distribution with probability $(1 - e_j)$ and $e_j$ respectively. We repeated this process 10,000 times and recorded the different numbers of second alleles found in the simulation. The critical value was chosen as the number of second alleles with a right-side p-value of 0.001.

**Validation of SNPs**

Multiplex assays targeting 458 SNPs in 250 *F. heteroclitus* individuals, 90 *F. grandis* individuals, 23 *F. majalis* individuals, and 21 *F. similis* individuals were attempted using the Sequenom MassARRAY technology. These consisted of 81 putative SNPs identified by the *F. heteroclitus* pyrosequencing, 350 putative SNPs previously identified in *F. heteroclitus* ESTs (Quackenbush *et al*., 2000), and 27 putative SNPs from 22 genes containing, amongst others, SNPs in the aryl hydrocarbon receptor (Hahn *et al*., 2004), lactate dehydrogenase B
Assays were designed using the MassARRAY Assay Design Software with the goal of maximizing multiplexing of 36 SNPs per well (Sequenom, San Diego, CA, USA). Only SNPs where 70 base pairs were annotated on either side of the polymorphism were included in the study. There were 14 SNPs previously identified with 454 pyrosequencing where this criterion was not met. If multiple SNPs were proximal (< 70 base pairs) to one another, one SNP was chosen and the other(s) was translated into a degenerate nucleotide (e.g., K = G or T). Reaction conditions were performed by iPLEX chemistry as recommended by Sequenom across 13 plates at the University of Minnesota’s BioMedical Genomics Center. SNP genotypes were called using the Sequenom System Typer 4.0 Analysis package. This software uses a three-parameter model to calculate the significance of each putative genotype. Based on the relative significance, a final genotype is called and assigned a particular name (e.g., conservative, moderate, aggressive, user call). Non-calls also were noted (e.g., low probability, bad spectrum).

Analysis of Genotype Data

Arlequin v.3.11 was used to calculate genetic diversity among populations (of *F. heteroclitus* and *F. grandis*) by calculating the percentage of polymorphic SNPs ($P_O$), observed ($H_O$) and expected heterozygosity ($H_E$), and the within-population fixation index ($F$) (Excoffier *et al.*, 2005). Fixation index deviations from zero were tested by 10,000 permutations of alleles between individuals. Hardy-Weinberg equilibrium also was tested in each population. An analysis of molecular variance (AMOVA) was performed to calculate the distribution of variance within populations, between North and South regions, and
between *F. heteroclitus* populations within North and South regions. For *F. grandis*, the AMOVA was performed to calculate the distribution of variance within populations as well as between populations longitudinally along the Gulf of Mexico. Since SNPs were initially identified from *F. heteroclitus* sequence data, a maximum of 5% missing data was used as a parameter for calculations involving *F. heteroclitus* and 10% for all others.

A Mantel test was performed to assess the assumption of isolation by distance using XLSTAT 2009 for *F. heteroclitus* and *F. grandis*.

STRUCTURE v.2.2 (Falush *et al.*, 2003; Pritchard *et al.*, 2000) was used to estimate the number of populations (K) in *F. heteroclitus*, *F. grandis*, *F. majalis* and *F. similis* along both the Western Atlantic and the Gulf of Mexico and to assign individuals to these populations. The Monte Carlo Markov Chain was run for $10^5$ iterations following a burn-in period of $10^5$ iterations for K = 1 to 14 using the correlated allele frequencies model and assumed admixture. Distruct v. 1.1 (Rosenberg, 2004) was used to generate bar plots to depict classifications with the highest probability under the model. JMP Genomics 3.2 for SAS 9.1.3 conducted principal component analysis on all samples to establish population structure.

**Results**

**GS FLX Sequencing and Assembly**

A total of 111,001 reads were obtained in one run of the GS FLX instrument producing 5,346,445 total bases of sequence (average read length of 218 bases) with 99.98% of bases having a quality score of 20 or greater. Across the eight regions of the plate, there
were on average 1,982 reads per individual. The third barcode produced many less reads per region (<1,000) amongst all regions. All other barcodes performed very similarly with respect to the number of reads per individual across regions. Only 46% of the number of expected reads (111,001 instead of 240,000) were obtained from sequencing. Prior to sequencing, the amplification success of loci on the beads was checked for quality using a Bioanalyzer 2100, and all samples passed. However, three of the eight regions produced half the expected number of reads and a fourth region produced only 15% the expected number of reads. This indicated local problems in sequencing with respect to particular regions and the samples in those regions rather than the plate as a whole. All control beads passed the filter control with an average percentage of 90% across all regions, whereas the percentage of samples passing the filter control varied between regions and averaged 36%: regions with fewer than expected reads had fewer samples passing the filter control. Two regions had very high failure rates due to mixed samples, indicating more than one amplicon per bead.

Upon alignment 1,464 contigs were obtained with an average length of 213 base pairs. The average coverage across all loci was 22 reads per contig (Figure 3). Due to the low coverage of any one contig per individual, the detection of a SNP within a contig was mainly based on its presence across populations rather than at the individual level. Of the 1,464 contigs obtained, 96 contained SNPs. Within these contigs, 261 SNPs were identified. Among those contigs containing SNPs, the average length was 243 base pairs with an average coverage of 184 reads per contig (Figure 3). The observed rate of SNP detection is a function of depth, so as read counts per contig increased so did the number of SNPs detected. One third of all contigs with identified SNPs had only one SNP and 57% had two or fewer
SNPs per contig. SNPs were distributed approximately evenly along the position in the contig ($R^2 = 0.01$).

**Genotyping success**

Of the initial 458 loci we attempted to amplify, 281 had a greater than 90% successful call rate among all individuals with no more than two alleles per SNP. In *F. heteroclitus* 74.4% of all loci amplified in greater than 95% of individuals. In *F. grandis*, 11% of SNPs did not amplify, and 74% of SNPs were monomorphic. 24% of the monomorphic SNPs in *F. grandis* also were monomorphic in *F. heteroclitus*, but for the alternative allele, indicating fixed differences between the two species.

On average, 80% (SD = ± 7.4%) of the putative SNPs identified with 454 pyrosequencing were amplified with MassARRAY in *F. heteroclitus* individuals: 72 of the 81 loci (88.8%) were polymorphic, 8 loci (9.8%) were monomorphic, and one locus did not amplify. Among all other putative SNPs genotyped with MassARRAY, 83% were successfully amplified. However, 13.5% of all loci in *F. heteroclitus*, 25.6% in *F. grandis*, 27.1% in *F. majalis* and 39.3% in *F. similis* did not amplify (Figure 4a). Many non-*heteroclitus* loci were also not polymorphic, and in *F. heteroclitus* 12.3% of all loci were monomorphic, as were 58.2% in *F. grandis*, 26.4% in *F. majalis*, and 29.7% in *F. similis* (Figure 4b). Due to the divergence between species resulting in unsuccessful amplification in non-*heteroclitus* individuals, locus amplification success was addressed on a species and population level for all remaining tests and not on the overall amplification success. Due to the low sample size, amplification rate, and predominant monomorphism of loci in *F. majalis* and *F. similis* samples, further characterizations of genetic parameters (with the exception of
population structure) were not carried out for these two species.

SNPs which were identified by Sequenom software as low probability in greater than 50% of all individuals were removed (17 SNPs in total). An additional 20 SNPs were excluded from analyses due to their excessive heterozygosity across individuals and populations of *F. heteroclitus*. These SNPs may represent segmental duplication where the two duplicate regions are identical, except that a SNP has been driven to high frequency or become fixed in one of the duplicates.

**Genetic Diversity**

The percentage of polymorphic SNPs (*P*<sub>O</sub>) ranged from 3.7% to 67% (Table 3) among populations and species. The percentages of polymorphic SNPs were significantly different between northern and southern populations of *F. heteroclitus* where levels decreased in populations further north and east (*p*=0.035). Among *F. grandis* populations, the percentages of polymorphic SNPs did not significantly differ along latitude (*p*= 0.143) or longitude (*p*= 0.415). Among populations, most loci were in Hardy-Weinberg equilibrium (Table 3). Observed heterozygosity (*H*<sub>O</sub>) among all populations ranged from 0.016 to 0.17 with a mean of 0.10 (Table 3). Observed heterozygosity was lower in northern *F. heteroclitus* in comparison to southern populations (*p*=0.04) and did not differ along latitude (*p*= 0.72) or longitude (*p*= 0.33) in *F. grandis*. Average expected heterozygosity (*H*<sub>E</sub>) ranged from 0.019 to 0.20 with a mean of 0.11 (Table 3). The average within-population fixation index, *F*, averaged over all polymorphic loci was on average 0.16 in *F. heteroclitus* and 0.20 in *F. grandis* (Table 3).

SNPs identified *via* 454 sequencing did not have genetic parameters that differed
from SNPs identified in ESTs with the exception of Hardy-Weinberg equilibrium. 454-derived SNPs had a higher percentage of SNPs not in Hardy-Weinberg equilibrium due to a lack of heterozygosity (22% versus 9%).

Many SNP loci (60%) in *F. heteroclitus* had a frequency greater than 0.10 and were considered common SNPs (Table 4). In contrast, 90% of SNPs in *F. grandis* had low minor allele frequencies below 0.10.

**Population Structure**

The two independent tests of population stratification (STRUCTURE and principle component analysis (PCA)) identified species and population differences in all samples (Figure 5). STRUCTURE analysis, which uses a Bayesian MCMC clustering approach to assign individuals to clusters, separated populations into eight different clusters (Pr(K) = 0.37; Figure 4a). At the most probable clustering of the data (K=8), ten runs produced nearly identical membership coefficients which had pairwise similarity coefficients greater than 0.98. *F. heteroclitus* clustered north to south and *F. grandis* as its own separate cluster. Among *F. heteroclitus*, individuals from Maine and Georgia, the most northern and southern collection sites, formed their own distinct clusters. Individuals from sites between Maine and Georgia clustered with others from geographically similar sites. *F. majalis* and *F. similis* clustered together and away from the other two species. Similarly, in the PCA analysis, which does not rely on modeling the data, northern and southern *F. heteroclitus* stratified by latitude and were distinct from each other (Figure 5b) and each *F. heteroclitus* population was clustered together (Figure 5c). *F. grandis* made its own cluster and *F. majalis* and *F. similis* clustered together apart from other species (Figure 5b).
In *F. heteroclitus*, AMOVA showed that most of the variation was distributed within populations (59.05%), but another large proportion of variation (31.1%) was distributed among northern and southern regions. The remaining 9.85% of variation was explained by differences among populations within regions. In *F. grandis*, most of the variation was distributed within populations (82.4%), and a smaller proportion (17.6%) of variation was distributed longitudinally between populations across the Gulf of Mexico.

A Mantel test showed significant isolation by distance among *F. heteroclitus* populations (p< 0.001) and *F. grandis* populations (p=0.032).

**Discussion**

We used high throughput sequencing and genotyping technology to identify and verify SNP markers in four non-model species within the *Fundulus* genus. Genotype data sharply differentiated northern and southern populations of *F. heteroclitus* as well as other species in this genus (*F. grandis*, *F. majalis*, and *F. similis*). Within the species where SNPs were originally annotated, most can be successfully verified and used to study population structure as well as the role and outcome of selection forces on a genome-wide scale.

Using the 454 FLX pyrosequencing system, we observed 111,001 reads yielding an average of 22x coverage across 1,464 contigs. Read lengths and quality scores were similar to many other studies using the 454 FLX system to sequence uncharacterized genomes (Novaes *et al.*, 2008; Wiedmann *et al.*, 2008), but we identified fewer SNPs. Two-hundred and sixty-one SNPs were identified in 96 of these contigs (81 were further verified with the Sequenom MassARRAY platform). The percentage of contigs containing SNPs did differ
between experiments: we obtained 0.07% of contigs containing SNPs while pyrosequencing of *Eucalyptus* ESTs identified 0.05% of contigs containing SNPs (Novaes *et al.*, 2008) and pyrosequencing of size selected, genomic DNA from swine identified 11.4% of contigs contained SNPs (Wiedmann *et al.*, 2008).

Our 454 pyrosequencing of genomic DNA was originally designed to both discover and genotype SNPs within and among populations of *F. heteroclitus*. Thus, we attempted to perform genome reduction with selective PCR reactions to approximately 200 loci that could be sequenced in 10 populations of 8 individuals. With 30,000 reads per one-eighth of a 454 sequencing plate, each region would have 15X coverage per individual or 980X coverage across all populations, enabling accurate genotype calls for most individuals. However, preselective amplification was not perfect, and many more than 200 loci were sequenced; most amplified only a single time in a single individual (these singlets therefore were not useful for variant detection). Furthermore, we obtained only 46% of the expected number of reads. In the end, these problems led to the inability to directly call individual genotypes. We were hoping to both identify SNPs and genotype individuals in a single step, but a more successful approach (as evidenced by the swine group (Wiedmann *et al.*, 2008)) is to make reduced representation libraries from many pooled individuals for SNP discovery followed by individual genotyping. Because a pool of individuals is used, this approach identifies few singlets and thus enhances the number of reads per contig. Furthermore, improvements in both the number and length of reads using the Titanium series FLX 454 system compared to the original FLX system we used will increase the number of identified SNPs.

To increase our ability to measure population genetic parameters within and among
populations, we verified SNPs identified through 454 sequencing and additional SNPs annotated from *F. heteroclitus* cDNAs using the MassARRAY system. Similar percentages of 454 pyrosequencing derived SNPs and SNPs identified from ESTs were verified (80% and 83%, respectively). Of the 458 putative SNPs, 379 (82.75%) were polymorphic, but only 264 had a greater than 90% successful call rate among all individuals. Among *F. heteroclitus*, most SNPs amplified (61.3% were called in >95% of individuals) indicating that differences in amplification rate between species led to the lower overall call rate. In white spruce, 91% of SNPs verified with the Illumina SNP bead array platform (Fan *et al.*, 2003; Shen *et al.*, 2005) were true. Comparable to *F. heteroclitus*, 70% of SNPs in spruce were called in greater than 95% of individuals (Namroud *et al.*, 2008). Overall, verification of SNPs was powerful in providing information over many markers and individuals and was able to provide data to determine differences within populations, between populations and between species.

Species differentiation was demonstrated using principle component analysis (PCA) as well as STRUCTURE analysis. Both analyses showed separation between *F. heteroclitus*, *F. grandis* and *F. majalis* and *similis* as well as population structure within *F. heteroclitus* (Figure 4). These analyses provided the most resolution (even among distinguishing populations) in *F. heteroclitus* because the SNPs were originally identified in this species (*i.e.*, due to an ascertainment bias). PCA and STRUCTURE did not differentiate sister species, *F. similis* and *F. majalis*, from each other or establish population structure within these species. Small sample sizes (1 to 10 individuals per population), high levels of monomorphism (average of 28% of all SNPs), and the fact that only 10% of SNP alleles
differed between these two species, decreased the power to detect such differences when analyzed in conjunction with *F. heteroclitus* and *F. grandis*. Population structure also was masked in *F. grandis* when data was analyzed with other species. However, when *F. grandis* individuals were analyzed separately, they also showed distinct population structure (data not shown). One other study has reported multiple fixed differences in mitochondrial sequences between *F. heteroclitus* and *F. grandis* (Whitehead, 2009), but no other study to date has evaluated differences at many loci between all four species used in this study.

Within *F. heteroclitus* and *F. grandis* species, within-population fixation indices (F$_{IS}$, averaged across all loci) ranged from 0.09 to 0.32. Among *F. heteroclitus*, all populations had an overall significant deficiency of heterozygotes indicated by positive F$_{IS}$ values. In these populations, approximately 10% of loci had similarly very large F$_{IS}$ values (>0.5) across populations causing the skew in the average F$_{IS}$ value for each population. Within a population, these loci were predominately homozygous for one allele with a complete absence of the heterozygote and one or a few individuals homozygous for the alternative allele. The loci which presented this pattern were called conservatively at both alleles by Sequenom software across all individuals indicating that genotyping error was not the main reason for this pattern. Furthermore, all northern populations were predominately homozygous for one allele and all southern populations were predominately homozygous for the alternative allele indicating strong demographic patterns in the data. The same demographic pattern was not found in *F. grandis*. Among *F. grandis* populations, most (70%) SNPs with high F$_{IS}$ values were different between populations. This is in contrast to *F. heteroclitus* populations where loci with high F$_{IS}$ values were shared across populations.
Within any one *F. grandis* population, one allele was predominant as a homozygote with one or a few individuals with the alternative homozygote. The most parsimonious explanation is that there is undetected substructure.

SNPs in Hardy-Weinberg were shown to be moderately polymorphic (average of 60%) in *F. heteroclitus*. In *F. grandis*, SNPs were shown to lack polymorphism (7.18%). The higher percentage of monomorphic loci in *F. grandis* likely is due to ascertainment bias in SNP discovery caused by only using *F. heteroclitus* populations. Many of the monomorphic loci (24%) represent fixed differences between *F. heteroclitus* and *F. grandis*. Thus, while SNP markers developed in *F. heteroclitus* are not necessarily polymorphic in other *Fundulus* species, they still can be used to differentiate *F. heteroclitus* from other species.

Among *F. heteroclitus* populations, genotype data revealed strong latitudinal clines between the Northern and Southern *F. heteroclitus* populations. PCA, STRUCTURE, *F*<sub>ST</sub> values, and the isolation by distance test identified that individuals from Northern populations (above 40-41°N) were distinct from Southern populations. This split is centered around the southern-most extent of the Atlantic coastal advancement during the late Pleistocene (Mickelson DM, 1983). Specifically, observed heterozygosity and allelic richness across all loci is significantly lower (p=0.043, p=0.042, respectively) in the north than in the south. These differences have been shown previously in morphological features (Able, Felley, 1986) numerous allozyme loci (Cashon *et al.*, 1981; Powers, Place, 1978; Powers *et al.*, 1986; Ropson *et al.*, 1990) and microsatellites (Adams *et al.*, 2006). The larger historical population size of *F. heteroclitus* in the south (Adams *et al.*, 2006) would...
maintain greater heterozygosity and allelic richness at shared loci; in the north, where population sizes are smaller, loci have a higher probability of becoming fixed.

Four STRUCTURE clusters encompass the six northern populations while only two clusters encompass the five southern populations (Figure 5A). Separate northern clusters may be driven by smaller population sizes in which drift is greater. When genetic drift has a larger effect it becomes easier to distinguish populations because the average difference in allele frequencies of a marker in different populations will be greater. This is illustrated by a larger average $F_{ST}$ of 0.20 among northern populations in comparison to that of an average $F_{ST}$ of 0.10 among southern populations. This statistic is also evident for the north and south split, where populations from respective regions had an extremely high $F_{ST}$ value of 0.44 when compared against one another. Similar genetic divergence has been reported for *F. heteroclitus* using microsatellites (0.196 among northern populations, 0.117 among southern populations and 0.330 for the two most divergent populations, Nova Scotia and Georgia (Adams et al., 2006)). Similar demographic patterns have been described in freshwater fish (Bernatchez, Wilson, 1998) and marine species such as goby (Gysels et al., 2004) and blue crab (McMillen-Jackson, Bert, 2004), and, as in *Fundulus*, these patterns are attributed to Pleistocene events.

A similar latitudinal cline occurs between populations of *F. grandis*, and a Mantel test shows significant isolation by distance. However, there were no significant differences between either levels of polymorphism or observed heterozygosity along latitude or longitude. Williams et al., 2008 reported significant isolation by distance as well as decreased allelic richness with increasing latitude. In this 2008 study, microsatellites were
used, and two additional sites southern to those used in our study were included. Since microsatellites have many more alleles than SNPs and two additional sites were found to have relatively higher allelic richness in comparison to all other sampling sites along the gulf, this may account for the differences found in levels of polymorphism.

**Conclusions**

By targeting SNPs contained in both coding and non-coding areas of the genome, we are able to better understand how evolutionary forces are shaping the *Fundulus* genome. Similar studies using high throughput methods to sequence SNP markers have been developed in Atlantic cod (Moen *et al.*, 2008), white spruce (Namroud *et al.*, 2008), *Eucalyptus* (Novaes *et al.*, 2008), and swine (Wiedmann *et al.*, 2008). Like our study, these studies expanded their own species’ knowledge base with respect to potential markers for studying evolutionary adaptation (in the case of cod and spruce), genome-wide assessment of diversity (*Eucalyptus*) or for use in breeding programs (swine).
Acknowledgments

The authors thank G. Bozinovic and M. Everett for assistance in the collection of samples and D. Crawford for valuable input into methodology. *Part of this work was carried out by using the resources of the Computational Biology Service Unit from Cornell University which is partially funded by Microsoft Corporation.* Funding was partially provided by NIEHS Training Grant ES007046 award from the Department of Environmental and Molecular Toxicology at North Carolina State University to LMW, NIH 5 R01 ES011588 to MFO, NSF DEB0948510 to ARB, and NIH R01 HG003229 CDB.
References


<table>
<thead>
<tr>
<th>Adapters</th>
<th>BspEI (5’ to 3’)</th>
<th>EcoRI (5’ to 3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GACGATGAGTCCTGAGC</td>
<td>CTGAGTCCTAGTAGCACCTCGTAGACTGCGTACC</td>
</tr>
<tr>
<td></td>
<td>CCGGGCTCAGGACTCATCGTC</td>
<td>AATTGGTACGAGTCTCAC*</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Preselective Primers</th>
<th>EcoRI (5’ to 3’)</th>
<th>BspEI (5’ to 3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CTGAGTCTAGTGACACC</td>
<td>GACGATGAGTCCTGAGC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Selective Primers</th>
<th>EcoRI (5’ to 3’)</th>
<th>BspEI (5’ to 3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>GACTGCGTACCAATTCAAG</td>
<td>GACGATGAGTCCTGAGCC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Barcoded Primers</th>
<th>EcoRI (5’ to 3’)</th>
<th>BspEI (5’ to 3’)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GCCTCCCTCGCGCCATCAGAGCCTAAGCTGACTGCGTACCAATTCAAG</td>
<td>GCCTTGCCAGCCCGCTCAGGACGATGAGTCCTGAGCC</td>
</tr>
<tr>
<td>2</td>
<td>GCCTCCCTCGCGCCATCAGAGTCTCGACTGCGTACCAATTCAAG</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>GCCTCCCTCGCGCCATCAGAGTCTCGACTGCGTACCAATTCAAG</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>GCCTCCCTCGCGCCATCAGAGTCTCGACTGCGTACCAATTCAAG</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>GCCTCCCTCGCGCCATCAGAGTCTCGACTGCGTACCAATTCAAG</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>GCCTCCCTCGCGCCATCAGAGTCTCGACTGCGTACCAATTCAAG</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>GCCTCCCTCGCGCCATCAGAGTCTCGACTGCGTACCAATTCAAG</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>GCCTCCCTCGCGCCATCAGAGTCTCGACTGCGTACCAATTCAAG</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>GCCTCCCTCGCGCCATCAGAGTCTCGACTGCGTACCAATTCAAG</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>GCCTCCCTCGCGCCATCAGAGTCTCGACTGCGTACCAATTCAAG</td>
<td></td>
</tr>
</tbody>
</table>

**Table 1.** Adapters and primers used in the amplification of genomic DNA. Star indicates location of 2’ O-methyl block.
Table 2. Genotyping success of SNP markers using the MassARRAY multiplex assay

<table>
<thead>
<tr>
<th>Category</th>
<th>Number of SNPs</th>
<th>Percentage of SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>SNPs called in &gt;95% of <em>F. heteroclitus</em> individuals</td>
<td>259</td>
<td>61.4</td>
</tr>
<tr>
<td>SNPs called in &lt;80% of all individuals</td>
<td>135</td>
<td>31.9</td>
</tr>
<tr>
<td>SNPs called in &gt;90% but &lt;95% of all individuals</td>
<td>101</td>
<td>23.9</td>
</tr>
<tr>
<td>Monomorphic SNPs called in &gt;95% of all individuals</td>
<td>23</td>
<td>5.4</td>
</tr>
<tr>
<td>Polymorphic SNPs called in &gt;95% of all individuals</td>
<td>163</td>
<td>38.6</td>
</tr>
<tr>
<td>SNPs called in &lt;90% of all individuals identified in 454</td>
<td>35</td>
<td>43.2</td>
</tr>
<tr>
<td>SNPs called in &gt;90% of all individuals identified in 454</td>
<td>46</td>
<td>56.8</td>
</tr>
</tbody>
</table>
Table 3. Genetic parameters of sampled populations in two species of *Fundulus*. † p≤ 0.01 based on 10,000 permutations between individuals within the same populations.

**Fundulus heteroclitus**

<table>
<thead>
<tr>
<th>Population</th>
<th>P₀</th>
<th>H₀</th>
<th>Hₑ</th>
<th>F</th>
<th>% Departure from HWE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maine</td>
<td>33</td>
<td>0.08</td>
<td>0.09</td>
<td>0.13†</td>
<td>7.0</td>
</tr>
<tr>
<td>Sandwich</td>
<td>48</td>
<td>0.12</td>
<td>0.14</td>
<td>0.13†</td>
<td>9.3</td>
</tr>
<tr>
<td>New Bedford Harbor</td>
<td>57</td>
<td>0.13</td>
<td>0.15</td>
<td>0.12†</td>
<td>7.9</td>
</tr>
<tr>
<td>Point Judith</td>
<td>44</td>
<td>0.11</td>
<td>0.13</td>
<td>0.18†</td>
<td>10.8</td>
</tr>
<tr>
<td>Clinton</td>
<td>59</td>
<td>0.12</td>
<td>0.13</td>
<td>0.09†</td>
<td>6.0</td>
</tr>
<tr>
<td>Newark</td>
<td>65</td>
<td>0.17</td>
<td>0.19</td>
<td>0.11†</td>
<td>6.4</td>
</tr>
<tr>
<td>Tuckerton</td>
<td>67</td>
<td>0.16</td>
<td>0.21</td>
<td>0.25†</td>
<td>12.5</td>
</tr>
<tr>
<td>Magotha</td>
<td>66</td>
<td>0.17</td>
<td>0.20</td>
<td>0.17†</td>
<td>11.5</td>
</tr>
<tr>
<td>Elizabeth River</td>
<td>67</td>
<td>0.16</td>
<td>0.20</td>
<td>0.23†</td>
<td>12.3</td>
</tr>
<tr>
<td>Manteo</td>
<td>65</td>
<td>0.16</td>
<td>0.20</td>
<td>0.19†</td>
<td>12.9</td>
</tr>
<tr>
<td>Georgia</td>
<td>51</td>
<td>0.13</td>
<td>0.16</td>
<td>0.19†</td>
<td>13.2</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>56.54</td>
<td>0.14</td>
<td>0.16</td>
<td>0.16</td>
<td>9.98</td>
</tr>
<tr>
<td><strong>Standard Deviation</strong></td>
<td>11.28</td>
<td>0.03</td>
<td>0.04</td>
<td>0.05</td>
<td>2.76</td>
</tr>
</tbody>
</table>

**Fundulus grandis**

<table>
<thead>
<tr>
<th>Population</th>
<th>P₀</th>
<th>H₀</th>
<th>Hₑ</th>
<th>F</th>
<th>% Departure from HWE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weeks Bay</td>
<td>9.0</td>
<td>0.016</td>
<td>0.019</td>
<td>0.13†</td>
<td>1.1</td>
</tr>
<tr>
<td>Dauphin Island</td>
<td>5.9</td>
<td>0.016</td>
<td>0.024</td>
<td>0.23†</td>
<td>2.8</td>
</tr>
<tr>
<td>Leeville</td>
<td>5.9</td>
<td>0.017</td>
<td>0.020</td>
<td>0.32†</td>
<td>2.0</td>
</tr>
<tr>
<td>Hackberry</td>
<td>10.1</td>
<td>0.023</td>
<td>0.032</td>
<td>0.10</td>
<td>3.0</td>
</tr>
<tr>
<td>Port O’Connor</td>
<td>8.5</td>
<td>0.018</td>
<td>0.026</td>
<td>0.27†</td>
<td>4.2</td>
</tr>
<tr>
<td>Port Aransas</td>
<td>3.7</td>
<td>0.021</td>
<td>0.031</td>
<td>0.23†</td>
<td>2.1</td>
</tr>
<tr>
<td><strong>Mean</strong></td>
<td>7.18</td>
<td>0.02</td>
<td>0.03</td>
<td>0.20</td>
<td>2.53</td>
</tr>
<tr>
<td><strong>Standard Deviation</strong></td>
<td>2.4</td>
<td>0.003</td>
<td>0.005</td>
<td>0.11</td>
<td>1.06</td>
</tr>
</tbody>
</table>
Table 4. SNP minor allele frequencies (MAF) within *F. heteroclitus* and *F. grandis* populations.

<table>
<thead>
<tr>
<th></th>
<th>Average MAF</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># SNP</td>
<td>Including Monomorphic</td>
<td>Excluding Monomorphic</td>
<td>Monomorphic</td>
<td>0≤MAF≤0.1</td>
<td>0.1≤MAF≤0.2</td>
<td>0.2≤MAF≤0.3</td>
<td>0.3≤MAF≤0.4</td>
<td>0.4≤MAF≤0.5</td>
</tr>
<tr>
<td><em>F. heteroclitus</em></td>
<td>398</td>
<td>0.2</td>
<td>0.25</td>
<td>0.21</td>
<td>0.19</td>
<td>0.11</td>
<td>0.16</td>
<td>0.15</td>
<td>0.18</td>
</tr>
<tr>
<td><em>F. grandis</em></td>
<td>410</td>
<td>0.03</td>
<td>0.13</td>
<td>0.74</td>
<td>0.16</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td><strong>Overall</strong></td>
<td>404</td>
<td>0.12</td>
<td>0.19</td>
<td>0.48</td>
<td>0.18</td>
<td>0.07</td>
<td>0.09</td>
<td>0.09</td>
<td>0.11</td>
</tr>
</tbody>
</table>
Figure 1. Sampling sites for *Fundulus* species. *F. heteroclitus* was collected along the east coast of the United States and *F. grandis* was collected along the Gulf of Mexico coast.
Figure 2. Design of 454 pyrosequencing contig generated from the digestion of genomic DNA with restriction enzymes (EcoRI and BspEI), the addition of restriction site specific linkers, an individual barcode and a 454 amplicon adapter.
Figure 3. Contig totals *versus* number of reads per contig amongst those contigs with identified SNPs (bars) and all contigs (squares).
Figure 4. Non-amplified and non-polymorphic loci among Fundulus species. (A) Numbers of loci, which did not amplify with the MassARRAY platform among the four species of Fundulus. Not shown: loci shared between F. majalis and F. similis (8) and F. heteroclitus and F. grandis (12). (B) Numbers of loci, which were not polymorphic among the four species. Not shown: loci shared between F. majalis and F. similis (9) and F. heteroclitus and F. grandis (1).
Figure 5. (A) Population structure as assessed by STRUCTURE. Bar plot was generated by DISTRUCT and depicts the classifications of the populations with the highest probability under the model. K indicates the number of clusters that maximized the probability of the model. Each individual is shown as a vertical bar. (B) Principal components PC1, PC2 and PC3 from all SNPs (as calculated in JMP Genomics 3.2) among all individuals. Species are separated from each other as well as northern and southern *F. heteroclitus* populations. Colors represent different species. (C) Principal components PC1, PC2, and PC3 from all SNPs among *F. heteroclitus* individuals. Colors represent different populations.
CHAPTER 3

Ecologically and Evolutionarily Important SNPs Identified in Natural Populations

Larissa M. Williams\(^1\) and Marjorie F. Oleksiak\(^2\)

1. Department of Environmental and Molecular Toxicology
   Box 7633, North Carolina State University
   Raleigh, NC 27695-7633 USA

2. Rosenstiel School of Marine and Atmospheric Sciences
   University of Miami
   4600 Rickenbacker Causeway
   Miami, FL 33149 USA

Corresponding Author:
Marjorie Oleksiak
Rosenstiel School of Marine and Atmospheric Sciences
University of Miami
4600 Rickenbacker Causeway
Miami, FL 33149
Fax: 305-421-4600
Email: moleksiak@rsmas.miami.edu
Abstract

Evolution by natural selection acts on natural populations amidst migration, gene-by-environmental interactions, constraints and tradeoffs, which affect the rate and frequency of adaptive change. We asked how many and how rapidly loci change in populations subject to severe, recent, environmental changes. To address these questions, we used genome-wide association approaches to identify SNPs with evolutionarily significant patterns in natural populations of *Fundulus heteroclitus* that inhabit and have adapted to highly polluted Superfund sites. Three statistical tests identified SNPs evolving by natural selection in three independent populations adapted to pollution: 1.6–4% of loci were significantly different from the neutral model in these populations. One SNP, in the xenobiotic metabolizing enzyme, cytochrome P4501A (CYP1A), was identified in all polluted populations using all tests. Extrapolating across the genome, these data suggest that rapid evolutionary change in natural populations will involve hundreds of loci, a few of which will be shared in independent events.
Introduction

Haldane estimated the mean rate of gene substitution as one per 300 generations (Haldane, 1990). Yet, due to human intervention, the global environment is changing rapidly (Kerr, 2007; Vitousek et al., 1997) making adaptation in the 300 generations envisioned by Haldane unsustainable. One species, which has adapted in many fewer generations, is the estuarine fish *Fundulus heteroclitus*: it has adapted to anthropogenic contaminants in less than 15 generations (Nacci et al., 1999) in at least three separate geographical locations along the east coast of the United States. These three populations are exposed to some of the highest concentrations of aromatic hydrocarbon pollutants of any vertebrate species (Wirgin, Waldman, 2004) and inhabit highly polluted Superfund sites (Elksus et al., 1999; Meyer J, 2002; Nacci et al., 1999; Ownby et al., 2002), hazardous waste sites mandated for clean up by the Comprehensive Environmental Response, Compensation, and Liability Act of 1980. Such rapid adaptation indicates a strong selective force. Genome wide association (GWA) studies were used to identify SNPs associated with these polluted environments and thus provide an approach to investigate and understand rapid adaptive changes.

We sought to identify adaptive molecular variations that appear to be evolutionarily important within and among populations subject to anthropogenic stress using three statistical tests: an $F_{ST}$ modeling approach, an association test, and a test on allele frequencies (MAF-$F_{MAX}$). We genotyped 367 SNPs from both coding and non-coding regions in 180 individuals from nine populations (Fig. 1). These populations have large effective population sizes and high genetic variation (Adams et al., 2006). SNPs in each polluted population were compared with those in two flanking reference site populations located north and south of
each polluted site (triads, Figure 1). Populations within triads show little to moderate
differentiation based on multilocus microsatellite estimates of $F_{ST}$, which range from 0.043 to
0.101 (Adams et al., 2006). Our experimental design distinguishes pollutant effects from
demographic ones because the genetic distance between the two clean reference populations
is greater than the genetic distance between the polluted population and either reference
population (Figure 1). Thus, the variation due to demography can be accounted for between
the two distant reference populations, and divergence in a polluted population compared to
both paired reference populations suggests that the contamination or other environmental
factors associated only with the Superfund sites serves as the causative selective force.

**Materials and Methods**

*F. heteroclitus* were collected using minnow traps during the spring of 2005. Fin
clips were sampled from 20 individuals from each of the nine collection sites along the
Atlantic Coast of the United States (Figure 1). Three of the nine collection sites were EPA
Superfund sites including New Bedford Harbor (EPA ID: MAD980731335), Newark (EPA
ID: NJD980528996) and Elizabeth River (EPA IS: VAD990710410). To control for random
processes such as drift, fish also were collected from populations from clean, control sites
flanking each polluted site population.

Genomic DNA from fin clips was extracted using a modified version of Aljanabi and
Martinez (Aljanabi, Martinez, 1997), and DNA was resuspended in 50 µL 0.1X TE buffer.
We genotyped 180 *F. heteroclitus* at 458 SNPs using the MASSARRAY platform at the
University of Minnesota as described (Williams et al., 2010). We analyzed a subset of these
SNPs (367): SNPs that amplified in greater than 80% of *F. heteroclitus* and did not show an excess of heterozygosity. SNPs were from both coding and non-coding loci: 295 were coding and 72 were non-coding.

Arlequin v.3.11 was used to calculate $F_{ST}$ values for each SNP within a population using the AMOVA function (Excoffier *et al.*, 2005). $F_{ST}$ values were modeled to detect outliers using the FDIST2 program (Beaumont, Nichols, 1996). Simulations were run for each pair of populations using the average heterozygosity of the empirical data with 20,000 iterations assuming 10 demes, 2 populations, 20 individuals per sample and a stepwise mutation model. The 99th percentile of simulation values was plotted against empirical data to determine the range of $F_{ST}$ values in the neutral model. Those empirical values which exceeded simulation values and were shared outliers between each of the pair wise comparisons (polluted *versus* both reference sites) were considered to be outliers and potentially under selection by pollution or linked to loci under selection.

We used JMP Genomics 3.2 for SAS 9.1.3 to conduct SNP case control trait association tests. Tests were used to identify SNPs associated with pollution (trait) using a chi-square test with the assumption that individuals are unrelated in recent generations. A second case control trait association test was used to determine whether the differences in allele frequencies were due to geographical differences between clean sites. A likelihood ratio test was used to determine which of the associations, that of the polluted or clean sites, was the best fit for the data. SNPs with $p$-values $< 0.01$ in the association test and likelihood $p$-values $< 0.01$ for the polluted site model were identified as outliers. A Bonferroni correction was applied to each triad to correct for multiple testing.
Allele frequencies for each SNP were compared in each triad to determine whether there were significant differences in allele frequencies in a polluted site versus both reference sites. In each triad, the minor allele was determined on a per SNP basis. Within each population, the allele frequency of that overall minor allele was calculated for a random sampling of 15 out of 20 individuals 100 times. One-hundred random samplings of 15 out of 20 individuals has less than 1% probability that the same combination of individuals will be chosen more than once. A one-way ANOVA was performed on the iteratively derived values. To control for type I errors among all 100 iterations, we computed multiple test-corrected critical values for the F-statistic by the one-step adjustment method (Westfall, Young, 1993). This test used a random sampling of individuals (assuming no population structure) in the iterative ANOVA process described above to calculate the maximum F-value for all ANOVAs (FMAX). Random sampling of individuals was carried out 1,000 times to determine the top 1% of the maximum F-statistics. F-statistics in the empirical data that exceeded the 1% FMAX values were considered to be outliers.

**Results and Discussion**

The first statistical test used an FST modeling approach: empirical FST values of each SNP were compared against the 99th quantile of simulated, neutral distributions of FST values along the range of heterozygosity values (Figure 2) (Beaumont, Nichols, 1996). Selective loci were identified as outliers in each Superfund site population relative to its two reference site populations: polluted versus both references, where the union of polluted versus reference 1 and polluted versus reference 2 but not comparisons between the reference site
populations had p-values <0.01 (Williams, Oleksiak, 2008). Thirty outlier loci (8.2%) were found in the New Bedford Harbor triad, 9 (2.5%) in the Newark Bay triad, and 43 (11.7%) in the Elizabeth River triad. While $F_{ST}$ tests relying on a simple island model of population differentiation have been shown to be robust, they are prone to a large excess of false positive loci when complex genetic structures exist (Excoffier et al., 2009). To further assess the strength of our outliers, we performed two additional statistical tests.

We did association tests for each SNP within each triad and calculated p-values based on the strength of association with polluted populations (red line; Figure 3) or reference populations (blue line; Figure 3). We then used a likelihood ratio test to determine whether the polluted or reference model was a better fit of the data. In the New Bedford Harbor triad, 30 SNPs (8.2%) were significantly associated with the polluted model ($p \leq 0.01$). In the Newark Bay triad, fewer SNPs were associated with the polluted model (7; 1.9%). In the Elizabeth River triad, 29 SNP (7.9%) were associated with the polluted model. This association test detected many demographic effects in the Newark Bay triad (blue line, Figure 3), which is located at the historical introgression zone between northern and southern $F. heteroclitus$ populations (Adams et al., 2006; Bernardi et al., 1993; Haney et al., 2009). This large demographic effect may mask some of the effects of directional selection due to pollution.

We used a third test (MAF-$F_{\text{MAX}}$ test) to determine whether the minor allele frequency (MAF) of each SNP was significantly different in a polluted site versus both the reference sites. For this test, we sampled 15 of the 20 individuals in each population 100 times and calculated the MAF of those subsets of individuals (15 polluted individuals and 30
reference individuals) on a SNP by SNP basis. Analysis of variance (ANOVA) was used on these 100 iterations to identify SNPs with MAF significantly different between the polluted and reference populations (Figure 4). An $F_{\text{MAX}}$ (Westfall, Young, 1993) was used to control for type I errors among all 100 iterations: the distribution of F-values from permutations of the data assuming random population differentiation was used to determine the critical F-values that occur less than 1% of the time among all permuted F-values ($1\% F_{\text{MAX}}$ values (Westfall, Young, 1993)). In the New Bedford Harbor population, 22 SNPs (6.0%) had significantly different MAF between polluted and both reference populations (Figure 4). In the Newark Bay and Elizabeth River populations, 9 and 18 SNPs (2.5% and 4.9%, respectively) had significantly different MAF between polluted and reference populations (Figure 4).

The $F_{\text{ST}}$ modeling approach, association test, and MAF-$F_{\text{MAX}}$ test all identified similar percentages of SNPs with non-neutral patterns in all three triads. However, the MAF-$F_{\text{MAX}}$ test often identified the fewest number of SNPs with non-neutral patterns (22, 9, and 18 for New Bedford Harbor, Newark Bay and Elizabeth River triads, respectively versus 30, 9 and 43 for the $F_{\text{ST}}$ modeling approach and 30, 7 and 29 for the association test in these three triads). The MAF-$F_{\text{MAX}}$ test asks whether a specific allele (the minor allele of the total population) occurs more frequently in the polluted population versus the flanking reference populations. This test differs from the association test, which associates an allele with the polluted population. Both tests differ from the $F_{\text{ST}}$ modeling approach, which is based on differences in heterozygosities between populations. These tests use different aspects of the polymorphism spectrum measured by SNPs (minor allele frequency, frequency of each allele
and heterozygosity).

SNPs identified as outliers (exhibiting non-neutral behavior) in more than one test are statistically more powerful and less likely to suffer from type I errors. Within each triad, 6-15 SNPs were identified as outliers in all three tests: the New Bedford Harbor triad had 12, the Newark Bay triad had 6, and the Elizabeth River triad had 15 (Figure 1b, Supplemental Table S1). Among all triads, 15 of these SNPs occur in coding regions. Only one of these 15 SNPs, a SNP in $\beta_2$-microglobulin, is non-synonymous. $\beta_2$-microglobulin is vital to the immune response and is noncovalently associated with the heavy chain of the major histocompatibility complex (MHC) class I antigens (Bernabeu et al., 1984). The G to A SNP in $\beta_2$-microglobulin changes an aspartic acid residue an asparagine predominantly in the polluted Newark Bay population. The other 14 SNPs that occur in coding regions all result in synonymous SNPs. Synonymous SNPs could affect mRNA splicing, translation or stability or could simply be linked to causative genetic polymorphisms. This can only be determined by functional, locus specific tests. Among the remaining annotated genes (Paschall et al., 2004), five SNPs are in 3’ untranslated regions and two are in 5’ upstream regions. These SNPs potentially affect RNA stability and transcription.

If one assumes that the loci identified in all three tests are in fact undergoing natural selection, then 1.6% to 4.1% of the loci that we examined in the three triads is selectively important or linked to areas of the genome that are selectively important. Among these SNPs, we found no fixed SNPs between polluted and reference populations suggesting selection in favor of alleles that have not yet reached fixation (Voight et al., 2006). This is not unexpected given the high migration rates among populations (Brown, Chapman, 1991).
and the recent selective change in the environment. In fact, despite the lethal concentrations of pollutants at these sites, these populations show no evidence of reduced genetic diversity (Williams, Oleksiak, 2008), most likely due to migrants.

A central question in outbred natural populations is whether similar or different solutions will evolve in response to comparable selective forces. Two SNPs that were significant in all three tests were shared between two different triads. A SNP in peroxiredoxin 6 was shared between the New Bedford Harbor and Elizabeth River populations, and a SNP in a sequence similar to mouse clone RP24-528E17 was shared between Newark Bay and Elizabeth River populations. Only one SNP was significant in all three tests and across all three triads: a SNP in the first intron of the phase I xenobiotic metabolizing enzyme CYP1A. CYP1A is integral to the detoxification pathway of many of the contaminants to which F. heteroclitus are continuously exposed in the three Superfund sites (Weis, 2002). These compounds, including polycyclic aromatic hydrocarbons (PAHs), dioxins and coplanar polychlorinated biphenyls (PCBs), induce CYP1A through the aryl hydrocarbon receptor (AhR) pathway (Hahn, 1998). In all three Superfund populations, CYP1A is refractory to induction by prototypical inducers (Bello et al., 2001; Elskus et al., 1999; Meyer, Di Giulio, 2002; Nacci et al., 1999), and this trait is associated with resistance to PAH, PCB and dioxin toxicity (Bello et al., 2001; Nacci et al., 1999). Potentially, the SNP in the first intron of CYP1A affects transcription or is linked to SNPs affecting transcription.

GWA studies have been used to relate polymorphisms to disease and phenotypic traits (Hindorff et al., 2009). For example, non-synonymous SNPs have been associated with
Crohn’s disease, arthritis, freckles, and height (Hindorff et al., 2009). Instead of relating SNPs to phenotypic traits, we asked which polymorphisms are associated with recent inhospitable environments (highly polluted Superfund sites). The analysis suggests that 1.6-4% of loci respond rapidly to change in the environment, 1.1-3% from mRNA encoding loci. These SNPs were randomly chosen and identified from expressed sequence tags (ESTs). If we assume 30,000 mRNA encoding genes, our results suggest that 330-900 loci have adaptively diverged in the last 50 years in each Superfund population. Relative to Haldane’s expectation (Haldane, 1990), this is a large number of loci affected by natural selection over a short time. Possibly, the 367 polymorphic loci assayed have an unusually high frequency of adaptive change. Assuming one SNP every ~1000 bp (Vignal et al., 2002), average gene sizes of ~1300 bp (Xu et al., 2006), and F. heteroclitus EST lengths of 400 bp (modal size is closer to 550 bp), we missed SNPs in the remaining 800 or 2/3 of nucleotides. If we also assume that all 2/3 are evolving by neutral processes, then we overestimated the percentage of non-neutral loci by 2/3 by only sampling 1/3 of each gene. Thus, 0.37-1% rather than 1.1-3% of genic loci respond rapidly, and conservatively, 110-300 genes are involved in adaptation in natural populations in a short time. These data suggest that adaptive divergence can occur rapidly and involve hundreds of loci.
Acknowledgements

Partial funding for this work was received from NIH 5 RO1 ES011588 and NSF OCE 1008542. The Authors thank Douglas L. Crawford for helpful discussions and review of the manuscript.
References


heteroclitus) from a creosote-contaminated site in the Elizabeth River, VA, USA.

Mar Environ Res 54, 621-626.


Table 1. Genes and loci with SNPs identified as outliers using the F\textsubscript{ST} modeling approach, Association Test, and Minor Allele Frequency-F\textsubscript{MAX} Test (MAF-F\textsubscript{MAX}). P-values for the association test, the likelihood ratio test for model fit and MAF-F\textsubscript{MAX} test are listed. P-values for the association test significant with a Bonferroni correction are bold. Genes and loci in blue are identified as outliers using both the F\textsubscript{ST} modeling approach and the MAF-F\textsubscript{MAX} test. Genes and loci in green are identified as outliers using the F\textsubscript{ST} modeling approach and the Association test. Genes and loci in purple are identified as outliers using the Association test and the MAF-F\textsubscript{MAX} test. Genes and loci in red are identified as outliers using all three tests.

<table>
<thead>
<tr>
<th>$F_{\text{ST}}$ modeling approach</th>
<th>Association Test</th>
<th>$p$-value</th>
<th>Likelihood Ratio Test</th>
<th>$p$-value</th>
<th>MAF-F\textsubscript{MAX} Test</th>
<th>$p$-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATP synthase D chain, mitochondrial</td>
<td>AMBP protein precursor</td>
<td>$1.2 \times 10^{-3}$</td>
<td>$1.1 \times 10^{-2}$</td>
<td>ATP synthase D chain, mitochondrial</td>
<td>$1.3 \times 10^{-4}$</td>
<td></td>
</tr>
<tr>
<td>ATP synthase subunit delta</td>
<td>Biotinidase fragment 2</td>
<td>$2.4 \times 10^{-4}$</td>
<td>$1.4 \times 10^{-3}$</td>
<td>ATP synthase subunit f, mitochondrial</td>
<td>$1.4 \times 10^{-4}$</td>
<td></td>
</tr>
<tr>
<td>ATP synthase subunit f, mitochondrial</td>
<td>CYP1A\textsubscript{PP} (707)</td>
<td>$1.2 \times 10^{-3}$</td>
<td>$2.1 \times 10^{-2}$</td>
<td>CYP1A\textsubscript{PP} (710)</td>
<td>$2.3 \times 10^{-3}$</td>
<td></td>
</tr>
<tr>
<td>Atrial natriuretic peptide</td>
<td>CYP1A\textsubscript{PP} (618)</td>
<td>$1.1 \times 10^{-5}$</td>
<td>$1.3 \times 10^{-5}$</td>
<td>CYP1A\textsubscript{PP} (268)</td>
<td>$7.6 \times 10^{-5}$</td>
<td></td>
</tr>
<tr>
<td>CYP1A\textsubscript{PP} (268)</td>
<td>Elastase 3</td>
<td>$2.6 \times 10^{-3}$</td>
<td>$4.1 \times 10^{-3}$</td>
<td>FABP, Fibrinogen, beta polypeptide</td>
<td>$1.3 \times 10^{-6}$</td>
<td></td>
</tr>
<tr>
<td>Elastase 3</td>
<td>FABP</td>
<td>$1.5 \times 10^{-4}$</td>
<td>$4.7 \times 10^{-4}$</td>
<td>Glyceraldehyde 3-phosphate dehydrogenase</td>
<td>$8.7 \times 10^{-5}$</td>
<td></td>
</tr>
<tr>
<td>FABP</td>
<td>Fibrinogen, beta polypeptide</td>
<td>$2.4 \times 10^{-10}$</td>
<td>$7.4 \times 10^{-10}$</td>
<td>Nudix diphosphate kinase B</td>
<td>$9.9 \times 10^{-5}$</td>
<td></td>
</tr>
<tr>
<td>Fibrinogen, beta polypeptide</td>
<td>Guanine nucleotide-binding protein subunit $\beta$-2-like 1</td>
<td>$1.2 \times 10^{-3}$</td>
<td>$1.5 \times 10^{-2}$</td>
<td>Peptidyl-prolyl cis-trans isomerase</td>
<td>$1.2 \times 10^{-5}$</td>
<td></td>
</tr>
<tr>
<td>Glyceraldehyde 3-phosphate dehydrogenase</td>
<td>NADH dehydrogenase [ubiquinone] 1 $\beta$ subcomplex subunit 2, mitochondrial precursor</td>
<td>$1.8 \times 10^{-3}$</td>
<td>$6.0 \times 10^{-3}$</td>
<td>Peroxiredoxin-6</td>
<td>$3.9 \times 10^{-3}$</td>
<td></td>
</tr>
<tr>
<td>Guanine nucleotide-binding protein subunit $\beta$-2-like 1</td>
<td>NADH dehydrogenase 1 $\alpha$ subcomplex subunit 4</td>
<td>$2.1 \times 10^{-5}$</td>
<td>$6.5 \times 10^{-4}$</td>
<td>Pleuromutilin-like peptide WF3 precursor</td>
<td>$4.6 \times 10^{-16}$</td>
<td></td>
</tr>
<tr>
<td>NADH dehydrogenase [ubiquinone] 1 $\beta$ subcomplex subunit 2, mitochondrial precursor</td>
<td>Nudix diphosphate kinase B</td>
<td>$7.3 \times 10^{-5}$</td>
<td>$8.2 \times 10^{-5}$</td>
<td>Ribosomal protein L7a</td>
<td>$1.1 \times 10^{-7}$</td>
<td></td>
</tr>
</tbody>
</table>
Table 1. Continued

| Nuclease diphosphate kinase B | Peptidyl-prolyl cis-trans isomerase | 9.8×10⁻⁵  | 1.0×10⁻⁵ | Ribosomal protein S10 | 4.9×10⁻⁵ |
| Peptidyl-prolyl cis-trans isomerase | Peroxiredoxin-6 | 8.1×10⁻⁷  | 1.0×10⁻³ | Ribosomal protein S29 | 5.5×10⁻³ |
| Peroxiredoxin-6 | Ribosomal protein S29 | 2.9×10⁻⁷  | 6.8×10⁻⁴ | Selenoprotein Pa precursor putative mRNA | 1.2×10⁻⁴ |
| Ribosomal protein L7a | Selenoprotein Pa precursor putative mRNA | 2.9×10⁻⁴  | 8.5×10⁻⁴ | Serotransferrin precursor | 2.5×10⁻³ |
| Ribosomal protein S15 | Serotransferrin precursor | 1.6×10⁻³  | 1.7×10⁻³ | Similar to yeast hypothetical protein DSM70294 | 5.3×10⁻⁸ |
| Selenoprotein Pa precursor putative mRNA | Similar to yeast hypothetical protein DSM70294 | 3.8×10⁻⁹  | 5.8×10⁻⁹ | Thioredoxin | 7.1×10⁻⁵ |
| Similar to yeast hypothetical protein DSM70294 | Thioredoxin | 2.1×10⁻⁷  | 2.3×10⁻⁸ | Unknown (AFLP 330_77) | 2.2×10⁻⁴ |
| Thioredoxin | Trypsinogen Y | 1.4×10⁻³  | 2.9×10⁻³ | Unknown (AFLP 886_60) | 4.1×10⁻³ |
| Trypsinogen Y | Unknown (AFLP 1230_196) | 2.8×10⁻⁴  | 9.3×10⁻⁴ | Vitellogenin I | 2.3×10⁻⁴ |
| Unknown (AFLP 210_146) | Vitellogenin | 1.6×10⁻³  | 3.8×10⁻³ | 60S ribosomal protein L41 | 5.0×10⁻⁶ |
| Vitellogenin | Vitellogenin | 5.1×10⁻¹  | 3.8×10⁻³ | Vitellogenin-1 precursor | 2.8×10⁻⁴ |
| Vitellogenin | 40S ribosomal protein S24 | 6.6×10⁻⁵  | 1.4×10⁻³ | 40S ribosomal protein S24 | 8.3×10⁻⁴ |
| 40S ribosomal protein S24 | 40S ribosomal protein S26 | 8.3×10⁻⁴  | 6.1×10⁻³ | 40S ribosomal protein S26 | 2.2×10⁻⁴ |
| 60S ribosomal protein L24 | 60S ribosomal protein L27 | 1.8×10⁻¹  | 1.2×10⁻² | 60S ribosomal protein L24 | 2.6×10⁻⁷ |
| 60S ribosomal protein L24 | 60S ribosomal protein L41 | 2.6×10⁻⁷  | 4.2×10⁻⁷ | 60S ribosomal protein L41 | 2.6×10⁻⁷ |

Newark Bay

<table>
<thead>
<tr>
<th>FST modeling approach</th>
<th>Association Test</th>
<th>p-value</th>
<th>Likelihood ratio test p-value</th>
<th>MAF-FMAX Test</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>B2-microglobulin</td>
<td>B2-microglobulin</td>
<td>1×10⁻⁵</td>
<td>2.3×10⁻⁵</td>
<td>B2-microglobulin</td>
<td>2.4×10⁻⁴</td>
</tr>
<tr>
<td>CYP1A_PP (+268)</td>
<td>CYP1A_PP (-518)</td>
<td>1.6×10⁻⁶</td>
<td>2.0×10⁻³</td>
<td>CYP1A_PP (+268)</td>
<td>1.5×10⁻³</td>
</tr>
<tr>
<td>Ferritin, middle subunit (Ferritin M)</td>
<td>CYP1A_PP (+268)</td>
<td>9.6×10⁻⁴</td>
<td>4.0×10⁻³</td>
<td>Ferritin, middle subunit (Ferritin M)</td>
<td>4.8×10⁻³</td>
</tr>
<tr>
<td>Lysozyme C precursor</td>
<td>Ribosomal protein</td>
<td>7.1×10⁻³</td>
<td>1.0×10⁻²</td>
<td>Lysozyme C</td>
<td>5.5×10⁻¹¹</td>
</tr>
</tbody>
</table>

139
Table 1. Continued

<table>
<thead>
<tr>
<th>Ribosomal protein S2</th>
<th>S2</th>
<th>precursor</th>
<th>Ribosomal protein S2</th>
<th>S2</th>
<th>precursor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similar to gDNA of Danio rerio clone #CH211-27352</td>
<td>6.2x10^-3</td>
<td>1.3x10^-2</td>
<td>Similar to gDNA of Danio rerio clone #CH211-27352</td>
<td>5.2x10^-4</td>
<td></td>
</tr>
<tr>
<td>Similar to mouse clone RP24-528E17</td>
<td>3.1x10^-3</td>
<td>5.0x10^-3</td>
<td>Similar to mouse clone RP24-528E17</td>
<td>6.0x10^-3</td>
<td></td>
</tr>
<tr>
<td>Zona Pellucida binding protein</td>
<td>5.3x10^-3</td>
<td>9.0x10^-2</td>
<td>Unknown (TC17861_587)</td>
<td>1.1x10^-4</td>
<td></td>
</tr>
<tr>
<td>Zona Pellucida binding protein</td>
<td>7.8x10^-3</td>
<td></td>
<td>Zona Pellucida binding protein</td>
<td>3.6x10^-4</td>
<td></td>
</tr>
</tbody>
</table>

### Elizabeth River

<table>
<thead>
<tr>
<th>Fat trimming approach</th>
<th>Association Test</th>
<th>p-value</th>
<th>Association Test</th>
<th>p-value</th>
<th>MAF-Fmax Test</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actin, alpha cardiac</td>
<td>Astacin like metallo-protease</td>
<td>2.6x10^-5</td>
<td>Astacin like metallo-protease</td>
<td>4.9x10^-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adult beta-type globin</td>
<td>B2-microglobulin</td>
<td>8.5x10^-5</td>
<td>B2-microglobulin</td>
<td>9.5x10^-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Astacin like metallo-protease</td>
<td>Chymotrypsinogen 2-like protein</td>
<td>2.6x10^-4</td>
<td>Chymotrypsinogen 2-like protein</td>
<td>5.3x10^-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B2/C2 protein</td>
<td>CYP1A_PP (+268)</td>
<td>3.7x10^-4</td>
<td>CYP1A_PP (+268)</td>
<td>3.2x10^-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chymotrypsinogen 2-like protein</td>
<td>CYP1A_PP (-618)</td>
<td>5.1x10^-3</td>
<td>CYP1A_PP (+268)</td>
<td>3.7x10^-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP1A_PP (+268)</td>
<td>Drtp1</td>
<td>1.7x10^-3</td>
<td>Drtp1</td>
<td>1.2x10^-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP1A_PP (-618)</td>
<td>Eukaryotic translation elongation factor 2</td>
<td>7.3x10^-3</td>
<td>Leukocyte elastase inhibitor</td>
<td>2.2x10^-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CYP1A_PP (-707)</td>
<td>Fibrinogen, gamma polypeptide</td>
<td>2.0x10^-6</td>
<td>Fibrinogen, gamma polypeptide</td>
<td>3.4x10^-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Drtp1</td>
<td>NADH dehydrogenase [ubiquinone] 1-0 subcomplex subunit 2, mitochondrial precursor</td>
<td>3.2x10^-3</td>
<td>Peroxiredoxin-6</td>
<td>3.9x10^-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eukaryotic translation elongation factor 2</td>
<td>Nascent polypeptide-associated complex subunit alpha</td>
<td>3.2x10^-5</td>
<td>Ribosomal protein L7</td>
<td>2.5x10^-3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fibrinogen, gamma polypeptide</td>
<td>Nucleosome</td>
<td>6.3x10^-3</td>
<td>Ribosomal protein S9</td>
<td>1.4x10^-4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leukocyte elastase inhibitor</td>
<td>Pancreatic progenitor cell differentiation and proliferation factor b</td>
<td>7.9x10^3</td>
<td>2.5x10^2</td>
<td>Ribosomal protein S20</td>
<td>2.1x10^4</td>
<td></td>
</tr>
<tr>
<td>----------------------------</td>
<td>---------------------------------------------------------------------</td>
<td>--------</td>
<td>---------</td>
<td>----------------------</td>
<td>---------</td>
<td></td>
</tr>
<tr>
<td>L-SF precursor</td>
<td>Peroxiredoxin-6</td>
<td>1.4x10^3</td>
<td>1.9x10^3</td>
<td>SEC61, gamma subunit isoform 1</td>
<td>9.7x10^5</td>
<td></td>
</tr>
<tr>
<td>NADH dehydrogenase [ubiquinone] 1-β subcomplex subunit 2, mitochondrial precursor</td>
<td>Ribosomal protein L7</td>
<td>2.6x10^3</td>
<td>3.2x10^2</td>
<td>Similar to chimp clone PTB-146D01</td>
<td>4.7x10^4</td>
<td></td>
</tr>
<tr>
<td>Nascent polypeptide-associated complex subunit alpha</td>
<td>Ribosomal protein S9</td>
<td>1.1x10^5</td>
<td>1.1x10^2</td>
<td>Similar to mouse clone RP24-525E17</td>
<td>5.3x10^5</td>
<td></td>
</tr>
<tr>
<td>Nuclease diphosphate kinase B</td>
<td>Ribosomal protein S20</td>
<td>4.0x10^6</td>
<td>8.0x10^5</td>
<td>Unknown (AFLP 122_129)</td>
<td>3.9x10^3</td>
<td></td>
</tr>
<tr>
<td>Pancreatic progenitor cell differentiation and proliferation factor b</td>
<td>SEC61, gamma subunit isoform 1</td>
<td>1.3x10^4</td>
<td>6.3x10^3</td>
<td>Unknown (AFLP 56_52)</td>
<td>3.9x10^3</td>
<td></td>
</tr>
<tr>
<td>Peroxiredoxin-6</td>
<td>Similar to chimp clone PTB-146D01</td>
<td>3.8x10^4</td>
<td>4.5x10^4</td>
<td>Similar to mouse clone RP24-525E17</td>
<td>3.5x10^2</td>
<td></td>
</tr>
<tr>
<td>Phosvitinless vitelligenin</td>
<td>TBT-binding protein</td>
<td>1.4x10^5</td>
<td>2.2x10^5</td>
<td>Unknown (AFLP 11_129)</td>
<td>3.9x10^3</td>
<td></td>
</tr>
<tr>
<td>Ribosomal protein L34</td>
<td>TBT-binding protein</td>
<td>2.2x10^3</td>
<td>2.3x10^2</td>
<td>Unknown (AFLP 136_94)</td>
<td>4.1x10^3</td>
<td></td>
</tr>
<tr>
<td>Ribosomal protein L7</td>
<td>Trypsinogen 2</td>
<td>5.5x10^4</td>
<td>6.2x10^3</td>
<td>Unknown (AFLP 56_52)</td>
<td>4.1x10^3</td>
<td></td>
</tr>
<tr>
<td>Ribosomal protein S9</td>
<td>Unknown (AFLP 11_129)</td>
<td>2.0x10^3</td>
<td>5.8x10^3</td>
<td>Unknown (AFLP 56_52)</td>
<td>4.1x10^3</td>
<td></td>
</tr>
<tr>
<td>Ribosomal protein S6</td>
<td>Unknown (AFLP 136_94)</td>
<td>2.8x10^5</td>
<td>3.9x10^3</td>
<td>Unknown (AFLP 11_129)</td>
<td>4.1x10^3</td>
<td></td>
</tr>
<tr>
<td>SEC61, gamma subunit isoform 1</td>
<td>Unknown (AFLP 136_94)</td>
<td>1.2x10^3</td>
<td>4.1x10^3</td>
<td>Unknown (AFLP 122_129)</td>
<td>4.1x10^3</td>
<td></td>
</tr>
<tr>
<td>Selenoprotein Pa precursor putative miRNA</td>
<td>Unknown (AFLP 122_129)</td>
<td>1.4x10^3</td>
<td>4.1x10^3</td>
<td>Unknown (AFLP 122_129)</td>
<td>4.1x10^3</td>
<td></td>
</tr>
<tr>
<td>Similar to chimp clone PTB-146D01</td>
<td>Unknown (AFLP 525_183)</td>
<td>5.2x10^4</td>
<td>4.2x10^3</td>
<td>Unknown (AFLP 122_129)</td>
<td>4.1x10^3</td>
<td></td>
</tr>
<tr>
<td>Similar to Gasterosteus aculeatus clone VMRC25-21C14</td>
<td>Unknown (AFLP 967_45)</td>
<td>1.7x10^3</td>
<td>7.4x10^3</td>
<td>Unknown (AFLP 967_45)</td>
<td>4.1x10^3</td>
<td></td>
</tr>
<tr>
<td>Similar to mouse clone RP24-525E17</td>
<td>Unknown (AFLP 1079_383)</td>
<td>2.1x10^4</td>
<td>3.4x10^3</td>
<td>Unknown (AFLP 1079_383)</td>
<td>4.1x10^3</td>
<td></td>
</tr>
<tr>
<td>TBT-binding protein</td>
<td>Unknown (AFLP 1461_65)</td>
<td>2.9x10^4</td>
<td>4.9x10^3</td>
<td>Unknown (AFLP 1461_65)</td>
<td>4.1x10^3</td>
<td></td>
</tr>
<tr>
<td>Unknown (AFLP 11_129)</td>
<td>Unknown (AFLP 56_52)</td>
<td></td>
<td></td>
<td>Unknown (AFLP 11_129)</td>
<td>4.1x10^3</td>
<td></td>
</tr>
<tr>
<td>Unknown (AFLP 122_129)</td>
<td>Unknown (AFLP 136_94)</td>
<td></td>
<td></td>
<td>Unknown (AFLP 122_129)</td>
<td>4.1x10^3</td>
<td></td>
</tr>
<tr>
<td>Unknown (AFLP 136_94)</td>
<td>Unknown (AFLP 280_83)</td>
<td></td>
<td></td>
<td>Unknown (AFLP 280_83)</td>
<td>4.1x10^3</td>
<td></td>
</tr>
</tbody>
</table>
Table 1. Continued

<table>
<thead>
<tr>
<th>Unknown (AFLP 510_115)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zona pellucida binding protein</td>
</tr>
<tr>
<td>Unknown (AFLP 987_45)</td>
</tr>
<tr>
<td>Unknown (AFLP 1197_172)</td>
</tr>
<tr>
<td>Unknown (AFLP 1303_220)</td>
</tr>
<tr>
<td>Unknown (AFLP 1328_135)</td>
</tr>
<tr>
<td>Unknown (AFLP 1461_65)</td>
</tr>
<tr>
<td>Unknown (TC16623_751)</td>
</tr>
</tbody>
</table>
Figure 1. *F. heteroclitus* sampling sites along the East coast of the United States. Polluted sites are starred and flanked north and south by clean reference sites (circles) to form a triad. Venn diagrams indicate the number of SNPs exhibiting non-neutral behavior using the three statistical tests: the $\text{F}_{\text{ST}}$-modeling approach ($\text{F}_{\text{ST}}$), Association (Assoc.), and MAF-$\text{F}_{\text{MAX}}$ (MAF).
Figure 2. F$_{ST}$-modeling approach to detect selection. Empirical F$_{ST}$ values are plotted against heterozygosity. The line demarks the 99th percentile estimated from a simulation model. Blue diamonds indicate SNPs that are significantly different between the polluted population and both reference populations but not different for reference versus reference. Red dots are superimposed on blue diamonds if the SNP was also significant in the other two statistical tests. Less interesting are the crosses and open diamonds. Black crosses are outliers also in the reference versus reference comparison. Open diamonds represent outliers where the polluted population was only significant in comparison to one reference population.
Figure 3. **Association test for detection of selection.** Likelihood of association of each SNP with either the polluted site (red line) or reference sites (blue line) as a $-\log_{10}p$-value. The $-\log_{10}p$-value of 2 is marked by a black line, and the Bonferroni correction for multiple testing is marked by the dotted grey line ($-\log_{10}p$-value of 4.56). SNPs are identified as outliers in polluted sites versus reference sites if the polluted association value is greater than 2 and the likelihood ratio test $p$-value of polluted versus reference association is $\leq 0.01$ (supplemental Table S1). Black squares indicate those SNPs where the likelihood model for pollution significantly exceeds the model based on divergence among reference sites. Red dots are superimposed on black squares if the SNP was also significant in the other two statistical tests.
Figure 4. MAF-$F_{\text{MAX}}$ test for detection of differences in SNP allele frequencies between polluted and reference sites. The allele frequency of the triad-wide minor allele was calculated and plotted for all SNPs. Columns are collection sites arranged north to south, and each row represents an individual SNP. SNPs with allele frequencies significantly different in an ANOVA using $F_{\text{MAX}}$ (Westfall, Young, 1993) to control for type I errors among iterations ($F_{\text{MAX}}$: empirical F-value exceeds the top 1% of all permuted F-values assuming random population differentiation) between polluted and both reference sites are plotted. Red dots denote SNPs exhibiting non-neutral behavior in all three statistical tests. The SNP exhibiting non-neutral behavior in all three triads and using all tests (CYP1A +268) is boxed.
CHAPTER 4

Cytochrome P4501A promoter polymorphisms and activity in natural populations

Larissa M. Williams¹ and Marjorie F. Oleksiak²

1. Department of Environmental and Molecular Toxicology  
   Box 7633, North Carolina State University  
   Raleigh, NC 27695-7633 USA

2. Rosenstiel School of Marine and Atmospheric Sciences  
   University of Miami  
   4600 Rickenbacker Causeway  
   Miami, FL 33149 USA

Corresponding Author:  
Marjorie Oleksiak  
Rosenstiel School of Marine and Atmospheric Sciences  
University of Miami  
4600 Rickenbacker Causeway  
Miami, FL 33149  
Fax: 305-421-4600  
Email: moleksiak@rsmas.miami.edu
Abstract

Cytochrome P4501A (CYP1A) has been shown to be refractory to prototypic inducers in populations of the estuarine minnow, *Fundulus heteroclitus*, adapted to chronic anthropogenic pollution. Two SNPs in the promoter and first intron of the CYP1A promoter were previously found to be under selection, indicating that natural selection was acting on the CYP1A promoter or loci linked to these SNPs. In order to understand the role of the CYP1A promoter and these selectively important SNPs, the promoter and first intron and exon were sequenced in multiple individuals in one polluted (New Bedford Harbor) population resistant to PCBs and other anthropogenic contaminants and two reference populations north and south of the polluted site. The CYP1A promoter was extremely variable (an average of 21% of the promoter nucleotides varied among all populations) and there were no fixed differences between populations. There was little variation in known and predicted transcription factor binding sites between populations. There was also no signature of selection along the promoter, and the SNPs found to be under selection in a prior study were not in linkage disequilibrium with the remainder of the promoter or the first intron and exon. The inducibility of the promoter was also explored; the promoter was induced by a prototypic PAH, 3-methylcholanthrene, in a dose-dependent manner. When promoters from multiple individuals per population were tested, the CYP1A promoter was most induced in the polluted New Bedford Harbor population as compared to both reference populations. This is in contrast to CYP1A expression *in vivo* which is refractory to induction in New Bedford Harbor individuals. Overall, the variation in the promoter does not explain the variation in *in vitro* promoter expression in all tested populations. These results indicate that
the underlying mechanism for the \textit{in vivo} transcriptional phenotype may lay further upstream or downstream of the CYP1A promoter region which was sequenced (1600bp), or involve proteins which were not available in the cell line (top minnow) in which the transfection assays were completed.

\section*{Introduction}

The estuarine minnow, \textit{Fundulus heteroclitus}, inhabit several of the most heavily polluted estuaries in the world (Wirgin, Waldman, 2004), including New Bedford Harbor, Massachusetts, a Superfund site polluted with high levels of polychlorinated biphenyls (PCBs) (Weaver, 1983). While this site has only been contaminated for approximately 50 years, \textit{F. heteroclitus} has adapted to its seemingly inhospitable conditions. Recently, two SNPs in the proximal promoter of the hepatic monooxygenase, cytochrome P4501A (CYP1A), were shown to be under selection in the New Bedford Harbor population as compared to clean, reference populations (Williams, Oleksiak, 2010) indicating a genetic underpinning for adaptation. Furthermore, \textit{F. heteroclitus} individuals from NBH are refractory to CYP1A induction by prototypic inducers (Bello \textit{et al.}, 2001; Nacci \textit{et al.}, 1999) and have lower CYP1A activity as compared to reference populations. The basis for this refractory phenotype has not been characterized, but could involve the SNPs under selection either directly or through linkage to areas responsible for modulating CYP1A transcription.

CYP1A has been used as a biomarker of environmental exposure to anthropogenic contamination (M. Nilsen \textit{et al.}, 1998) such as polycyclic aromatic hydrocarbons (PAH; (Arinc \textit{et al.}, 2000; Collier \textit{et al.}, 1998; Hahn, 1998; Huggett \textit{et al.}, 2006; M. Nilsen \textit{et al.},
CYP1A transcription is mainly regulated through the aryl hydrocarbon receptor (AHR) pathway (Whitlock et al., 1996). The AHR is a highly conserved ligand-activated transcription factor that is primarily found in the cytoplasm, complexed with Hsp90 and immunophilin-like Ara9 (also known as XAP2 or AIP). Upon interaction with xenobiotic ligands such as PCBs and PAHs (Denison, Nagy, 2003), AHR translocates to the nucleus, disassociates with Hsp90, and dimerizes with the aryl hydrocarbon receptor nuclear translocator (ARNT) (Reyes et al., 1992). The AHR-ARNT heterodimer interacts with gene-regulatory elements commonly referred to as dioxin response elements (DRE) or xenobiotic response elements (XRE) (Murre et al., 1989), disrupting chromatin structure and facilitating the interaction of additional transcription factors and co-regulatory proteins with the promoters of target genes, including CYP1A (Beischlag et al., 2008).

CYP1A transcription, total protein and enzyme activity has been well described in the estuarine minnow species, *F. heteroclitus*. *F. heteroclitus* lives along the east coast of the United States in both urban and pristine estuaries where it is one of the most abundant intertidal marsh fishes (Burnett et al., 2007). *F. heteroclitus* is a non-migratory species (Skinner et al., 2005), living in large subpopulations (Adams et al., 2006). These large subpopulations maintain high standing genetic variation (Oleksiak, 2010), allowing quick
adaptation to changing environmental conditions. Over the last 150 years, the east coast has become highly urbanized, leading to anthropogenic contamination of the waters and sediment of the estuaries. Many of these urban sites have been designated as Superfund sites: hazardous waste sites mandated for clean up by the Comprehensive Environmental Response, Compensation, and Liability Act of 1980.

One notable Superfund site is New Bedford Harbor (NBH), also known as the Acushnet River Estuary. From the 1940s to 1978, two capacitor manufacturing facilities discharged PCBs and heavy metals both directly and indirectly (through the sewer system) into the harbor (Weaver, 1983). This activity made NBH one of the most PCB contaminated sites in the United States, with concentrations of PCBs in the sediment exceeding 100,000 parts per million (Weaver, 1983) in some areas and 100-2,000 μg/g dry sediment in other areas of the estuary (Lake et al., 1995; Pruell et al., 1990). NBH also contains high levels of chlorinated dibenzofurans, PAHs, and metals such as Cr, Cu, Cd, and Pb (Pruell et al., 1990; Shine et al., 1995; Weaver, 1983). While much of the flora and fauna that lived in NBH pre-urbanization can no longer survive there, *F. heteroclitus* has flourished in this seemingly toxic environment (Nacci et al., 2001; Nacci et al., 2002b), despite bioaccumulating whole body PCB concentrations of 1,370 μg/g dry weight and non-*ortho* and mono-*ortho* PCB congener concentrations of 30-35 μg/g dry weight in the liver (Black et al., 1998; Lake et al., 1995). The mechanism(s) by which *F. heteroclitus* has adapted to this environment has been the attention of much research for several decades and is focused around the CYP1A xenobiotic metabolizing gene and associated pathways.
*F. heteroclitus* individuals from NBH are refractory to CYP1A induction by prototypic inducers (Arzuaga, Elskus, 2010; Bello *et al.*, 2001; Nacci *et al.*, 1999) and have lower CYP1A activity as compared to reference populations. NBH fish and offspring are also insensitive to the lethal effects of PCBs (Nacci *et al.*, 2002a) and accumulate fewer DNA adducts when exposed to the PAH, benzo(a)pyrene (Nacci *et al.*, 2002b). The mechanism of differential sensitivity between NBH and reference populations has been proposed to involve the AHR-ARNT-CYP1A signaling pathway and may involve both heritable and non-heritable mechanisms. The genetic basis of the resistance phenotype has not been described. Efforts have been made to explore the role of the AHR pathway; no changes in gene expression of AHR (1 and 2), ARNT, or repressor (AHRR) have been found between NBH and reference populations (Karchner *et al.*, 2002; Powell *et al.*, 2000). While the AHR1 (Hahn *et al.*, 2002; Hahn *et al.*, 2004) and AHR2 (Hahn *et al.*, 2005) loci are polymorphic, and there are different allele frequencies in NBH and reference populations that may have been driven by selective pressures related to polluted environments (Hahn *et al.*, 2002; Hahn *et al.*, 2004), no variants have been shown to be functionally important *in vitro* (Hahn *et al.*, 2004).

Single nucleotide polymorphisms in the CYP1A promoter have recently been shown to be under selection or linked to areas of the genome under selection in *F. heteroclitus* from NBH as compared to two reference sites (Williams, Oleksiak, 2010). One selectively important SNP is located at -670 bp upstream of the +1 transcriptional start site and the other at 173 bp into the first intron of the CYP1A gene. These selectively important SNPs may
alter transcription factor binding or be linked to areas affecting transcription and may be responsible for the refractory induction of CYP1A in the NBH population.

To further explore the CYP1A promoter and the role of these SNPs in resistance to anthropogenic contaminant mixtures, we sequenced 1,600 bp of the promoter and the first (non-coding) exon and intron in multiple individuals from NBH and reference sites to assess promoter-wide patterns of variation and to conduct evolutionary analyses among and between populations. We also tested the inducibility of each promoter to a prototypic PAH using a transient luciferase transfection assay to determine CYP1A promoter activities within and between populations.

**Materials and Methods**

**Genomic Clones**

*F. heteroclitus* were collected using minnow traps during the spring of 2005. Fin clips were sampled from eight individuals from each of the three collection sites along the Atlantic Coast of the United States (Figure 1). One of the three collection sites was an EPA Superfund site: New Bedford Harbor (EPA ID: MAD980731335) and the other two represent clean reference sites which flanked the polluted site equidistant north and south. *Fundulus grandis*, a sister species to *F. heteroclitus*, was collected from Port O’Connor, Texas in the spring of 2008.

Genomic DNA from fin clips was extracted using a modified version of Aljanabi and Martinez (Aljanabi, Martinez, 1997), and DNA was resuspended in 50 µL 0.1X TE buffer. 2kb of the CYP1A promoter and first intron and two exons were amplified using PCR from
the genomic DNA of each individual. The forward primer (5’-AGTTATAGCCACAGTCCAGTCATTT-3’) was located 1575 bp upstream of the transcriptional start site and the reverse primer (5’-CAAGGCTATCAAACCCTCAGACAC-3’) was 526 bp downstream of the transcriptional start site in the second exon (Powell et al., 2004). Loci were ligated into a Promega pGEM-T vector overnight at 16°C with T4 DNA ligase, transformed into electrocompetant JM109 E.coli cells, and grown on carbenicillin selective plates. Colonies were screened for insert with PCR using vector specific primers. Insert PCR products from three clones per individual were sequenced in both directions using ABI BigDye terminator chemistry providing 6x coverage of each nucleotide. Sequences were aligned in MacVector 8.0 (Olson, 1994) and CAP3 (Huang, Madan, 1999).

**Sequence Analysis**

CYP1A sequences from each of eight individuals in NBH, Point Judith and Sandwich and one *F. grandis* individual were analyzed with DnaSP (Librado, Rozas, 2009) to determine patterns of sequence variation along the promoter as well as in the first and second intron and first exon of the CYP1A gene. To examine evolutionary processes affecting the CYP1A promoter, Tajima’s D was calculated across the promoter in all three populations, as well as for each individual nucleotide (Tajima, 1989). Fu and Li’s test for natural selection (Fu, Li, 1993) was also performed and the D and F statistics were calculated using *F. grandis* as the outgroup. Average nucleotide substitutions per site within (P1:1 and P1:2) and between populations (Dxy) was also calculated and plotted using a sliding window with a 50 bp-wide window and a 10 bp step size. A modification of the McDonald-Kreitman test (Mcdonald, 1996)
Kreitman, 1991) was used to detect selection following a procedure described by Crawford et al. (Crawford et al., 1999) which compares the patterns of sequence variation among functional and non-functional regions of the promoter. Functional regions (xenobiotic and glucocorticoid response elements) were ascribed a priori (Karchner et al., 1999; Powell et al., 1999; Powell et al., 2004) but not tested for functionality in this study. Additional potential regulatory regions within the promoter were determined with AliBaba 2.1 (http://www.gene-regulation.com/pub/programs/alibaba2/index.html), which predicts binding sites of transcription factors in unknown DNA sequences using binding sites collected in TRANSFAC (Matys et al., 2006), a database containing eukaryotic cis-acting regulatory DNA elements and trans-acting factors. The McDonald-Kreitman test distinguishes between natural selection and neutral evolutionary processes by testing the ratio between fixed functional differences to fixed non-functional differences relative to the ratio of polymorphisms between taxon. A two-tailed Fisher’s exact test was used to test the null hypothesis that sequence variation is random. Arlequin 3.5 (Excoffier, Lischer, 2010) was used to determine linkage disequilibrium among nucleotides along the promoter.

Phylogenetic analysis for the whole promoter, functional and non-functional regions was applied to the proximal promoter for each of the eight F. heteroclitus individuals from the three collection sites as well as the outgroup species, F. grandis using maximum parsimony methods in PAUP* 4.0 (Swofford, 2003).

**Reporter Gene constructs**

Reporter gene constructs for transfection experiments were generated by digesting
CYP1A inserts from the pGEM-T vector first with ApaI, blunting the insert with T4 DNA polymerase and further digesting with SacI restriction enzymes. Promega pGL3-Basic vector containing the Firefly luciferase gene was linearized by digesting with KpnI, blunting with T4 DNA polymerase and a final digestion with SacI restriction enzymes. These cuts assisted in the directional cloning of the CYP1A insert into the pGL3-Basic vector upstream of the luciferase gene. The promoter construct was 6,831 bp in length, with 2013 bp being the CYP1A promoter, first intron and first exon (Supplementary Figure 1). The 2013 bp which consisted of the CYP1A promoter, first intron and first exon, the start of transcription occurred at 1630 bp. There was an intact TATA box at 1596bp. Upstream of the transcriptional start site there were three XRE sites at 799bp, 848bp and 1405bp. Digested insert DNA and pGL3-Basic vector was gel purified using Zymoclean Gel DNA recovery kit and ligated together overnight at 16°C. Ligation product was transformed into electrocompeant JM109 E.coli cells, and plated on carbenicillin selective plates. Colonies were screened for insert with PCR using vector specific primers.

**Cell Culture and transfection**

Fish hepatoma cells PLHC-1 (*Poeciliopsis lucida* hepatoma cell) (Ryan, Hightower, 1994) were grown in Gibco CO₂ independent media supplemented with 5% FBS and 50 U ml⁻¹ of penicillin G/50 μg ml⁻¹ streptomycin in a 30°C incubator. PLHC-1 cells have an intact and inducible CYP1A system and possess properties similar to those observed in *F. heteroclitus* liver (Hahn et al., 1993). Cells were usually split every 4 days by dissociating with 0.05% (w/v) trypsin and 0.5 mM EDTA and subcultured at 2x10⁷ cells per each 75cm²
tissue culture flask. Using Fugene 6 transfection reagent (Boehringer), cells at 50–80% confluence were transiently co-transfected in 12-well plates at a density of 2x10^6 cells per well with the reporter gene vector and an internal control vector expressing Renilla luciferase under the control of a cytomegalovirus promoter (pRL-CMV) (Promega). Cells were treated with 3-methylcholanthrene (3-MC) dissolved in DMSO, or with DMSO only. 3-MC is a prototypic inducer of CYP1A through the AHR pathway and evokes a TCDD and related halogenated aromatic hydrocarbon pattern of induction (Poland, Knutson, 1982). Luciferase and Renilla activities were measured using the Promega Dual-Luciferase Reporter Assay System.

Results

Proximal and First Exon and Intron Sequences of CYP1A

1630 nucleotides of the CYP1A proximal promoter were sequenced in eight individuals from each of three populations (Sandwich, New Bedford Harbor, and Point Judith). The promoter has a substantial amount of variation, with 21% of the promoter containing variable nucleotides between individuals and populations although there are no fixed differences between populations. One hundred and fifty-seven parsimony informative sites occur along the promoter upstream of the transcriptional start site: 17 are fixed differences between F. heteroclitus and its sister species F. grandis (Figure 2). The three xenobiotic response elements (Powell et al., 2004) were mainly conserved and had few polymorphisms across populations and species (Supplementary Figure 1). XRE3 had several polymorphisms present only in the New Bedford Harbor population at nucleotides 1406 and
1407, where individuals NBH6 and NBH13 have a TT rather than an AC. The average nucleotide diversity (excluding gaps; N= 1474 bp) for the proximal promoter (Jukes, Cantor, 1969) for *F. heteroclitus* among all three populations is 0.0150 or theta (θ) per site of 0.027. Per population, the average nucleotide diversity is 0.012 for Sandwich (N=1514 bp), 0.021 for New Bedford Harbor (N=1494 bp), and 0.0016 for Point Judith (N=1538 bp).

The first exon and intron also were sequenced, as one of the SNPs under selection (Williams, Oleksiak, 2010) was located in the first intron (at nucleotide 1892). The first exon contained 11 parsimony informative sites (of which one was a fixed difference between *F. heteroclitus* and *F. grandis*), and the first intron contained 34 parsimony informative sites (of which two were fixed difference between *F. heteroclitus* and *F. grandis*). No polymorphisms in the first exon resulted in a nonsynonymous change to the deduced amino acid sequence. The average nucleotide diversity for the first exon was 0.018 for Sandwich (N=111 bp) and 0.009 for New Bedford Harbor (N=112 bp); Point Judith contained no polymorphisms. The average nucleotide diversity for the first intron was 0.017 for Sandwich (N=268), 0.040 for New Bedford Harbor (N=261), and 0.005 for Point Judith (N=264).

Sequence variation distinguished population divergence and established genetic distance by geographical distance (Table 1). Between the two most geographically close sites, Sandwich and New Bedford, the $G_{ST}$ was 0.028 and $F_{ST}$ was 0.054. Between New Bedford Harbor and Point Judith, the $G_{ST}$ was 0.037 and $F_{ST}$ was 0.405. Sandwich and Point Judith, the two furthest sampling sites from each other, had a $G_{ST}$ of 0.040 and a $F_{ST}$ of 0.410. The average number of pairwise nucleotide differences ($K_{xy}$) were similar in all three
pairwise comparisons, with 42.11 for Sandwich vs. New Bedford, 41.14 for Sandwich vs. Point Judith, and 46.09 for New Bedford vs. Point Judith (Table 1).

Linkage disequilibrium was compared for all polymorphic nucleotides along the promoter and the first exon and intron. For Sandwich, 558 of the total 12,483 pairwise comparisons of polymorphic loci (4.5%) were in significant LD (Figure 3). New Bedford had 2,175 significant pairwise comparisons in LD, out of a total of 33,871 pairwise combinations (6.4%) (Figure 4). Point Judith had the fewest significant pairwise comparisons (158) but had the highest percentage of significant comparisons of 46.76% among the 338 pairwise comparisons (Figure 5). Sandwich and New Bedford Harbor sequences showed strong LD among nucleotides in the 5’ end of the promoter and also among a conserved area of the 3’ end of the first intron. At the 3’ end of the first intron, the sequence ‘ACTT’ spanning nucleotides 1970 to 1973, was in LD with a large portion of the promoter. This was also true in the Point Judith population, although the LD of this area with the rest of the promoter spanned fewer nucleotides than was the case in Sandwich and New Bedford Harbor. Point Judith sequences have a large, strong, block of LD in the 5’ end of the promoter where a conserved, ancestral insertion occurs. There is no LD among the SNPs (929, 1892 bp) found to be under selection in the previous study (Williams, Oleksiak, 2010) with the remainder of the promoter, intron one or exon one. No functional regions (Supplementary Table 1) were found to be in significant LD with other nucleotides along the promoter or beyond the transcriptional start site with the exception of XRE3, which was in significant LD in the New Bedford Harbor population with nucleotides 1336-1361 (just upstream of the binding XRE3 binding site at 1405-1410).
Evolutionary Analyses of Promoter

Evolution by natural selection does not seem to be a driving force behind the sequence variation found within and among *F. heteroclitus* populations. A McDonald-Kreitman test, which compares functional and non-functional, fixed and polymorphic changes between species (or in this case, also between populations), failed to reject the null hypothesis that the CYP1A promoter is not evolving by natural selection (Table 2). Most of the changes in the promoter were found to be polymorphic in nonfunctional regions.

The variation between *F. heteroclitus* can be visualized in sliding window comparisons (Figure 6). In the comparison of Sandwich to New Bedford Harbor, similar areas of the promoter are equally variable between populations except between nucleotides 1553-1619 (upstream of the TATA box at 1596 bp) where there is a difference between the nucleotide substitutions per site between the two populations. In this area there are several polymorphisms that only exist in one individual in one population, increasing the Dxy value. There is also significant variation within the Point Judith population, although this variation decreases at the very 3’ end of the promoter. When compared with New Bedford Harbor, Dxy exceeds the nucleotide substitution per site among populations in the 3’ end of the promoter spanning from 1466-1555. In this area, Point Judith individuals do not have any sequence variation, and there is moderate variation in three individuals (4, 6, 13) in the New Bedford Harbor population. When clean, reference sites, Sandwich and Point Judith, are compared, the same area that had a large Dxy value in the Point Judith vs. New Bedford Harbor comparison (1466-1555bp) exhibits the same pattern. In this area, there is no
sequence variation in the Point Judith population, but there is moderate variation in the Sandwich population. There are no clear patterns of variation near functional regions of the promoter between New Bedford Harbor and its clean, reference sites.

A similar lack of a clear pattern in variation along the CYP1A promoter can be visualized by plotting Tajima’s D and Fu and Li’s D statistics by population (Figure 7). The Tajima test for data from a single locus (Tajima, 1989) compares the estimate of $\theta$ based on the average number of pairwise differences to that based on the number of segregating sites. The test of Fu and Li (Fu, Li, 1993) with an outgroup (F. grandis in this case) compares estimates of the number of segregating sites to the number of mutations on external branches expected under neutrality. Tajima’s D was -0.202 for Sandwich, 0.034 for New Bedford Harbor and 0.496 for Point Judith populations. None of these values were significant at a p-value of 0.05 and no sliding window region was significant along the promoter. Fu and Li’s D also was calculated for the promoter in each population and was 2.10 for Sandwich (p <0.02), 2.04 for New Bedford Harbor (p<0.02) and 0.96 for Point Judith (p >0.10) populations. Sandwich, New Bedford Harbor and Point Judith sequences have generally positive D statistics along the promoter signifying low levels of both low and high frequency polymorphisms and indicating a decrease in population size and/or balancing selection. Fu and Li’s D statistic was found to be significantly positive in the Sandwich population at the sliding window encompassing nucleotides 1492-1541. This area falls upstream of the TATA box but does not contain any known or predicted transcription factor binding sites. The New Bedford Harbor population also had a significantly positive Fu and Li’s D statistic in the sliding windows of nucleotide regions 1241-1353 bp and 1406-1490 bp. Notably, XRE3
falls in the second window. In the Point Judith population, similar patterns of positive D values were found in the Tajima’s D and Fu and Li’s calculation, although this pattern of variation was not significant.

Phylogenetic analysis also revealed apparently random variation along the promoter. Maximum parsimony was used to construct relationships between individuals, populations and species for the CYP1A promoter (Figure 8). When the full length promoter is used to construct a phylogenetic analysis, Point Judith individuals (which have little sequence variation) fall out together and New Bedford Harbor and Sandwich individuals are interspersed with each other (Figure 8A). One individual from the New Bedford Harbor population (NBH307) falls out with the outgroup, *F. grandis*. This individual has many ancestral nucleotides and indels, which sets it apart from the other individuals both in its own population and in comparison to other populations. A similar, random, pattern is seen when non-functional regions of the promoter are used to construct the phylogenetic analysis (Figure 8B). As noted in the McDonald-Kreitman test, there is little sequence variation in the functional regions of the CYP1A promoter. This is reflected in the phylogenetic analysis, where all *F. heteroclitus* form a monopyletic group (Figure 8C).

**Functional Assays**

The induction of luciferase expression from two *F. heteroclitus* CYP1A proximal promoter reporter gene constructs by several doses of 3-MC in PLHC-1 cells was tested (Figure 9). Transiently transfected cells were dosed with 3-MC or treated with vehicle (DMSO) control as described in Williams *et al.* (Williams *et al.*, 2000), at concentrations
ranging from 0.001 μM to 2 μM. This test was conducted to assess whether luciferase
eexpression increased in a dose-dependent manner in the clean (Sandwich individual S27)
and/or polluted (New Bedford Harbor individual NBH6) CYP1A proximal promoter
constructs. At lower doses (<0.05 μM), luciferase expression was similar for both constructs,
and not significantly different from vehicle-treated control. At 0.05 μM the New Bedford
Harbor construct had significant fold induction over control. At doses of 0.1 μM and above,
both Sandwich and New Bedford Harbor constructs had significant fold induction over
control. New Bedford also had a significantly higher (p ≤ 0.05) fold induction in comparison
to Sandwich, ranging from 1.5 to 1.9 times greater than Sandwich. A horizontal asymptote
occurs at the 1μM dose, indicating that maximal induction of the promoter occurred at or
around that dose.

To determine if there are individual and population differences between the
inducibility of CYP1A promoters, luciferase activities from four individual promoter
constructs were compared within and between the three populations (Figure 10) at a dose of
1 μM 3-MC. The pattern observed was very similar to that of the dose-response curve:
reference populations, Sandwich and Point Judith, had on average a fold induction over
control of 2.66 and 2.39, respectively, which was lower than the New Bedford Harbor
population which had an average fold induction over control of 4.54. Within population
induction variability was minimal, ranging from 0.14 to 0.30 over control. A one-way
ANOVA determined a significant difference between populations (p < 0.0001), and a
Tukey’s HSD post-hoc test determined that New Bedford Harbor was significantly different
from Sandwich (p< 0.0001) and Point Judith (p< 0.0001), but Sandwich and Point Judith were not significantly different from each other (p= 0.258).

**Discussion**

The role of the CYP1A proximal promoter in the resistance phenotype to environmental, anthropogenic contamination in the New Bedford Harbor *F. heteroclitus* population had not been investigated prior to this study. Furthermore, an in-depth survey of genetic variation in this promoter in natural populations also had not been conducted. This study sought to determine which regions, if any, of the CYP1A promoter are evolving by natural selection and whether the promoter plays a functional role in the refractory CYP1A inducibility phenotype that has been described in the New Bedford Harbor population (Bello *et al.*, 2001; Nacci *et al.*, 1999).

The CYP1A proximal promoter, cloned from eight individuals from each of three populations, is extremely variable. Up to 21% of the nucleotides in the promoter vary between individuals. A McDonald-Kreitman test indicated that the promoter was not evolving by natural selection due to the fact that there were no nucleotides that were fixed or near fixation. The variation observed for the CYP1A promoter is much greater relative to other promoters that have been sequenced between populations: *Ldh-B* proximal promoter in *F. heteroclitus* (10%; (Crawford *et al.*, 1999); the first intron of *Adh* (1.7%; (Kreitman, 1983), the *eve* enhancer (0.87%; (Ludwig, Kreitman, 1995), and *G6pd* (1.0%; (Eanes *et al.*, 1993) in *Drosophila melanogaster*; *G* protein-coupled receptor kinase 4 (0.94%; (Hasenkamp *et al.*, 2008) and CCR5 (2.1%; (Bamshad *et al.*, 2002) in humans; endo16 in the purple sea
urchin (10%; (Balhoff, Wray, 2005); the chalcone synthetase promoter in Arabidopsis thaliana (1%; (de Meaux et al., 2005). However, unlike the studies listed above, which showed many of the polymorphisms in functional regions of the promoter, almost all of the variation (92%) for CYP1A exists in regions without a described function.

Plots of variation between populations (Figure 6) revealed that each population surveyed had a substantial and similar amount of variation. It is possible that the observed changes do not alter the pattern of transcription-factor binding and are thus entirely neutral. It has also been hypothesized that stabilizing selection on transcriptional output allows slightly deleterious mutations to persist, compensated for by adaptive changes elsewhere in the promoter and resulting in continuous binding-site turnover (Balhoff, Wray, 2005; Ludwig et al., 2000). This alternative hypothesis is supported for Sandwich and New Bedford populations by the Fu and Li’s D test for selection which found significantly positive values for the promoter. Significantly positive values for this statistic reflect an excess of intermediate-frequency alleles, which can result from population bottlenecks, structure and/or balancing selection. Conversely, Tajima’s D for each population was not significant. Tajima’s D has been found to be more powerful than Fu and Li’s D test to test for very strong, recent directional selection (Braverman et al., 1995; Simonsen et al., 1995). However, it has been shown that Tajima’s D and Fu and Li’s D test have very low power to detect balancing selection (Charlesworth et al., 1995). The one test that does have slightly higher power to detect this type of selection, Fu and Li’s F-test (Fu, Li, 1993), found no significant patterns of variation along the CYP1A promoter in any population (data not shown). Given that only one of the three tests for selection was significant, balancing
selection most likely is not acting on the CYP1A promoter to compensate for slightly deleterious mutations, but rather the pattern of variation is neutral. Variation is most likely maintained due to large population sizes and migration. The pattern, but not the significant nature, of variation was entirely different for known functional regions of the promoter.

The regions of known or described function were mainly conserved across populations and species (between *F. heteroclitus* and its sister species *F. grandis*). There were no fixed differences between species or populations and only 13 polymorphisms total in functional regions among all individuals and populations. Despite very high levels of conservation within functional sites, there was no significant selection signal on those areas in Tajima’s D or Fu and Li’s D. If balancing selection were maintaining those sites, one would also expect to find high linkage disequilibrium between the nucleotides of each site and other functional sites (Charlesworth, 2006; Wall, 1999); this was not the case. In the functional regions where there were polymorphisms, the majority of the polymorphisms were in one or a few individuals across populations. There was a polymorphism in the XRE3, which binds the AHR-ARNT complex less tightly than XRE1 (Powell *et al.*, 2004). In New Bedford Harbor individuals 6 and 13, the XRE site was “C*TTG*CGA” rather than the consensus sequence “C*ACG*CGA”. However, the difference in the XRE sequence did not significantly alter the inducibility of the CYP1A promoter construct *in vitro* (for individual NBH6) compared to other NBH individuals assays, which have the consensus XRE3 consensus sequence. While the XREs described in the CYP1A promoter do bind the AHR-ARNT complex *in vitro* (Powell *et al.*, 2004), the functional significance of each of these XREs is still unclear with respect to inducing CYP1A transcription. Deletion constructs
containing all, two, one or none of the XREs coupled to a reporter gene would have to be conducted in vitro in order to assess their ability to support AHR-regulated transcription. Furthermore, in order to more accurately assess their role in binding AHR-ARNT in vivo, an in vivo footprinting assay (Schulte et al., 1995) would be necessary.

For a DNA sequence that has to bind many conserved transcription factors in order to induce the transcription of CYP1A, the promoter is quite variable. However, transfection assays with CYP1A promoter constructs coupled to the luciferase gene showed that the CYP1A promoter functioned in vitro to stimulate transcription and was inducible in a dose-responsive manner to a prototypic PAH, 3-MC. That inducibility of luciferase expression differed between two different individual constructs, one from Sandwich and one from New Bedford Harbor given several dosages of 3-MC. Fold induction over control was similar between the constructs below 0.05 μM 3-MC, but at 0.05 μM 3-MC the New Bedford Harbor construct had significant fold induction over control. At doses of 0.1 μM 3-MC and above, both Sandwich and New Bedford Harbor constructs had significant fold induction over control. New Bedford also had a significantly higher (p ≤ 0.05) fold induction in comparison to Sandwich, ranging from 1.5 to 1.9 times greater than Sandwich. This NBH individual had a similar dose- response curve to CYP1A promoter constructs derived from European flounder (Platichthys flesus) which were transfected into HepG2 cells (Williams et al., 2000). In flounder it was shown that doses as low as 0.05 μM 3-MC had a significant fold induction of luciferase expression over vehicle control. Flounder CYP1A was cloned from fish unexposed to environmental contaminants. These observations indicate that New Bedford Harbor derived-CYP1A promoters may function in a manner similar to other teleosts,
regardless of exposure to environmental contaminants. However, since the fold induction of the Sandwich individual was significantly less than that of the New Bedford Harbor individual at the same doses, individual variation in inducibility due to genetic differences (individual or population-level) may be the driving force behind CYP1A promoter activity. Cardiac gene expression variation between *F. heteroclitus* individuals within a population is very high but has been shown to predict variation in metabolism (Oleksiak *et al.*, 2005). Similarly, genetic variation among the CYP1A promoter may predict the inducibility of the promoter by PAHs and PCBs. To explore this theory, we performed a series of transfection assays on constructs derived from four individuals from each of the populations to determine if the pattern of inducibility was based on individual differences or was a shared characteristic of a population.

The CYP1A promoter inducibility was not significantly variable within a population, but was significantly different between polluted and reference populations. New Bedford Harbor had significantly higher fold induction over control as compared to each reference population (Figure 10). The reference populations did not have significantly different average fold induction from each other. There are no fixed differences along the entire promoter or in known or described transcription factor binding sites to explain the significant difference in average fold induction over control between New Bedford Harbor populations and its reference site populations. Given the refractory CYP1A transcriptional phenotype exhibited by New Bedford Harbor fish (Bello *et al.*, 2001; Nacci *et al.*, 1999), it was expected that the CYP1A promoter for this population would be less inducible by prototypic PAHs as compared to populations without environmental contamination histories—the
opposite was found. CYP1A transcription is controlled mainly through the AHR pathway. While there are no known functional differences found in the ability of AHR1 variants to bind TCDD between New Bedford Harbor and Sandwich (Hahn et al., 2004), there are variants specific to individuals from polluted sites that have yet to be tested for functionality in the AHR2 gene as well as the AHR repressor (Hahn et al., 2005). Variants of either of these two proteins may function solely or with other proteins to give both the inducibility and CYP1A refractory phenotypes described. There are also many other enhancers and repressors that may serve to enhance the inducibility of the CYP1A gene while still maintaining the refractory transcriptional phenotype. In humans, the vasoactive intestinal peptide was shown to be regulated by numerous enhancers and repressors in a 600 bp region 4kb upstream of the transcriptional start site (Liu et al., 2001). Thus, the CYP1A promoter may be regulated through many different proteins that have not been identified because they are further up- or downstream from the promoter sequence used for this study, and these areas may contain variants specific to the populations. These could also be in linkage disequilibrium with the two SNPs found to be under selection (Williams, Oleksiak, 2010) but shown in this study not to be in LD with any other nucleotides in the sequenced area. LD has been measured in humans from a few kilobases to in some instances greater than 100 kb (Patil et al., 2001). Without a sequenced genome, the detection of such long distance LD in F. heteroclitus is not possible. Thus, if other SNPs responsible for modulating CYP1A transcription are in LD with our selectively important SNPs but are beyond the 1.5 kb sequenced area of this study, they will not be detected. To determine whether the SNPs identified as being under selection were important to the transcriptional phenotype, site-
directed mutagenesis could be used. If the mutagenesis of either or both of the SNPs resulted in a polluted individual having a reference phenotype and vice-versa, the role of these SNPs in the transcriptional regulatory phenotype could be attributed to those particular SNPs.

Several other genes, such as basic leucine zipper nuclear factor 1, CYP1B1, and guanidinoacetate N-methyltransferase, amongst others, have shown a similar gene expression profile in the New Bedford Harbor population as compared to reference populations as CYP1A (Oleksiak et al., unpublished). This pattern indicates that there may be shared transcriptional regulation of these genes and others which may involve a singular, or combination of several different, transcription factors, enhancers and/or repressors.

Epigenetic factors may also be causative in this phenotype. One particular epigenetic mechanism, that of cytosine methylation at CpG sites, was explored for the CYP1A promoter in *F. heteroclitus* populations residing in the Elizabeth River, VA Superfund site (Timme-Laragy et al., 2005). Cytosine methylation was not detected at any of the 34 CpG sites examined in any of the total eight adult fish examined, including 3 CpG sites that are part of putative XREs. Thus, this particular epigenetic mechanism could not explain the refractory phenotype of the CYP1A gene in the Elizabeth River polluted population. Similarly Arzuaga *et al.* (Arzuaga et al., 2004) found that the DNA de-methylating agent 5-azacytidine does not restore CYP1A induction in the polluted *F. heteroclitus* population in the Superfund site of Newark Bay suggesting DNA methylation is not responsible for the refractory CYP1A phenotype in that population. These studies do not rule out methylation as a potential mechanism in the New Bedford Harbor population. Due to the shared phenotype between
the three polluted populations, it does however imply it may not be the underlying factor. There are other epigenetic mechanisms which could contribute to this refractory phenotype including gene silencing, position effect, bookmarking, maternal effects, regulation of histone modification and heterochromatin structure, paramutation, and transvection. In order to explore these possibilities, among other possibilities, the use of chromatin immunoprecipitation, epigenetic microarray technologies, DNA adenine methyltransferase identification, fluorescent in situ hybridization, and bisulfite sequencing could be used.

The high inducibility of the New Bedford Harbor CYP1A constructs result could also be explained as an artifact of the cell line type. It is possible that the PLHC-1 cell type does not contain all the same transcription factors as does F. heteroclitus. Thus, if there is a necessary repressor found in F. heteroclitus but not top minnow (Poeciliopsis lucida) which would create a lower inducibility of the promoter in the New Bedford Harbor fish, we would not observe this in our study. However, this study did show population differences between New Bedford Harbor and its reference sites, which are not an artifact of the cell culture.

The CYP1A promoter is extremely variable, mainly in non-functional areas and is not evolving through natural selection based on the McDonald-Kreitman test, Tajima’s D calculations, and Fu and Li’s D calculations. Patterns of variation differ between each of the tested populations; however all are equally variable. Known and predicted functional transcription factor binding sites are well conserved across populations. One site that differed, XRE3 in two New Bedford Harbor individuals, did not seemingly contribute to the inducibility of the CYP1A promoter in vitro to a prototypic inducer. Inducibility of the
promoter was dose-dependent, and significantly higher in New Bedford Harbor individuals. This study opens the door to the investigation of other mechanisms involved in the refractory CYP1A transcriptional phenotype observed in the New Bedford Harbor population.
Acknowledgements

Partial funding for this work was received from NIH 5 RO1 ES011588 and NSF OCE 1008542. The authors thank the University of Miami’s Molecular Core for sequencing.
References


Table 1. Measures of population divergence between *F. heteroclitus* populations. 

\( K_{xy} = \pi D \) (average number of pairwise nucleotide differences), \( G_{ST} \) is Nei’s coefficient of gene variation and defined as \( 1-(H_S/H_T) \) where \( H_S \) and \( H_T \) are the mean heterozygosity within populations and in the entire species, respectively, and \( F_{ST} \) is Wright's inbreeding coefficient and is defined as \( (H_T-H_S)/H_T \).

<table>
<thead>
<tr>
<th>Population 1</th>
<th>Population 2</th>
<th>( K_{xy} )</th>
<th>( G_{ST} )</th>
<th>( F_{ST} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sandwich</td>
<td>New Bedford</td>
<td>42.11</td>
<td>0.028</td>
<td>0.054</td>
</tr>
<tr>
<td>Sandwich</td>
<td>Point Judith</td>
<td>41.14</td>
<td>0.040</td>
<td>0.410</td>
</tr>
<tr>
<td>New Bedford</td>
<td>Point Judith</td>
<td>46.09</td>
<td>0.037</td>
<td>0.405</td>
</tr>
</tbody>
</table>
Table 2. Neutrality test for the pattern of sequence variation (McDonald, Kreitman, 1991). Probability that this pattern of sequence variation is due to random (H₀) or selective (H₁) evolutionary processes. A two-sided Fisher’s exact test failed to reject the null hypothesis (p=0.4906) for table A which tests between species. A two-sided Fisher’s exact test failed to reject the null hypothesis (p=1.0) for table B which tests between polluted and reference populations of F. heteroclitus.

<table>
<thead>
<tr>
<th></th>
<th>Fixed</th>
<th>Polymorphic</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Functional</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Non-functional</td>
<td>17</td>
<td>142</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>Fixed</th>
<th>Polymorphic</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Functional</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Non-functional</td>
<td>0</td>
<td>142</td>
</tr>
</tbody>
</table>
Figure 1. *F. heteroclitus* collection sites. Superfund site (New Bedford Harbor, MA) is denoted by a star, and reference sites north and south denoted by a circle.
Figure 2. 221 parsimony informative sites among and between *F. heteroclitus* populations and between *F. heteroclitus* and *F. grandis* in the sequenced portion of the CYP1A promoter, exon and intron 1. Within the promoter region (upstream of basepair 1630), there are 157 parsimony informative sites. There are no fixed differences between *F. heteroclitus* populations. 20 fixed differences between *F. heteroclitus* and *F. grandis* are marked by a star. SNPs (929 and 1892bp) found to be under selection in Williams and Oleksiak (2010) are starred over the nucleotide number.
<table>
<thead>
<tr>
<th>Taxon</th>
<th>Nucleotide Position</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sandwich 6</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 7</td>
<td>CCCGGGCAAAGTGCCGAGTTAATGATTCTTCAAGTGTTGAT</td>
</tr>
<tr>
<td>Sandwich 8</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 9</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 10</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 11</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 12</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 13</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 14</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 15</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 16</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 17</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 18</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 19</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 20</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 21</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 22</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 23</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 24</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 25</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 26</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 27</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 28</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 29</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>Sandwich 30</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 3</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 4</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 5</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 6</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 7</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 8</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 9</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 10</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 11</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 12</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 13</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 14</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 15</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 16</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 17</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 18</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 19</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 20</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 21</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 22</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 23</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 24</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 25</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 26</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 27</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 28</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 29</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>NBH 30</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 1</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 2</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 3</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 4</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 5</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 6</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 7</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 8</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 9</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 10</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 11</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 12</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 13</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 14</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 15</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 16</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 17</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 18</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 19</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 20</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 21</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 22</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 23</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 24</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 25</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>FU 26</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
<tr>
<td>F. grandis 13</td>
<td>CGTCTTTAAGTGTGAGACTTTAAACCTCCAAACAGTTGAT</td>
</tr>
</tbody>
</table>
**Figure 3.** Linkage disequilibrium (LD) among polymorphic nucleotides along the proximal promoter and first exon and intron of CYP1A for Sandwich. Green indicates no significant LD, red indicates significant LD ($p \leq 0.05$), and the blue line indicates the diagonal.
Figure 4. Linkage disequilibrium (LD) among polymorphic nucleotides along the proximal promoter and first exon and intron of CYP1A for New Bedford Harbor. Green indicates no significant LD, red indicates significant LD (p≤0.05), and the blue line indicates the diagonal.
Figure 5. Linkage disequilibrium (LD) among polymorphic nucleotides along the proximal promoter and first exon and intron of CYP1A for Point Judith. Green indicates no significant LD, red indicates significant LD (p≤0.05), and the blue line indicates the diagonal.
Figure 6. Sliding window comparisons of average nucleotide substitutions per site within *F. heteroclitus* populations (Pi) and between populations (Dxy). Plots for the sliding window use 50bp-wide window and 10-bp step size.
Figure 7. Tajima’s D calculations (panels A,C,E) and Fu and Li’s D calculations (panels B, D, and F) using a sliding window with a 50bp-wide window and a 10-bp step size for each F. heteroclitus population.
Figure 8. Phylogenetic analyses of the CYP1A proximal promoter using PAUP*4.0 (Swofford, 2003). A. Phylogenetic tree using all nucleotides of the proximal promoter. B. Phylogenetic tree using functional regions of the proximal promoter. C. Phylogenetic tree using non-functional regions of the proximal promoter. Bootstrap values (N=500) are listed. Probabilities are maximum parsimony values.
A. Full length proximal promoter

B. Non-Functional regions of the proximal promoter

C. Functional regions of the proximal promoter
**Figure 9.** Induction of luciferase activity from *F. heteroclitus* CYP1A reporter gene constructs from a New Bedford and Sandwich individual by 3-MC in PLHC-1 cells plotted on a non-logarithmic (A) and logarithmic scale (B). A star indicates a significant fold induction over control (p<0.05).
Figure 10. Average fold induction of the CYP1A promoter over control for three populations (N=4) at a 1 μM dose of 3-MC. A one-way ANOVA found a significant difference in the average fold induction between populations (p< 0.0001) and a Tukey’s HSD post-hoc test determined that New Bedford Harbor was significantly different from Sandwich (p< 0.0001) and Point Judith (p< 0.0001), but Sandwich and Point Judith were not significantly different from each other (p= 0.258). Significant induction in the New Bedford Harbor population over the reference populations is denoted by a star.
**Supplementary Table 1.** Predicted transcription factor binding sites along the CYP1A promoter by AliBaba2.1 by constructing matrices from TRANSFAC 4.0. XRE and GRE consensus sequences were ascribed and tested for functionality as described in Powell et al. (2004).

<table>
<thead>
<tr>
<th>Nucleotide Position</th>
<th>Transcription Factor</th>
<th>Consensus Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>121-130</td>
<td>GATA-1</td>
<td>TTTTTTTAT/GC</td>
</tr>
<tr>
<td>230-238</td>
<td>C/EBPα</td>
<td>TAAGCAATC</td>
</tr>
<tr>
<td>399-404</td>
<td>GRE 1/2 Site</td>
<td>AGGACA</td>
</tr>
<tr>
<td>403-413</td>
<td>Oct-1</td>
<td>CATGTAACAA</td>
</tr>
<tr>
<td>729-740</td>
<td>Sp-1</td>
<td>CTCTCCCCCTCC</td>
</tr>
<tr>
<td>799-805</td>
<td>XRE 1</td>
<td>CACGCAA</td>
</tr>
<tr>
<td>848-854</td>
<td>XRE 2</td>
<td>CACGCAA</td>
</tr>
<tr>
<td>1027-1036</td>
<td>MEB-1 and GLO</td>
<td>TTATTTTTAA</td>
</tr>
<tr>
<td>1101-1111</td>
<td>CREB AND C/EBPα and AP1</td>
<td>CCTGCCTCAGC</td>
</tr>
<tr>
<td>1212-1221</td>
<td>C/EBPα</td>
<td>TAAATTAAT</td>
</tr>
<tr>
<td>1324-1333</td>
<td>OCT2.1</td>
<td>GCTGATTTTGC</td>
</tr>
<tr>
<td>1391-1401</td>
<td>GR</td>
<td>ACACCTTTTGA</td>
</tr>
<tr>
<td>1405-1410</td>
<td>XRE 3</td>
<td>CACGCGAA</td>
</tr>
<tr>
<td>1528-1539</td>
<td>Sp-1</td>
<td>GGAGGAGGGGAGA</td>
</tr>
<tr>
<td>1538-1546</td>
<td>CREB</td>
<td>GATGATGTC</td>
</tr>
<tr>
<td>1596-1601</td>
<td>TATA Box</td>
<td>TATAAA</td>
</tr>
</tbody>
</table>
Supplementary Figure 1. Full length proximal promoter and exon 1 and intron 1. Star indicates conservation in the nucleotide across all individuals, P indicates a parsimony informative site, and F indicates a fixed difference between *F. heteroclitus* and *F. grandis*. GRE and XRE binding sites are noted, as well as the TATA box, start of transcription, and intron and exonic boundaries. Big stars on nucleotides 929 and 1892 indicate SNPs under selection as described in Williams and Oleksiak (2010).
<table>
<thead>
<tr>
<th>Taxon</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sandwich 6</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>Sandwich 7</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>Sandwich 8</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>Sandwich 11</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>Sandwich 12</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>Sandwich 14</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>Sandwich 27</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>NBI 3</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>NBI 4</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>NBI 5</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>NBI 7</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>NBI 9</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>NBI 12</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>NBI 13</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>NBI 307</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>FJ 1</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>FJ 2</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>FJ 5</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>FJ 6</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>FJ 19</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>FJ 26</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>FO Grand 13</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>F</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>F</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>F</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>F</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
</tbody>
</table>

F

<table>
<thead>
<tr>
<th>Taxon</th>
<th>Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sandwich 6</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>Sandwich 7</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>Sandwich 8</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>Sandwich 11</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>Sandwich 12</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>Sandwich 14</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>Sandwich 27</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>NBI 3</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>NBI 4</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>NBI 5</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>NBI 7</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>NBI 9</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>NBI 12</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>NBI 13</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>NBI 307</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>PJ 1</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>PJ 2</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>PJ 5</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>PJ 6</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>PJ 19</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>PJ 26</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>FO Grand 13</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>F</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>F</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>F</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
<tr>
<td>F</td>
<td>ATACAAAACCTTTAAATTATACACCGATATCATGATATGAGGTCGCAAAATATATACATTAAATATTTGCA</td>
</tr>
</tbody>
</table>

F F F P
| Sandwich 6 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 7 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 8 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 9 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 10 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 11 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 12 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 13 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 14 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 15 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 16 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 17 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 18 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 19 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 20 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 21 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 22 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 23 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 24 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 25 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 26 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 27 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 28 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 29 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 30 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 31 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 32 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 33 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 34 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 35 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 36 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 37 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 38 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 39 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 40 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 41 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 42 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 43 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 44 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 45 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 46 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 47 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 48 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 49 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 50 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 51 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 52 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 53 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 54 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 55 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 56 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 57 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 58 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 59 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |
| Sandwich 60 | GATTCAGACTCGTTAGGCACTAAAT | GATTCAGACTCGTTAGGCACTAAAT |

** +1 (Exon 1) **
| Sandwich 6 | TTGCACAACAC |
| Sandwich 7 | TTGCACAACAC |
| Sandwich 8 | TTGCACAACAC |
| Sandwich 11 | TTGCACAACAC |
| Sandwich 12 | TTGCACAACAC |
| Sandwich 14 | TTGCACAACAC |
| Sandwich 27 | TTGCACAACAC |
| Sandwich 32 | TTGCACAACAC |
| NBH 3 | TTGCACAACAC |
| NBH 4 | TTGCACAACAC |
| NBH 6 | TTGCACAACAC |
| NBH 7 | TTGCACAACAC |
| NBH 8 | TTGCACAACAC |
| NBH 12 | TTGCACAACAC |
| NBH 13 | TTGCACAACAC |
| NBH 307 | TTGCACAACAC |
| PJ 1 | TTGCACAACAC |
| PJ 2 | TTGCACAACAC |
| PJ 3 | TTGCACAACAC |
| PJ 4 | TTGCACAACAC |
| PJ 5 | TTGCACAACAC |
| PJ 6 | TTGCACAACAC |
| PJ 18 | TTGCACAACAC |
| PJ 26 | TTGCACAACAC |
| PO Grand 13 | TTGCACAACAC |

************
SUMMARY AND DISCUSSION

This thesis sought to understand whether adaptation in *F. heteroclitus* to anthropogenic pollution has genetic basis in three independent populations. Prior to this thesis, there was very little knowledge of the overall genetic variation in populations of *F. heteroclitus* populations living in polluted areas as compared to clean, reference sites. This thesis used genome-wide scans to establish whether there may be genetic adaptation through the process of natural selection acting on these populations and if so, to identify those signatures of selection.

Summary of Chapters

Chapter one focused on establishing and carrying out a molecular technique to accurately capture genetic variation in nine populations of *F. heteroclitus*. The *F. heteroclitus* genome is not sequenced, and so a technique called amplified fragment length polymorphisms (AFLP) was employed. In total, 296 loci were scored among *F. heteroclitus* from New Bedford Harbor and its reference sites, 336 loci in *F. heteroclitus* from Newark and its two reference sites, and 299 loci in *F. heteroclitus* from Elizabeth River and its two reference sites. Among all populations, 450 distinct loci were scored. *F_{ST}* values were calculated for each locus among each pairwise comparison and compared against simulated values (Beaumont, Nichols, 1996). When empirical *F_{ST}* values were compared to a simulated distribution of *F_{ST}* values, 24 distinct outlier loci were identified among pairwise comparisons of pollutant impacted *F. heteroclitus* populations and both surrounding
reference populations. Two outlier loci were shared between New Bedford Harbor and Elizabeth River populations, and two different loci were shared between Newark Bay and Elizabeth River populations. In sum, 1% to 6% of loci are implicated as being under selection or linked to areas of the genome under selection in three *F. heteroclitus* populations that reside in polluted estuaries. Shared loci among polluted sites indicate that selection may be acting on multiple loci involved in adaptation, and loci shared between polluted sites potentially are involved in a generalized adaptive response. This chapter illustrated the utility of AFLP analyses to detect selection in natural *F. heteroclitus* populations.

Due to the anonymous nature of AFLPs, they are difficult to sequence using traditional methods (*e.g.*, gel extraction, reconstitution and sequencing). Chapter two sought to employ a new, high-throughput sequencing method to establish its utility in sequencing AFLP fragments to identify single nucleotide polymorphisms (SNPs) for population genetic studies in *F. heteroclitus* and related species *F. grandis* and *F. similis* (Gulf of Mexico) and *F. majalis* (Atlantic coast). High-throughput genotyping methods were also used to verify previously identified SNPs in expressed sequence tags (ESTs) among populations and species of the *Fundulus* genus. Using the 454 FLX pyrosequencing system, 1,464 distinct contigs were obtained. Two-hundred and sixty-one SNPs were identified in 96 of these contigs using a new statistical framework developed by the Bustamante lab at Cornell University. We further verified 81 of these SNPs on the MassARRAY genotyping platform. SNPs showed latitudinal clinal variation separating northern and southern populations and established isolation by distance in *F. heteroclitus* populations. In *F. grandis*, the SNPs isolated from *F. heteroclitus* were less polymorphic (*i.e.*, 74% were monomorphic versus
11.2 in *F. heteroclitus*) but still established isolation by distance. Markers differentiated species and populations. These approaches were used to quickly determine differences within the *Fundulus* genome and provide markers for population genetic studies. The process by which this data was collected also provided a framework with which researchers in the future could sample and carry out population genetic studies on non-model species. This is especially important, because a majority of species do not have a sequenced genome, since it is still cost prohibitive. Thus establishing a method which can generate hundreds of markers in a non-sequenced genome accurately is necessary. While our method was not perfect, given several changes such as making reduced representation libraries from many pooled individuals for SNP discovery followed by individual genotyping, laboratories studying adaptation in any number of species would benefit from the method which could provide valuable information with respect to how genetics and molecular mechanisms affect gene frequency.

SNP data collected in chapter two was further analyzed in chapter three to detect signatures of selection in polluted populations of *F. heteroclitus*. Three statistical tests, which tested different polymorphic parameters of the loci (minor allele frequency, frequency of each allele and heterozygosity), were used to detect outlier behavior among 367 SNPs across nine populations (3 polluted and 6 reference). The first test, an F<sub>ST</sub> test, identified 30 outlier loci (8.2%) in the New Bedford Harbor triad, 9 (2.5%) in the Newark Bay triad, and 43 (11.7%) in the Elizabeth River triad. To further test the robustness of our outliers, an association test, which calculates p-values based on the strength of association with polluted populations was used. The association test identified 30 SNP (8.2%) which were
significantly associated with the polluted New Bedford Harbor model ($p \leq 0.01$), 7 (1.9%) in the Newark Bay triad, and 29 (7.9%) in the Elizabeth River triad. Differences between minor allele frequencies (MAF) also were tested between polluted and reference sites. In the New Bedford Harbor population, 22 SNPs (6.0%) had significantly different MAF between polluted and both reference populations. In the Newark Bay and Elizabeth River populations, 9 and 18 SNPs (2.5% and 4.9%, respectively) had significantly different MAF between polluted and reference populations. Within each triad, 6-15 SNPs were identified as outliers in all three tests: the New Bedford Harbor triad had 12, the Newark Bay triad had 6, and the Elizabeth River triad had 15. Among all triads, 15 of these SNPs occur in coding regions. Only one of these 15 SNPs, a SNP in $\beta_2$-microglobulin, is non-synonymous. One SNP was significant in all three tests and across all three triads: a SNP in the first intron of the phase I xenobiotic metabolizing enzyme CYP1A. In total, 1.6% to 4.1% of the loci that we examined in the three triads is selectively important or linked to areas of the genome that are selectively important. Extrapolating across the genome, 110-300 genes are involved in adaptation in natural populations in a short time. This is the first estimate that has been made for $F.\ heteroclitus$ populations residing in polluted areas, and also has implications for understanding adaptation in other species adapted to complex environments.

Few studies on natural populations have shown that a single gene is responsible for a given phenotype; rather many phenotypes in natural populations are the composite of genotypes across many loci. In order to address this, many researchers are now using quantitative trait loci (QTL) studies to identify multiple markers that may contribute to a particular phenotype. However, a prerequisite for a QTL study is a linkage map, which
assists in the determination of associate phenotypes with specific identifiable regions of the genome (Stinchcombe, Hoekstra, 2007). There are several non-model species where linkage maps have been made by maintaining species in captivity and breeding crosses in the laboratory including butterflies (*Heliconius, Bicyclus*), sticklebacks (*Gasterous*), deermice (*Peromyscus*), monkeyflowers (*Mimulus*), and columbines (*Aqueligia*). An example where QTL studies have identified candidate genes underlying a phenotype in a natural population is in the three-spine stickleback, where the QTL study identified a 10 Mb region containing a large effect contributing to adaptive variation in pelvic morphology between oceanic and lake populations, and the gene *Pitx1* was identified as the underlying genetic cause for this difference (Shapiro *et al.*, 2004). However, QTL studies often only focus on the large effect regions and do not characterize small effect regions, which may be contributing to a phenotype. In the case of *F. heteroclitus*, where the phenotype is resistance to complex mixtures of contaminants, a QTL study may identify large effects (such as CYP1A transcription or protein activity), but will most likely fail to investigate small effects (if they are identified at all). The study conducted in chapter three provides evidence that the phenotype may be composed of many small genotype effects across the genome, making it harder to point to one responsible, underlying adaptive locus. That said, chapter three did identify several SNPs in the CYP1A promoter as being selectively important. The last chapter of this thesis sought to understand the role of these SNPs in a known phenotype for *F. heteroclitus* residing in polluted populations.

In an attempt to link an outlier SNP to a functionally significant phenotype, the role of the CYP1A SNP found to be under selection in all three triads was explored in the fourth
data chapter. In all three Superfund populations, CYP1A is refractory to induction by prototypical inducers (Bello et al., 2001; Elskus et al., 1999; Meyer, Di Giulio, 2002; Nacci et al., 1999), and this trait is associated with resistance to PAH, PCB and dioxin toxicity (Bello et al., 2001; Nacci et al., 1999). Potentially, the SNP in the first intron of CYP1A affects transcription or is linked to SNPs affecting transcription. Thus, data chapter 4 sought to quantify the amount of transcription in vitro between populations containing different CYP1A promoter haplotypes linked to the selectively important SNP in the New Bedford Harbor triad. Haplotypes were determined by sequencing 1.5 kb of CYP1A promoter and an additional 500 basepairs containing the first exon and first intron for 8 individuals in the New Bedford polluted population and flanking reference sites, Point Judith and Sandwich. Representative promoters from each population were cloned into a luciferase-containing pGL3-basic vector and transfected into PLHC-1 cells. Luciferase activity was quantified under control conditions (vehicle) or after dosage with a prototypic CYP1A inducer, 3-methylcholanthrene (3-MC).

The study found significant variation in the CYP1A promoter. The vast majority of this variation fell in regions that have no described function in binding transcription factors. While sequence variation distinguished population divergence and established genetic distance by geographical distance, when analyses for evolutionary relationships were conducted on the entire CYP1A promoter through phylogenetic analysis, there was no pattern. Point Judith was the only population that formed its own monophyletic group. This population is physically isolated (with a man-made seawall). As a result, there was little variation in the promoter, most likely due to inbreeding in the population and a lack of
migrants into the population. This study, aside from describing the variation within the promoter, also sought to determine if the SNPs detected as being under selection in chapter three were in linkage disequilibrium with other SNPs along the promoter. Selective sweeps often act upon regions of a genome where a selectively important SNP causes hitchhiking effects. Thus, while our SNPs may not cause functional effects, they may be linked to SNPs or regions that have an underlying functional effect. This study, however, did not find any strong LD between these selectively important SNPs and the rest of the promoter, the first exon or the first intron. LD may in fact be with other SNPs or regions up or downstream of these SNPs. Tests for natural selection on the CYP1A promoter also failed to reject the null hypothesis that the promoter was evolving by random genetic forces.

In addition to understanding the overall genetic variation and evolutionary forces acting upon the promoter, this chapter sought to determine if there was a difference in the inducibility of the promoter to a prototypic PAH. CYP1A has been found to be refractory to induction in *F. heteroclitus* residing in polluted estuaries. If the promoter had a part in this phenotype, the expectation is that the inducibility would be less for the promoters from polluted individuals as compared to clean, reference individuals. A dose response curve on two individuals (one from Sandwich and one from New Bedford Harbor) showed that each of the promoter constructs was able to significantly induce luciferase expression over control when treated with doses of 3-MC higher than 0.05 μM for New Bedford and 0.1 μM for Sandwich. The promoter constructs also induced luciferase expression in a dose-dependent manner, and hit an asymptote at a dose of 1 μM. This 1 μM dose was then used to determine if there were population specific differences in the promoter constructs to induce luciferase
expression. There were significant population differences found: New Bedford Harbor had higher average fold induction over both its reference sites. This result does not explain why CYP1A is refractory to induction in the New Bedford Harbor population. It is possible that there are repressors not present in the cell line in which the transfection assays were carried out, or transcription factor binding sites well upstream of the 1.5kb that was cloned, which are necessary for the refractory phenotype. It is, however, interesting that the portion of the CYP1A promoter that was cloned from New Bedford Harbor was more sensitive to induction by 3-MC as compared to the reference sites. This sensitivity may reflect changes in the ability of the promoter to bind the heterodimer, AHR-ARNT, or may reflect a change to other transcription factor binding sites that were outside the scope of this study.

This thesis demonstrated the utility of high-throughput molecular biology and population genetics and genomics to examine selective forces in multiple natural populations. Anthropogenic selective forces have shaped the genomes of *F. heteroclitus* along the east coast of the United States by changing allelic frequencies of 1-7% of the genome in individuals exposed to constant high levels of pollution as compared to reference populations. Since such a large number of loci are under selection, it is hard to imagine that only one to a few of the selectively important loci are contributing to the resistance phenotype, but that it is the sum of the interactions between loci that allow *F. heteroclitus* to adapt to a very hostile environment.
Future Directions

AFLP Study (Chapter 1)

The AFLP study identified 24 outlying loci among pairwise comparisons of pollutant impacted *F. heteroclitus* populations and both surrounding reference populations. These loci should be sequenced and characterized in order to understand their overall function in the *F. heteroclitus* adaptive phenotype to pollution.

**Sequencing:** Loci were initially amplified through a PCR reaction of ligated DNA with primers containing a selective extension of three base pairs. A fourth and fifth selective nucleotide could be added to the selective extension to further decrease the total number of products generated and selectively amplify the candidate loci. With each additional selective nucleotide, the total number of amplified fragments should decrease by 75%. For example, if 70 loci are generated after the first selective extension (*primer +3* selective nucleotides), a reaction using a primer with four selective nucleotides will generate approximately 18 fragments. If those 18 fragments are further amplified with a primer with five selective nucleotides, approximately four fragments will remain. Since each outlying locus is characterized by the primer set from which it was amplified as well as its size, it can be easily identified and isolated from other loci. Products could be resolved on a high percentage agarose gel to allow for isolation and purification. Loci could then be cloned into a TA vector, PCR-amplified using vector specific primers, and sequenced in both directions.

**Identifying outlying loci:** Sequences initially could be compared against GenBank using the Basic Local Alignment Search Tool ([http://www.ncbi.nlm.nih.gov/BLAST](http://www.ncbi.nlm.nih.gov/BLAST)) to determine similarity to sequences in the database. Loci sequences with no match will be
mapped to a *F. heteroclitus* bacterial artificial chromosome (BAC) library to identify proximal genes due to the absence of a sequenced genome.

**Characterizing outlying loci:** All identified loci represent a polymorphism among populations in the genomic DNA by virtue of the way they were identified, that being outlier distribution based on the presence or absence of a PCR product. Loci may represent one of three areas of genomic DNA: coding regions, regions proximal to coding regions, and non-coding (distal) regions.

**Coding Regions:** If a locus is within the coding region of a gene, it should be determined if the locus results in a synonymous or non-synonymous change to a codon. If the polymorphism changes the deduced, primary, amino acid sequence of the protein, the protein could be expressed using a bacterial or baculovirus expression system such as the VariFlex™ system (commercially available through Stratagene) to determine whether the altered amino acid results in an altered protein function. This will be more or less feasible depending on the protein and whether it has a defined function. Appropriate assays could then be developed to test the (altered) function of the protein. For example, ligand binding assays might be done with receptors while DNA binding assays might be run with transcription factors.

If the polymorphism is synonymous, the gene’s expression should be determined. For genes with altered expression between populations, the contribution of the gene to a pollutant response could be tested with dose-response experiments using a mixture of pollutants found in the particular Superfund site and quantitative, real-time PCR.
*Proximal Regions:* Constructs could be made in sequences from the polluted population to assess changes in promoter strength (reporter constructs expressing luciferase activity) as was done in chapter four.

*Distal Regions:* Sequences could be compared against known, functional binding sites using TRANSFAC software as was completed in chapter four. If binding sites are determined, column affinity chromatography with site specific oligonucleotides could be used to purify transcription factors, which could be further characterized. To determine if the binding is biologically relevant, an *in vivo* footprinting assay could be run. These assays would determine if distal regions bound distinct transcription factors in polluted populations versus reference populations which could account for some of the resistance phenotype.

*SNP Characterization (Chapters two and three)*

For chapter two, several additional resources would have been useful for SNP identification and characterization. First, a fully sequenced genome would have provided a backbone through which the SNPs could have been mapped and polymorphisms within and between populations more easily identified. This also would have provided a context through which linkage and linkage disequilibrium could have been determined between SNPs. A recent study in the mosquito (*Anopheles gamnai*) used measures of LD (LD-based haplotype diversity analysis) to determine selection acting upon insecticide resistance mutations in several populations (Lynd *et al.*, 2010). If a genome was available for *F. heteroclitus*, a similar study could have been conducted with the SNP data collected in chapters two and three, providing additional statistical tests for selection. This may have also
identified larger genomic regions that were under selection, making those regions targets for future studies. Because not all of the SNPs identified in our study are directly under selection, but rather are the result of hitchhiking effects, a sequenced genome would allow for the calculation of LD between our SNPs and the rest of the genome in order to identify other loci that may be directly under selection by pollution. It is the hope of the Fundulus consortium to get a sequenced genome in the near future.

The SNPs identified as being under selection are rich targets to explore the resistance phenotype in the three independent polluted populations. In the Newark population, a SNP in β2-microglobulin changes an aspartic acid residue to an asparagine. β2-microglobulin is a component of MHC class I molecules. Future studies could look at the functional consequence of this non-synonymous change and whether it affects the function of the protein by modeling the folding of the gene, determining if the change affects the way the protein interacts with others in the MHCI complex or establishing whether there is a difference in the immune function of wild type or mutant β2-microglobulin F. heteroclitus to pathogen challenges. The other 14 SNPs that occur in coding regions all result in synonymous changes. These SNPs may be optimal or suboptimal changes to the codon because of codon bias and could be determined with statistical tests (Comeron, Aguade, 1998). It is thought that optimal codons help to achieve faster translation rates and high accuracy (Shields et al., 1988; Sorensen et al., 1989). The rate of translation could also be tested between wild and mutant-types through a system developed by Björnsson and Isaksson (Bjornsson, Isaksson, 1988). This test may elucidate a potential role for these changes in
translational efficiency. These SNPs could also be linked to other selectively important SNPs, but without a sequenced genome determining these linkages is not possible.

**CYP1A Promoter (chapter 4)**

For the CYP1A promoter, the functional activity of each of the XRE’s has yet to be tested. In order to do so, deletion constructs containing none, one, and two XREs could be used to test if the inducibility of the promoter changes in the presence or absence of any one of the XREs. An *in vivo* footprinting study would also be useful to identify functional cis-acting elements along the promoter between populations. Furthermore, the genetic variation and promoter inducibility by xenobiotics needs to be tested in the other two triads. The SNP in the first intron was under selection in all three triads, so its LD in the other two polluted populations must be evaluated for the same region sequenced in chapter four. With a sequenced genome, the LD of these SNPs with more distal regions could also be evaluated. Additionally, potential enhancer and repressor regions much further upstream of what was sequenced could be investigated to provide insight into the high inducibility of the CYP1A promoter found in the New Bedford Harbor population *versus* its reference sites.

**Overall Conclusion**

This thesis established that *F. heteroclitus* living in polluted areas along the east coast of the United States are adapting to contamination through natural selection. Natural selection seems to be acting on many loci along the genome, and several of these areas are shared between polluted sites indicating a conserved evolutionary response. There are also
selectively important loci specific to each polluted site, indicating a specific selective force given a particular contaminant mixture. In total, without *a priori* sequence information, genome scans proved useful in identifying loci that have been selected for or are linked to areas that are selectively important and may be contributing to the resistance phenotype.
References


